The special issue on Advances in Computer Science and Electronic Systems contains the accepted papers in the 2011 Iberoamerican Conference on Electronics Engineering and Computer Science (CIIECC 2011), which was held on April 13-15, 2011, in the city of Aguascalientes, Mexico at Universidad Panamericana campus Bonaterra. The conference aim was to bring together researchers, industry professionals, and academics working in any discipline of electronics engineering and computer science to exchange ideas and discuss solutions that contribute to the technological development of the iberoamerican countries.

CIIECC 2011 featured technical sessions, keynote speeches, tutorials, panel sessions, and workshops on topics of current interest, as well as industry exhibitions.

CIIECC 2011
www.ciiecc.org

UNIVERSIDAD PANAMERICANA

SEP

ISSN: 1870-4069
www.ipn.mx
www.cic.ipn.mx

INSTITUTO POLITÉCNICO NACIONAL
"La Técnica al Servicio de la Patria"

Advances in Computer Science and Electronic Systems

Carlos Delgado Mata
Carlos A. Gutiérrez
Ramiro Velázquez Guerrero
Juan Humberto Sossa Azuela
(Eds.)

Vol. 52

RESEARCH IN COMPUTING SCIENCE

ISSN: 1870-4069

Advances in Computer Science and Electronic Systems

Carlos Delgado Mata
Carlos A. Gutiérrez
Ramiro Velázquez Guerrero
Juan Humberto Sossa Azuela
(Eds.)

Vol. 52

RCS Research in Computing Science

Table of Contents

Índice

Page/Pág.

Computer Science

Electronic Systems

Storyline for Adaptive Learning Virtual Environment (SALVE)

Frederic Kleinermann, Jordi Ribera

Department of Computer Science, Vrije Universiteit Brussel, 1050 Brussel, Belgium
{frederic.kleinermann, ribera.jordi}@gmail.com

Abstract. Serious games can be very engaging by being visually attractive and by simulating a professional environment. But they are often too predictable, impersonal and not so user-centric. To address these issues, a number of research works have been performed on adaptations where the virtual scenario adapts according to what and how the players need to learn in a given context. But most of these works focus only on adaptation for single user. It is therefore important to look at serious games with multi-user capabilities. This is challenging as the virtual scenario should adapt according to a group of learners having different skills and having to learn a common goal. This paper describes an approach called SALVE that allow authors to create a storyline (or scenario) in which they can define the type of adaptations they want to introduce for a specific learner or a group of learners so that the storyline adapts according to what the learners do inside.

Keywords: Serious Games, Virtual Environment, E-learning, Multi-Learner Virtual Environment.

1 Introduction

Today, Virtual Environment is appearing through the internet as an online community that often takes the form of a computer-based simulated environment, through which users can interact via avatars with one another and use and create objects. An example of that is Second Life [1] where people are represented by avatars and they can meet, socialize and explore the virtual world all together. These sort of virtual environments are often referred as Collaborative Virtual Environments (CVEs) or Multi-User Virtual Environments (MUVEs).

Recently, these Virtual Environment (VE) platforms have been used to develop games with training and learning goals. These sorts of games are called Serious Games and can be seen as a Virtual Learning Environments (VLE). They provide sort of Virtual Environments based on Games with the aims that the user is learning something. To some degree, Second Life could be seen as a serious games and a number of research works have been conducted to use Second Life as a Multi-learning Virtual Environments. An example of such works is the PIVOTE project [2]. However, the richness of such a VLE can also become its weakness. The learner may be overwhelmed or get lost in the VLE [3], not knowing what to do first or next, or may be distracted too much and not be able to focus on the actual learning task. For people not familiar with VEs (novice users), the time required to get acquaint with

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 1-13
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

such a VLE (i.e. learnability) may be long and therefore their short-term satisfaction may be low. On the other hand, youngsters used to play video games, may spend their time in activities not very much related to the learning activities, especially if they have low motivation for learning. This then results in a low effectiveness. These concerns are confirmed in [4]. One way to solve these problems is by providing VLE with an adaptive way, e.g., adapted to the individual learner and to the progress that he makes during the learning.

A number of research work already exists (see section 2) and use adaptation for improving the learning phase. But they are mostly made for single learner. Today with the possibility of Collaborative Virtual Environments (CVE) over the internet, it is also becoming important to look at the possibility of using adaptation in the context of multi-learner environment. But this is challenging as it is important to be able to distinguish between a specific learner and a group of learners. In that kind of multi-learner virtual environment, learners share the same space and as a result any modifications are seen by all.

Based on our previous work [5] [6] aiming at bringing adaptation to VE in the context of E-learning based on single learner (i.e. VLE), we have developed a new approach called Storyline for Adaptive Virtual Learning Environment (SALVE) aiming at creating adaptive storyline (or scenario) for Multi-learner Virtual Learning Environment (MVLE). The SALVE approach provides authors a way to create a storyline (or a scenario) in which they can add a number adaptations they want to introduce for a specific learner or a group of learners so that the course adapts according to what the learners do inside. The rest of the paper is structured as follows. In section 2, we consider related work. Section 3 introduces the SALVE approach. In section 4, we give some description of the software architecture implementing the SALVE approach. Section 5 presents a case study. Finally, section 6 presents conclusions and future work.

2 Related Work

This section starts by introducing existing works dealing with storyline (or scenario). It then reviews research works that use adaptation in the context of E-learning.

2.1 Storyline (or Scenario)

The *alVRed* [7][8] is a project aiming at the creation of a set of tools for designing non-linear and interactive stories in virtual environments. This approach considers *scenario* to be a 'story'. They have developed an authoring tool for scriptwriters that allows authors to model objects participating in the story. Although this work is interesting, it does not allow author to model adaptation somehow into the scenario.

In [9], they presented an approach on modelling technique for urban traffic which uses Environment Description Framework (EDF) and Scenario Description Language (SDL). EDF is a modeling technique concerned with modeling the static portion of the scene while SDL is a scripting language to describe scenarios in the world. The

work presented is limited to a specific domain i.e. urban traffic. The scripting language is not so intuitive. Furthermore no adaptation is taken into account.

In [10] a new approach called Simulation Behavior Specification Diagrams (SBSD) has been developed and it uses a scenario description language tailored towards one particular type of VE namely military mission simulation. An interesting aspect of this work is that the approach uses a visual language to model the scenario. But this approach does not allow to model full scenarios.

In [11], they have developed a scenario description language which is reminiscent of UML use cases. Instead of modeling all possible courses of a scenario, the action frame scenario language only models a single execution of it. Although what they have presented is interesting, they have not used adaptation.

In [12], they look at adaptive and intelligent Web-based Education Systems (WBES) that take into account the individual student learning requirements, by means of a holistic architecture and Framework for developing WBES. They have developed a framework that includes an authoring tool, a semantic web-based evaluation, and a cognitive maps-based student model.

2.2 Adaptation of the content/presentation for single user

Brusilovsky et al. [13] have integrated some adaptive hypermedia methods into virtual environments by developing an approach that supports different navigation techniques in the context of 3D E-Commerce. This work is interesting because it has extended some of the adaptive hypermedia methods (such as direct guidance, hiding, sorting) to 3D environments.

In 2000, Chittaro and Ranon [14] have described how to introduce adaptation inside e-commerce. Their approach is called ADVIRT. A set of personalization rules exploits a model of the customer to adapt features of the VR store. They have also customized and personalized the navigation and different layouts of the store. In 2002 [15], they have introduced a software architecture for adaptive 3D web sites called Awe3D (Adaptive Web 3D) which can generate and deliver adaptive Virtual Reality Modeling language (VRML). In 2007, the same authors [16] has explained that adaptation can happen for navigation and interaction in order to help the users in finding and using information more efficiently. Finally, and based on their previous work [17], Chittaro and Ranon have extended the E-learning platform EVE [18]. They introduced Adaptive EVE that is tailored to the knowledge level of a student and to their preferred style of learning.

Santos and Osorio [19] have introduced another approach for adaptation in VR. Their approach is called AdapTIVE (Adaptive Three-dimensional Intelligent and Virtual Environment) and is based on agents, called Interactive and Virtual Agents that assist the users and help them to interact with the environment.

Celentano and Pittarello [20] have developed an approach for adaptive navigation and interaction where a user's behavior is monitored in order to exploit the acquired knowledge for anticipating user's needs in forthcoming interactions.

3 SALVE Approach

MVLE are made of virtual objects that have visual appearance, behaviours and for which interaction can happen either with the learner or with other virtual objects. In our case adaptation may happen on the visual appearance of virtual object, on the behaviours or on the interactions. In other words, in the SALVE approach, we state that a virtual object can be in different states. Each state can be seen as an adaptation happening at a particular time. We call these states Virtual Environment Adaptations State (VEAS).

To trigger on or off a VEAS, a number of condition will need to be fulfilled like for instance if the learner A has seen a previous virtual objects before this one, then trigger this adaptation. This can be seen as a relation type in which a virtual object is in a certain adaptation state before the condition is fulfilled and then in a different adaptation state after the condition is fulfilled. We call these relation types Virtual Environment Relation Type (VERT). VERTs will also be based on the user's profile. The third element in our approach is the adaptive storyline itself where the authors can compose the storyline using not only these VEAS and VERT, but also other concepts like *next step*, *parallel step* and *pedagogical objects*. Furthermore all these VEAS need to take into account the fact that there are several learners at the same time (i.e. multi-learner environment). In other words these adaptations can be specific to a learner or to a group of learners. To achieve this, we have developed in our approach the notion of leaner's ID using *ID*, *colour* and *sound*. Each of these components will now be described in this section.

3.1 Identification of learners

To **extend** from single learner to multi-learner, it is important to be able to distinguish between a specific learner and a group of learners. Indeed in that kind of multi-learner virtual environment, learners share the same space and as a result any modifications are seen by all. The SALVE approach allows three types of identifications that can be combined namely colour, sound and a user ID. By bringing colour, sound and user ID as identifiers, we can then make adaptations targeted to a specific learner or to a group of learners. For instance, a building A is suddenly highlighted in the red colour. This will create not only an adaptation of the Virtual World, but it will also be a way to have a specific adaptation to a learner identified by the red colour as his ID. Sound can also be used in that way. Nevertheless, there can be situations where the same adaptation happens at the same time as several learners do the same action at the same time. For instance, suppose that different learners go to see the same virtual painting at the same time, then perhaps *enableAnnotation* is triggered. Since there are several learners, the annotation will be highlighted in several colours at the same time. This gives the problem that learners have no idea if the annotation is meant for him. To overcome this sort of problem, the approach uses:

- talkInChat: Normally, chats are text in which the learner can communicate through a panel. This can be also used to communicate to a learner or a group of learners.

- HeadLearnerDisplay (HLD): It shows a panel above the head of the avatar representing the learner inside the MVLE and this can be used to prompt messages that are specific to a learner.

Another ways is to use sound. There, the sound can adapt to a specific learner or a group of learners, by assigning a different sound to each of them.

3.2 Virtual Environment Adaptation State (VEAS)

When an adaptation is applied to a virtual object, we can say that the virtual object is in a certain state. These adaptations are called Virtual Environment Adaptation State (VEAS). The SALVE approach provides a library of VEAS that an author of a storyline can choose from to introduce adaptations in his storyline (or scenario). In this paper we outline some of them as an example. They are different types and they are not all related to only adaptation of content and presentation. For instance, to visually indicate that an object has not yet been studied, we may want to give it a different colour to attract the attention of a learner and we may change the colour to the one that identifies him. We will now outline some of the adaptations which have been defined in our library of VEAS.

The first category of adaptation types for objects is concerned with the adaptation of the visualization of an object, i.e. how to display it and how to hide it. Because of the length of paper, we will only mention some of them here.

– *semiDisplay*: this adaptation type is used to display the object in a semi-manner, by having a semi-transparent bounding box around it.

– *changeSize*: this adaptation type is used to change the visual appearance of an object by changing its size.

A type of adaptation which is more specific to pedagogy is marking objects to draw the attention of the learner. Marking an object can be done in different ways and we distinguish two different adaptation types for marking :

– *spotlight*: this adaptation type allows to mark an object by putting a spotlight on the object; in this way the object becomes more visible and can be used to draw the attention of the learner to this object.

– *highlight*: this adaptation type allows to mark an object by drawing a box around the object where only the edges of the box are displayed.

Another type of adaptation which is not related to visual appearance is the ability to change behaviours and interactions. Possible adaptation types for behaviors are:

– *enableBehavior:* this adaptation type allows enabling a behavior associated with an object. Note that we have also the opposite which is *disableBehaviour*.

- *enableInteraction*: this adaptation type allows enabling an interaction type (given as parameter) for an object. Not that the opposite exists and is defined as *disableInteraction*.

So far, we have discussed possible adaptations for individual components (object, behavior, interaction, …) of a MVLE. But we have adaptations that can have an impact on several components of the VE or on a part of the VE. We call them *adaptation strategies*, as they can be used as strategies to adapt a VE.

- *navigationWithRestrictedBehavior*: this adaptation strategy allows restricting the possible behaviors of objects while navigating. The restricted behavior can apply on all objects (in the VLE) or on a specified list of objects.

- *freeWithSuggestions*: this adaptation strategy will allow the learner to navigate freely in the VLE but in addition some objects will be "suggested". Suggesting is done by using marking (i.e. spotlight or highlight).

- *displayAtMost*: this adaptation strategy allows to specify when some objects should not be displayed anymore. The condition can be given by means of some pedagogical criteria like the knowledge level the learner currently has for the object or by setting a limit on the number of times the object should be displayed.

The adaptation strategies presented here are only a subset of possible adaptation strategies that we have developed in our approach.

3.3 VERT and Learner's Profile

VEAS should be activated under certain conditions. These conditions correspond to some kind of rules that are called Virtual Environment Relation Types (VERT). The VERT allows us to also encode and retrieve information on a learner or a group of learners such as how many times a certain learner has interacted with an object, how fast does the learner go or has he found certain clues. Note this can also be applied to a group of learners. More can be found in [6].

3.4 Adaptive StoryLine

The SALVE approach helps authors to conceptualize a story like for instance a course. The conceptualization of a story is done through the use of VEAS and VERT. Our approach has been inspired by the work done in the video game community and the way they actually generate a storyline [21]. In our approach, the conceptualization of a story is done from a high-level (i.e. away from the code level) and then the code itself is automatically generated.

In video games, there are different ways to look at a storyline. Often, a storyline is split among different levels where a player goes from one level to the next. For instance, when the player has collected enough points in one level, he can then progress to the next one. This way is very much linear. However inside a level, the

storyline can be very non-linear as in some video games for instance, the player has to collect different types of coins. The way he collects them can be in a non-linear way as he may collect two red coins and then suddenly three green coins and then another red coins. This helps the writer of a game to go from a linear way in his storyline (levels) to a non-linear way [21]. In our approach, we want to follow a similar way where the author of a story may want to see a global view of his storyline much more like sections in a linear way, but he also wants to give some freedom the way learners explore each sections. For instance, he wants to say that the learner should visit the modern arts gallery before going to the abstract art gallery. However, the way the learner explores the different paintings in the modern art gallery can be random.

To achieve this, our approach uses the notion of *next Step* that provides a general view of the storyline much more like a linear way. To make, the storyline be non-linear when several learners are involved, we have introduced the notion of *parallel steps*. Parallel steps are made by sub steps (like for instance next step) that can be run in parallel i.e. there are no precise orders in which they are executed. This depends only on the learners and how they explore the virtual world.

To introduce adaptation in the conceptualization of a storyline, each sub steps can use VEAS and VERT (described above). This will provide a way to say that if learner A does now this, adapt the virtual world like this. However, if the learner B does another thing, then the virtual world should also be adapted. The non-linearity will happen as these two types of adaptations can happen in parallel and as a result there is no way to know in advance if learner A does something before learner B or vice-versa.

4 Software Architecture

Our approach has been implemented through a software architecture shown in figure 1. They are made of different elements.

The first element is an authoring tool that allows an author to create the adaptive storyline. It provides the adaptation types (i.e. VEAS) that are used for this storyline, the VERTs are translated into rules and facts that can then be uploaded on the reasoning engine. The second element is the reasoning engine which reasons, using these rules and facts, on the storyline and what the learners do. The third element is the interpreter engine responsible to interpret the results given by the reasoning engine and communicate to the MVLE what adaptation state should be applied to which virtual object. It also does the opposite by receiving information from the MVLE and then knows what to relate to the reasoning engine based on the learner's profile. The fourth element is indeed the Multi-learner Virtual Learning Engine. Note that MVLE has been implemented using OpenSim [22] and the reasoning engine has been implemented using Prolog Engine [23]. The interpreter has been written as Web Service written in Java. The authoring tool has been developed using Air Adobe builder [24]. The MVLE architecture is a distributed architecture.

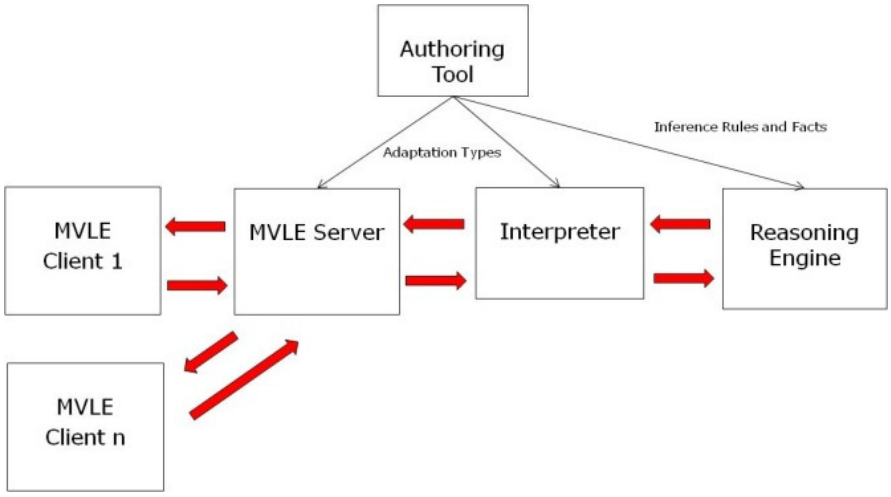


Fig. 1. Diagram of the structure of the system

5 Case Study

In this section, we will show some initial results based on our software architecture which implements the SALVE approach. The storyline is a course aiming at introducing the Etterbeck campus of the Vrije Universiteit Brussel (located in Brussel, Belgium) to new students. Every new students are represented by an avatars.

5.1 Creating the storyline

To create an adaptive storyline, the author of the course uses our authoring tool to compose his storyline (i.e. the course). We will now introduce elements of that storyline based on that case study. From figure 2, it can be seen that the storyline is deployed over a timeline axis. This gives the author an overview of the course in a linear way. This timeline is split into a number of stages identified by a number. Each stage corresponds to a step in our approach. We will now review them. The first stage is a *Start Step* which is represented graphically in the storyline by a red triangle. It allows the author to introduce the time to complete the course and the initial indications on how to reach the targets. It also contains a number of initial parameters like for instance, colours used to identify group of learners or individual learners.

The second stage in the storyline corresponds to a *Next Step* in our approach which is represented graphically by a blue round. In the case study, it contains the fact of visiting the *WISE* department.

The third stage corresponds to *parallel steps* which has a number of sub-steps that can be run in parallel bringing to non-linear way of the story. They are represented graphically in the storyline by a green square placed perpendicularly to the storyline In each sub-steps, a number of VEAS and VERTS can be added to bring adaptations.

For instance in one of these sub-steps, the author can define first the conditions that will trigger adaptations. Here if the learner represented by the blue colour, has not visited the athletic track (see figure 3), then some annotation should be displayed to the learner (see figure 4) telling him what to visit. Note that in our approach, the storyline is also displayed in full text (see bottom of figure 2, *Description of the course*).

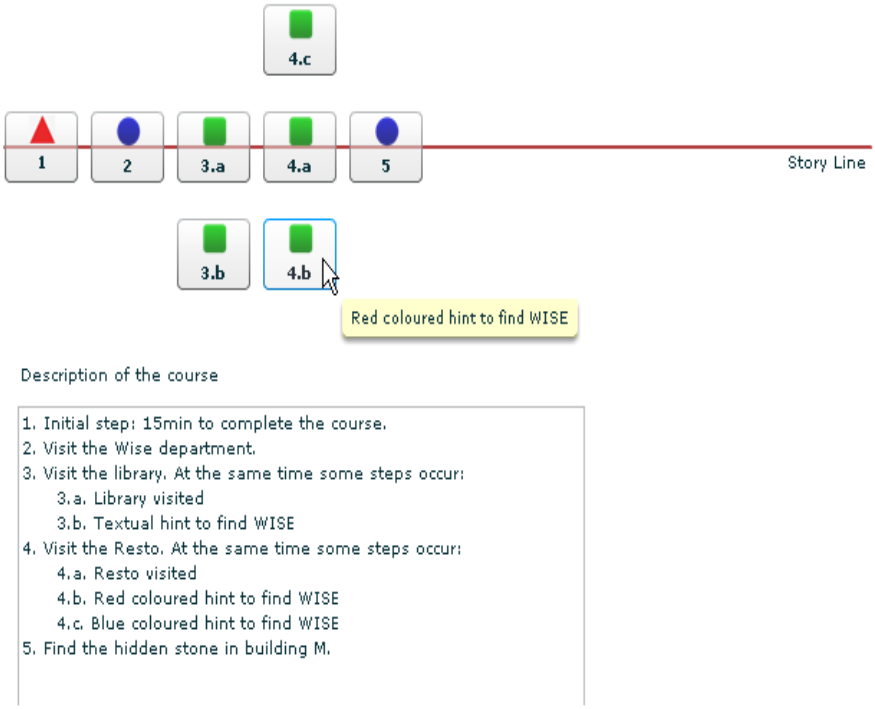


Fig. 2. Storyline and textual description of the course

In the SALVE approach, for each condition, we can also have several VEAS. We have represented that by means of a graphical notation to make a simple visual representation for the author (see figure 5). For each VEAS added, there will be an arrow from the condition to the VEAS, which is represented by a rhombus and its label (see figure 5).

5.2 Running the storyline

The course is ready to be run once the authoring tool has generated all the necessary adaptation (VEAS) for the OpenSim platform and for the interpreter. And once the different VERTs has been translated into prolog rules and uploaded to the reasoning engine. Figure 6 shows the first kind of adaptation where a specific learner receives a text message telling that he has visited the library and what he should visit next depending on his profile. Figure 7 shows an adaptation where the *WISE* department is

highlighted in the colour specific to a group of learners to attract their attention by showing them where the *WISE* department is located. This adaptation is shown only when the time has passed a certain threshold so that they can hurry and stop loosing time wandering where the *WISE* department is located.

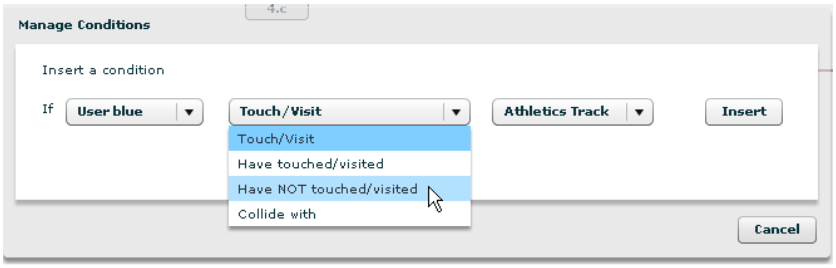


Fig. 3. Introduce a condition

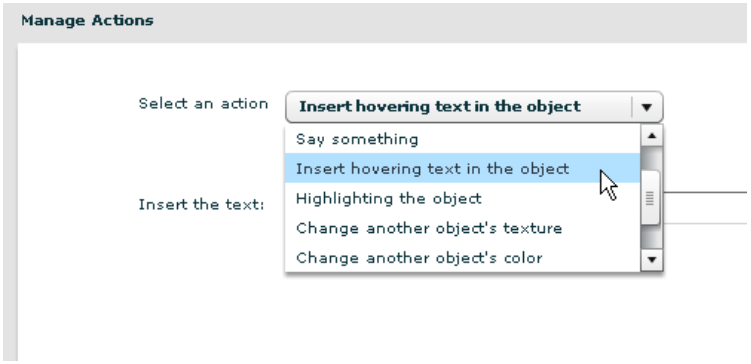


Fig. 4. Add a VEAS

6. Conclusions and Future Work

This paper presented an approach called SALVE aiming at creating adaptive storyline for Virtual Learning Environment in the context of multi-learners. It presented a number of adaptation (VEAS) and how they can be used. It also presented some initial results. The use of adaptation in multi-learner virtual environment is challenging as learners share the same space and any change will influence all the learners. It is important to find ways to make sure that adaptation can be tailored not only to a specific learner but also to a group of learners. Future work will be to address these challenges further and assess the usability of the authoring tool.

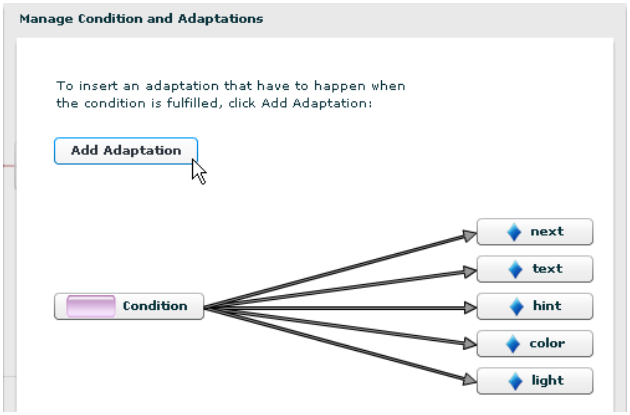


Fig.5. Add several VEAS

Fig. 6. Textual hint to the learners

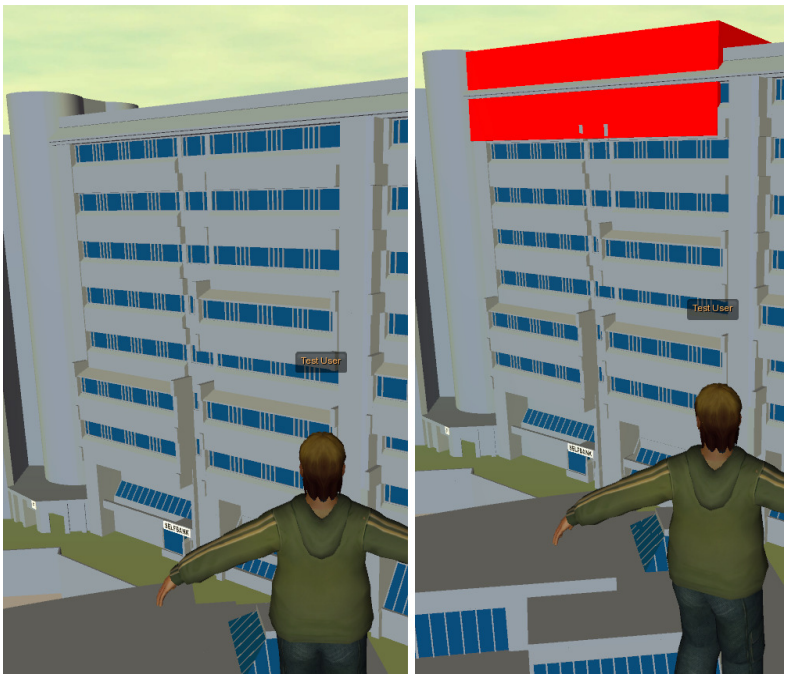


Fig. 7. WISE department highlighted in red

References

1. SecondLife, http://secondlife.com/
2. PIVOT, http://www.daden.co.uk/pivote.html
3. Bricken, M., Byrne, C. M.: Summer students in virtual reality: A Pilot Study on Educational Applications of Virtual Reality Technology. In: Wexelblat, A. (ed.), Virtual Reality: Applications and Exploration. New York: Academic Press, Inc., pp.199--218 (1993)
4. Virvou, M., Katsionis, G: On the usability and likeability of virtual reality games for education: The case of VR-ENGAGE. In: Computers & Education, Volume 50, Issue 1, January 2008, pp. 154--178 (2008)
5. De Troyer, O., Kleinermann, F., Ewais, A.: Enhancing Virtual Reality Learning Environments with Adaptivity: Lessons Learned. In: HCI in Work & Learning, Life & Leisure(USAB 2010), Austria, pp. 244--265 (2010)
6. De Troyer, O., Kleinermann, F., Pellens, B. Ewais, A.: Supporting Virtual Reality in an Adaptive Web-Based Learning Environment. In: EC-TEL 2009 : Learning in the synergy of Multiple Disciplines, LNCS, vol. 5794, Springer-Verlag, France pp. 627--632 (2009)
7. Grutzmacher, B., R. Wages, Trogemann, G.: An Authoring System for Non-Linear VR Scenarios. In: The 9th International Conference on Virtual Systems and Multimedia. Montreal, Canada, pp. 8--16 (2003)

8. Wages, R., B. Grutzmacher, Trogemann, G.: A Formalism and a Tool for Diverging Requirements in VR Scenario Modeling. In : The 13th International Conference on Artificial Reality and Telexistence. Keio University, Tokyo, Japan, pp. 8--14 (2003)
9. Willemsen, P.J.,: Behavior and Scenario Modeling for Real-Time Virtual Environments. PhD Thesis, University of Iowa (2000)
10. Bartley, C.R.: A Visual Language for Composable Simulation Scenarios. Master Thesis, Air Force Institute of Technology, Ohio (2003)
11. Ohnishi, A., Potts, C.: Grounding Scenarios in Frame-Based Action Semantics. In: Proceedings of the International Workshop on Requirements Engineering: Foundation of Software Quality, Switzerland pp. 12-18. (2001)
12. Canales, A., Peña, A., Peredo, R., Sossa, H.,, Gutiérrez, A.: Adaptive and intelligent web based education system: towards an integral architecture and framework. Expert Systems with Appplications. 33(4), 1076--1089 (2007)
13. Brusilovsky, P., Hughes, S., Lewis, M.: Adaptive Navigation Support in 3-D E-Commerce Activities. In: Proceedings of Workshop on Recommendation and Personalization in eCommerce at the 2nd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'2002), Malaga, Spain, pp. 132-139. (2002)
14. Chittaro, L., Ranon, R.: Adding Adaptive Features to Virtual Reality Interfaces for E-Commerce. In: Proceedings of International Conference on Adaptive Hypermedia and Adaptive Web-based Systems. LNCS, vol. 1892, pp. 86-97. Springer-Verlag, Berlin (2000)
15. Chitttaro, L., Ranon, R.: Dynamic Generation of Personalized VRML Content: a General Approach and its Application to 3D E-Commerce. In: Proceedings of 7th International Conference on 3D Web Technology, Web3D 2002, pp. 145-154. ACM Press (2002)
16. Chittaro, L., Ranon, R.: Adaptive 3D Web Sites. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web - Methods and Strategies of Web Personalization, LNCS, vol. 4321 pp. 433-464. Springer-Verlag, Berlin (2007)
17. Chittaro, L., Ranon, R.: Adaptive Hypermedia Techniques for 3D Educational Virtual Environments. IEEE Intelligent Systems, vol. 22, issue 4, pp. 31-37 (2007)
18. Chittaro, L., Ranon R.: An Adaptive 3D Virtual Environment for Learning the X3D Language. In: Proceedings of the 2008 International Conference on Intelligent User Interfaces (IUI 2008), pp. 419-420. ACM Press, New York (2008)
19. Santos, D., C. T., Osorio, F. S.: AdapTIVE: An Intelligent Virtual Environment and Its Application in E-Commerce. In: Proceedings of 28th Annual International Computer Software and Applications Conference (COMPSAC'04), pp. 468-473. (2004)
20. Celentano, A., Pittarello, F.: Observing and Adapting User Behaviour in Navigational 3D interface, In: Proceedings of 7th International Conference on Advanced Visual Interfaces (AVI 2004), pp. 275-282. ACM Press (2004)
21. Ince, S.: Writing for video games. A & C Black Publishers Limited (2006)
22. OpenSim, http://opensimulator.org
23. Prolog,: http://www.swi-prolog.org
24. Adobe AIR, http://www.adobe.com/es/products/air

Clustering for Virtual Environments Using Open Source Tools

Felipe Gómez-Caballero[1], Carlos Delgado-Mata[2], and Jesús Ibáñez[3]

[1] Furui Laboratory, Department of Computer Science,
Tokyo Institute of Technology, Tokyo, Japan
[2] IPIT (Instituto Panamericano de Investigación Tecnológica),
Escuela de Ingenieria, Universidad Panamericana campus Bonaterra,
Aguascalientes, México, C.P. 20290 cdelgado@up.edu.mx
[3] Grupo de Tecnologias Interactivas,
Departmento de Información y Tecnologias de Comunicación,
Universitat Pompeu Fabra,
Barcelona, España

Abstract. This paper presents the design and development of Virtual Reality and Virtual Environment technologies destined to foment computer science knowledge and to increase the use of top range technologies among undergraduate students in a university. The prototype developed is capable of creating virtual environments for diverse analysis and is based on computer clusters (distributed systems) of type *n servers-client* for virtual reality and object-oriented programming in C++ language, using Open Source software. Thus, significantly reducing the investment needed to use virtual reality technology.

1 Introduction

The interest for making a virtual environment project was born when a myriad of feasible applications on diverse fields was envisioned. One such field is the interaction of human beings through computer systems. For an extensive overview of interaction with Virtual Environments, the reader is referred to [8]. Moreover, these systems provide a very useful tool to visualize and to understand objects and phenomena from diverse fields of study (some examples are: climatological phenomena, medical studies, new product prototypes, building visualization).

The virtual environments can go further, for example: by introducing time dependent processes and phenomena, and self dependant elements within the environment. We can also create applications that don't need external intervention to represent certain level of "life". An example of this is the use intelligent agents by applying the concepts of Artificial Intelligence (AI), for more information about intelligent agents in virtual environments, please refer to [2].

However, Virtual Reality and Artificial Intelligence are tasks that require a lot of computer power to run on a virtual environment. The kind of processing power that is needed is provided, in most cases, by expensive computers; therefore implying a high

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 14-25
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

investment. In order to solve this problem we propose the use of computer clusters [1], which are made up of several PC's working together to perform a common task, dividing the processing load among them.

To carry out the research, by means of this technique, we implemented the solution with the smallest feasible cost, and that was achieved by using Open Source [12] tools as our primary development platform. This platform includes the operating system and the necessary software to manage a computer cluster [6][7].

The project described herein is relevant at this moment because Universidad Panamericana campus Bonaterra has research related to the Virtual Reality and fields derived from it [4][5][10][3]. For example, three-dimensional modelling and video game development [11]. However, the resources needed to complete these projects were not available to obtain the results that were pursed.

Because of that, it was necessary to create a virtual environment in which research within CINAVI (Research Centre for Intelligent Virtual Environments, by its abbreviations in Spanish) could be carried out. Furthermore, the aim is also to make this kind of technology available for other fields that can take advantage in using the provided technology. This gives a technological and academic advantage for Universidad Panamericana campus Bonaterra and thus can be applied with the regional industry. Therefore the technology becomes more accessible.

2 Project overview

2.1 System description

The final prototype of the project was developed so that it could be used as a tool that allow to demonstrate the advantages of Virtual Reality technologies by using low cost computers with average processing power (desktop computers). This tool lets us visualize any 3D model created following the VRML standard and later loaded within the application. The model is visualized through the displays of the computers working as graphical servers. The application allows 2 types of navigation within the virtual environment to let the user feel free to explore the "world". The framework developed in this project is expected to be used in further research applications and projects and thus increasing its potential.

2.2 Restrictions

To develop this project we had four main restrictions, which were addressed by means of different development methods.

The first and most important, was the cost. The available budget was very reduced, so a study was carried out to find alternatives to accomplish the final prototype using limited resources. The chosen option was to use existing computers in one of the labs

in our university.

Because of that, the second restriction arose, which is that the processing power of the available computers in the labs was average, an study was carried out to analyse a solution that could allow the development of an application that should use the computers' processing power available to the maximum.

The third restriction was the time available for the creation of the virtual 3D models for the environments of the University. Thus, the used models were developed by members of CINAVI.

The last restriction was that we wanted to let the students use the project and continue to support and improve it with the aid from other researchers. The licensing model to enable this is free software and Open Source [12], therefore it was chosen.

The project development was conducted as follows: Creation of 3D virtual models for the tests, comparison of clustering technologies, application development and performance tests.

As result, a prototype was created with the following characteristics:

- Any three-dimensional model created or exported on the VRML language can be loaded.
- Separation of processing loads by means of a cluster architecture, in such way that can work in several desktop computers to visualize simple and complex 3D models.
- The number of cluster nodes within the architecture can be changed without having to make modifications to the system.
- Several visualization windows in a single computer can be created.
- The navigation in the virtual environment follows the well known metaphor of mouse and keyboard.
- There are two types of navigation: "fly" navigation and "walk" navigation.

We followed the next requirements to build up the system. These are listed by priority:

1. The system should be able to use the maximum processing power of each computer within the cluster
2. The system should be able to display real-time images when responding to user movements.
3. The system should be able to load different three-dimensional models.
4. The system should provide an adequate navigation within the environment.
5. The system should be easy to use.
6. The system should be able to use as many servers as needed.
7. The system should be multi-platform.

As it can be seen, the project was designed and documented with the intention of being improved in future projects.

2.3 Creation of 3D Virtual Models

Different three-dimensional virtual models of Universidad Panamericana campus Bonaterra were developed (Engineering School, main building, etc.). The models were created with AC3D [9], a 3D modelling, multi-platform, software with a simple and intuitive interface. The models were created by students –members of CINAVI.

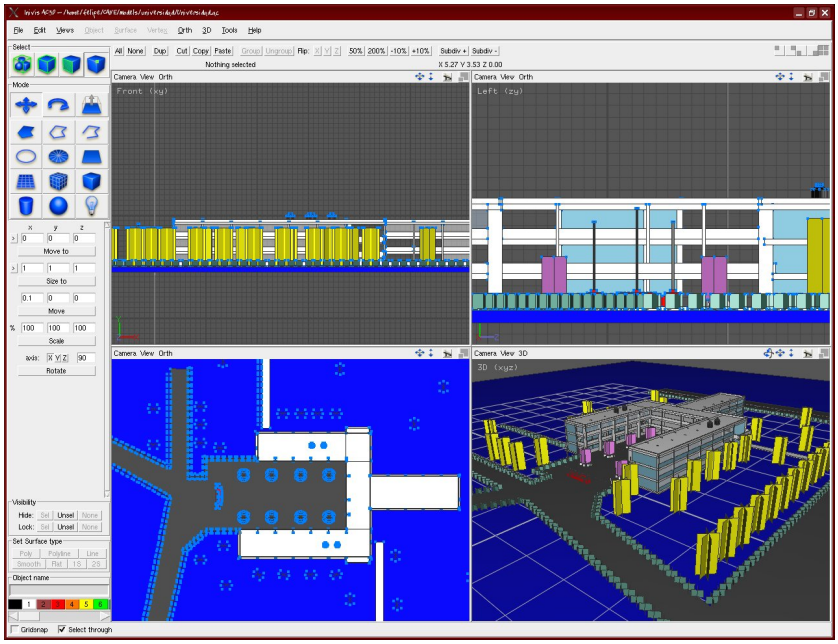


Fig. 1. Main building 3D model.

All the models were created based on the available blue-prints of each one of the university's buildings in order to capture all their details. Fig. 1 and Fig. 2 show some of the models created for the virtual world.

2.4 Comparison of clustering technologies

The reason for implementing a cluster configuration for this project was the demand of processing power required to handle 3D models, render tasks, visualization and manipulation in real time within the virtual environment. This is usually done in high performance computers [15], which are expensive.

Several applications and architectures for cluster implementation [6][7] were analysed during this stage, and it was found that there are a wide variety of cluster configurations as well as other options in software for high performance data processing. The

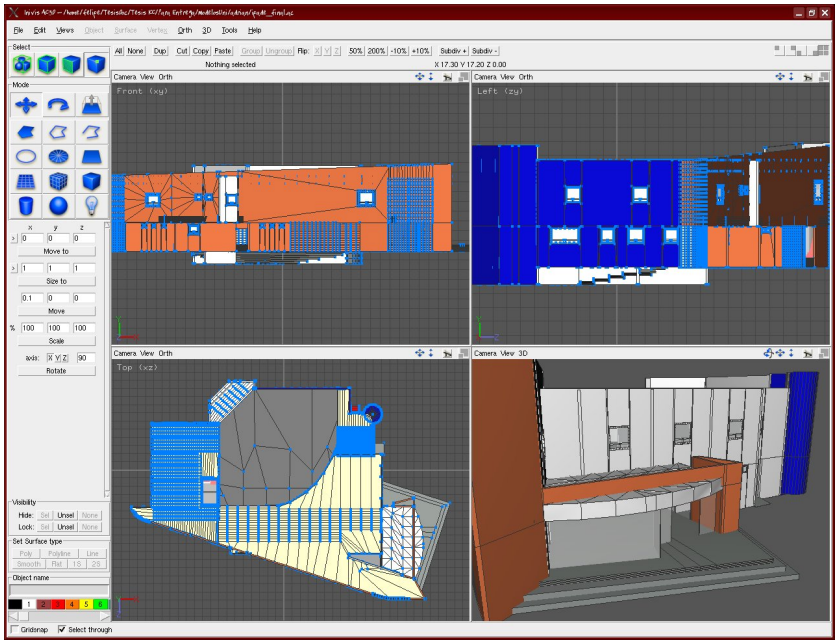


Fig. 2. IPADE building 3D model.

options vary in cost, platform and kind of use for the final application. Among popular applications that were found, are Beowulf, OSCAR and some proprietary solutions from Hewlett Packard and IBM, but none of them fulfilled the requirements to cover the expectations of the final application, because these required an extra configuration for the data flow between the different PC's to create a "n Servers-Client" configuration (Fig. 3). this configuration is similar to the one used for the prototype application on this project.

In order to cover the specifications, we used a cluster implementation from OpenSG scene-graph system [13], which sits on top of OpenGL[14]. OpenSG was chosen as the main tool to run the virtual environment scenes and handle the 3D graphics due to advantages on its competitor OpenGL Performer, like cost and portability. OpenSG was developed following Open Source principles and free software licenses (LGPL). Further, it's a multi-platform software. Another advantage of using OpenSG over the mentioned solution is that the graphical interface and the cluster administration is unified in a single application, reducing the configuration and implementation difficulty.

2.5 Project codification

The project was coded in C++ language using GNU/g++ libraries from GNU/Linux Slackware 10.2 operative system, as well as graphic and clustering libraries from OpenSG.

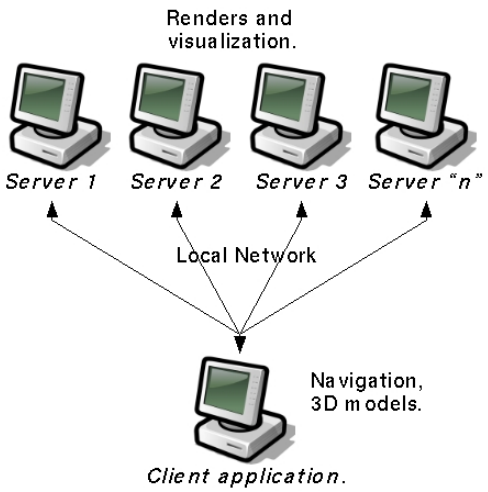


Fig. 3. Cluster architecture.

The project is based on two main modules: Server program and client application. In order to use the n Servers configuration, the client application was coded in such way that it could receive extra parameters through the command line, like the names or IP's of the servers that are being executed in the network and thus to distribute the process load of graphic visualization tasks among them.

One task of the client application is the interaction with the user of the virtual environment, using the GLUT library (GL Utility Library). This library creates the window where the movements of mouse and the events of the keyboard are captured. On the other hand, we created the scene-graph using OpenSG on the client to manage the resources used by the objects within the environment, and therefore we could access any of them.

The process to optimize the resources consists on constant examination to recognize non visible objects and then omit the render process for these objects in order to perform a better and faster visualization.

Another task performed in the client program, is the organization of every object into nodes, the assignation and management of all the matrix transformations related to them, and that are going to be performed within the environment. The process is carried out in the following order:

Firstly, a root node is created and thence all objects are linked to it (land geometry, creature control, navigation control and camera). The land node contains a single geometry file for the land, in such way that there is just one contact point for the navigator, making it easier to manage collision detection of the user. At the same level, OpenSG

creates a node for navigation to manipulate the navigation process.
A hierarchic node tree is created like the one shown in Fig. 4. Also, the client program manages the network connections between itself and the server programs, sending the necessary information of the 3D models through the connection, as well as the events that modify the visualization, this is shown in Fig. 5.

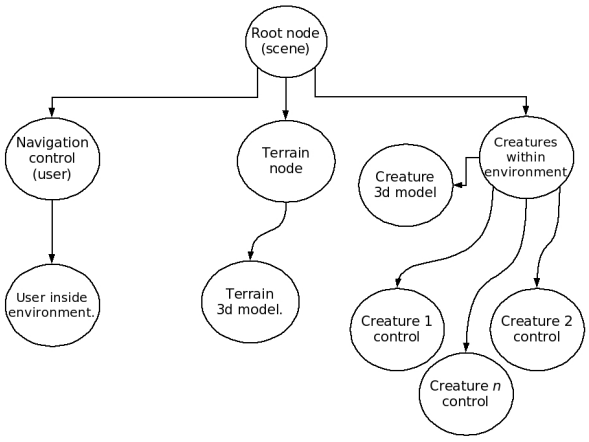


Fig. 4. Environment node tree.

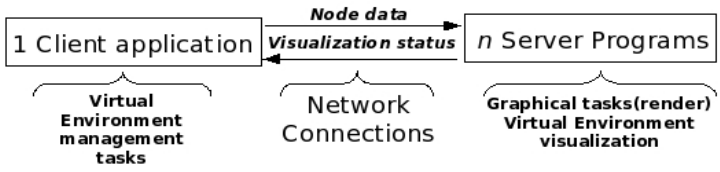


Fig. 5. Network connections client/server.

Secondly, a program was developed for the computers that work as servers. Those computers just need to respond to requests from the client program and receive data through the network connection. Each one of these servers is executed specifying a unique name, so that they can be recognized by the client program when it is executed and it tries to make the connections.

Each server program is executed as an independent process on the network, so each one could carry out the graphical tasks, and it initializes the visualization processes as soon as the execution of the client application is detected. The latter one sends the three-dimensional model data to the servers.

The server programs are always listening into the network looking for possible running client applications, in this way different environments can be loaded from the client with no need to reinitialise the execution of each server program.

The client application and the servers, won't show any visualization windows until the connection between all the cluster nodes are established, after completing this operation, an automatic synchronization between all the involved elements is performed in order to begin the visualization.

However, if there is more than one client in execution, the servers would take care of requests from the first client that handles the connection and synchronization successfully. At the end of the execution of the first client, they would initiate the connections with another one in standby mode. Fig. 6 shows a possible scenario to illustrate the cycle that the cluster would follow in the previously stated case.

There are two types of network connections: Multicast connection (automatic search for servers) and stream sock connection (using server addresses). The multicast connection tries to find the servers checking the whole network for online computers running the server program, this operation can cause a reduction in connection speed, in the worst case scenario a deadlock can be caused when not finding a suitable connection.

Therefore, it was decided to use StreamSock connections, because the cluster was tested and since the application will be used within a university, we tried to affect as little as possible the network traffic. Besides, this kind of connection is safer and more reliable than multicast and therefore the network reliability is not affected.

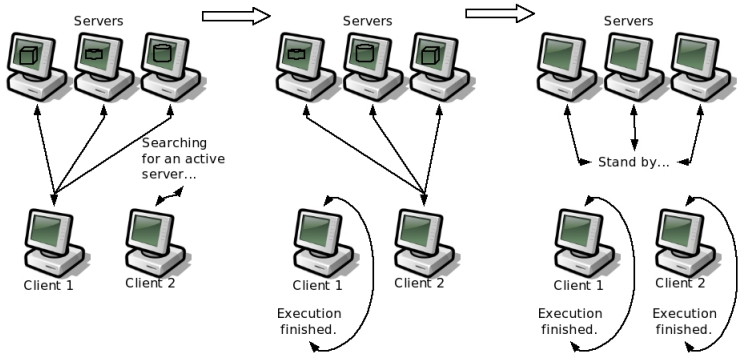


Fig. 6. Life cycle scenario.

3 Performance tests

Several tests for different architecture configurations, different cluster servers and different 3D models of the university were carried out. The architectures studied were using a single computer and from two to four computers processing the task load as a single system.

We obtained better results using a four computer configuration: three of them processing graphical tasks (servers) and one of them working as the client application or navigator for the virtual environment. This configuration was chosen among the others tested because the visualization and distribution of processing loads were optimal to show detailed 3D models on common desktop computers with average processing power.

For the project test and implementation three desktop computers were used to work as servers, with the following technical specifications:

Hardware: Intel Pentium 4 1.6 Ghz processor, 256 MB RAM, network card 10/100, nVidia GEForce AGP 64 MB video card and 80 GB hard disk.

Software: GNU/Linux Slackware 10.2 operative system , GNU gcc/g++ 3.3.4 compiler, OpenSG (cvssnapshot 02-02-06).

Also, one computer was used to work as client with the following specifications:

Hardware: Intel Pentium 4 2.6 Ghz processor, 512 MB RAM, network card 10/100, nVidia GEForce 4 mx440 AGP 128 MB video card and 40 GB hard disk.

Software: GNU/Linux Slackware 10.2 operative system , GNU gcc/g++ 3.3.4 compiler, OpenSG (cvssnapshot 02-02-06) and AC3D 4.0.

All the virtual three-dimensional models of Universidad Panamericana campus Bonaterra were loaded on the computer running the client application.

4 Results

At the end of this research project, a virtual environment that shows Universidad Panamericana campus Bonaterra was developed, using existing resources within the engineering school. Furthermore, this prototype is going to be used as a virtual tour guide application. Moreover, the project is seen as a foundation for new projects within CINAVI.

The clustered VE is shown in Fig. 7. The servers work together to carry out the visualization of the virtual environment which is divided between the number of graphic servers within the cluster architecture. For example, if there are three servers running, the whole view of the environment will divide itself in three screens. Also, there is an extra computer working as a client (or navigator) within the environment. This does not have any visualization, and the movement within the VE depends on the mouse interface connected to the client and it is reflected on the screens of the graphical servers.

The work described herein demonstrates that a virtual environment can be implemented without the need to acquire computers with high processing power. The processing power required can be achieved by dividing the processing loads between the servers and the client. Highly detailed and very exact three-dimensional models can be visualized, this technology can be used for any other application in which high processing power is needed.

A side result of this project was the creation and/or conversion to VRML of 3D models of Universidad Panamericana campus Bonaterra. The models used on this project are available for future projects.

This project opens a possibility for a future research, as similar kind of technology [16] can be used with the approach of a cooperative virtual environment, in which different users can participate and interact within the environment with the purpose of making studies or tasks with a common aim. Besides the processing loads between the different computers connected (as clients or servers) can be distributed in a cooperative VE.

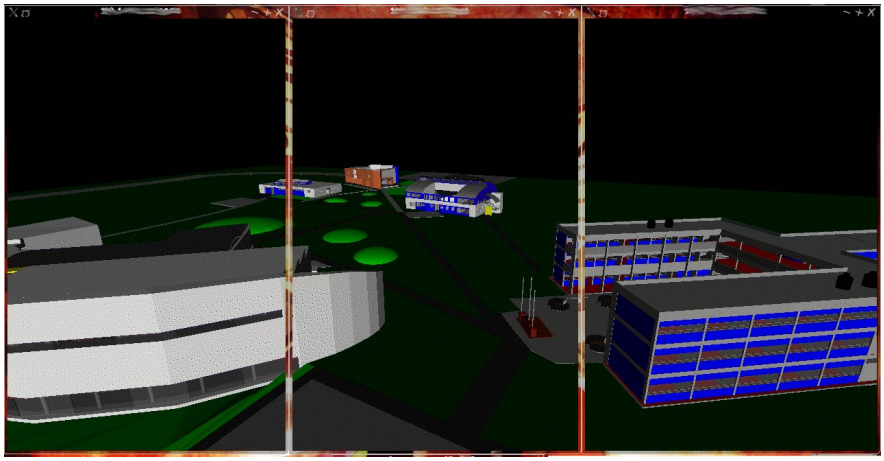


Fig. 7. Windowed cluster view.

5 Conclusions and future work

The design and implementation of this project has brought advance on the computer science academic field within the university. It has brought the possibility for further research projects and the liaison with the regional industry. Also, the interest from university students has arisen and this has caused the growth of the CINAVI group. An

important point about this research is the generated cost for the university was minimal for the implementation and tests. The only acquisition were AC3D licenses at a discount price of $29 USD each. Everything else was coded and created with free Open Source software.

On the other hand, the cluster architecture allows the use of average desktop computers from the university labs to perform the tests and execution of the virtual environment, and thus eliminates the necessity for new equipment to access of high processing power like the one required to run virtual environment applications. Furthermore, this project is a technological contribution not only for computer science field and the engineering school, due to the nature of this technology, but to other fields of study that can create ”virtual worlds” to extend its use inside the university in particular and the region in general.

References

1. Baker, M.: Cluster computing white paper. University of Portsmouth. United Kingdom (2000)
2. Delgado-Mata, C., Ibanez, J., Bee, S., Ruiz, R., Aylett, R.: On the use of virtual animals with artificial fear in virtual environments. New Generation Computing Journal Vol. 25 (2007)
3. Delgado-Mata, C., Cosio, B.M.L.: Hmm and nn for gesture recognition. In: IEEE CERMA. pp. 56–61. IEEE Computer Society, Cuernavaca, Mexico (September 2010)
4. Delgado-Mata, C., Ibáñez Martínez, J., Gómez-Caballero, F., Guillén-Hernández, O.M.: Transactions on edutainment i. chap. Behavioural reactive agents to define personality traits in the videogame Uberpong, pp. 135–149. Springer-Verlag, Berlin, Heidelberg (2008), http://portal.acm.org/citation.cfm?id=1809375.1809388
5. Delgado-Mata, C., Ruvalcaba-Manzano, R., Quezada-Patino, O., Gomez-Pimentel, D., Ibanez-Martinez, J.: Low cost video game technology to measure and improve motor skills in children. In: IEEE Africon. pp. 1–6. Nairobi, Kenya (September 2009)
6. Harbaugh, L.G.: Building high-performance linux clusters. Appro. USA (2004)
7. Hochstetler, S.: Linux clustering with csm and gpfs. IBM. USA (2004)
8. Ibanez, J., Delgado-Mata, C.: Lessons from research on interaction with virtual environments. Journal of Network and Computer Applications 34(1), 268 – 281 (2011), http://www.sciencedirect.com/science/article/B6WKB-50J9H7B-1/2/1fc48e7fa2b044cdfc8e9a9d73de9fde
9. Inivis-AC3D: http://www.inivis.com/
10. Ibáñez Martínez, J., Delgado-Mata, C.: From competitive to social two-player videogames. In: Proceedings of the 2nd Workshop on Child, Computer and Interaction. pp. 18:1–18:5. WOCCI ’09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1640377.1640395
11. Policarpo, F.: 3D Games Real-time rendering and Software Technology. ACM Press, USA (2001)
12. Raymond, E.S.: The cathedral and the bazaar. O’Reilly Media (2001)
13. Reiners, D.: Open source tools for virtual environments: Opensg and vrjuggler. UPGRADE, Vol. VII, issue no. 2 (2006)
14. Shreiner, D., et al.: OpenGL:Programming Guide. Addison Wesley, 6th edition edn. (2008)

15. Singhal, S., Zyda, M.: Networked Virtual Environments, Design and Implementation. ACM press SIGGRAPH Series (1999)
16. Yu, L., Yu, C., Wang, Y.: Multi-projector seamless displays system based on pc-cluster. Proceedings of the 16th international conference on Artificial Reality and Telexistence pp. 125–130 (2006)

Developing a Virtual Environment for Learning Geometry

Venustiano Soancatl, María-Luisa Cruz, Luis Huerta, Antonio León, Andrea Herrera, Lucina Torres*, William Zurita, and Inti Reyes

Universidad del Istmo, Departamento de Informática,
Carretera Chihuitan. S/N, 70110. Ciudad Ixtepec, Oaxaca, México
{venus,mlcruz,luisdh2,leonborges,
aismenee,ezurita,inti}@bianni.unistmo.edu.mx
luz83_83@hotmail.com*
http://www.unistmo.edu.mx

Abstract. In Mexico and in many other countries, students have many problems in the mathematical learning process, particularly in problem solving. The purpose of this project is to design and develop an educational computer-game system using a virtual environment to teach junior high school students how to solve geometric problems. It is based on two main approaches: game-based learning and question-led learning. The system includes a database that stores geometric problems and their step by step solutions, thereby leading students through the process of problem solving while keeping track of the students' performance. The design of this system permits the separation of the domain area and the game programming.

Keywords: virtual environment, game-based learning, question-led learning, geometry, problem solving.

1 Introduction

The fast growth in computer processing power has opened the door to new communication and entertainment forms, which includes virtual environments, complex computer games, social networks and immersive environments. Concerning computer games, they have evolved from simple 2D graphic interfaces in the 70's to 3D models in the 90's with virtual environments arriving at the beginning of the XXI century [1]. Now, the challenge is to use these new technologies to improve the quality of life. In education, the use of computer games to strengthen the learning process is becoming more and more popular [3]. This approach is called game-based learning (GBL) which is defined by Conolly and Stansfield as "the use of a computer games-based approach to deliver, support, and enhance teaching, learning, assessment, and evaluation" [4]. Many educational games have been developed that have their own characteristics and cover domains such as adventure, arcade, simulation, strategy, logic, puzzle and immersion [2]. Some research groups and companies are developing different kinds of educational games with great success. The aim of GBL is to use the engaging property

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 26-36
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

of the games in the learning process, such as: Virtual U to teach administration students [5], Timez Attack for learning the multiplication tables [6], Flooding Control Trainer *FCT* a game for training US Navy [7], UFractions (Ubiquitous Fractions) an educational mathematical mobile game, developed in SouthAfrica [8], and 80days a cultural game, developed in Europe [9]. These computer games have proven that it is possible to mix fun and knowledge to the benefit of many segments of society. The functionality of the GBL approach has not been proven, however, there are some reports about the impact of games in the learning process. The Teachers Evaluating Educational Multimedia (TEEM) company in the United Kingdom has done some research about the educational use of games [10]. They analyzed about 20 different games including different surveys applied to parents, teachers and students. The results are divided into four key stages and each stage represents a different set of knowledge or skills that the game helps to develop. The results show there are some topics that the parents think are very important for their children to learn. Parents were asked which subject they preferred in student games. Math, spelling and reading were most frequently mentioned. The prioritized skills by the parents were decision making (40%), design (25%), problem solving, strategy and cooperation (20%), and mouse control and taking turns (10%). In these surveys about (30%) of the students played computer games every day while the rest played once, 2-3 times or 4-6 times per week.

In Mexico, math competence is one of the greatest problems that students face, since many students score poorly in accordance with some national and international assessments such as the Assessment for Academic Achievements in School Centers (ENLACE), Educational Quality and Achievement Tests (EXCALE) and the Programme for International Student Assessment. They report that more than 50% of Mexican students are in a low proficiency level in the mathematics scale [11]. We propose using the GBL approach to help mexican students improve their mathematical skills. In this work we present the development of a virtual environment to teach geometry problem solving. Our goal is to join the fun of a videogame and the learning of geometry. Also we focus on the question-led learning approach to guide students through the process of problem solving [17]. To achieve this goal, we analyzed some important features of videogame content creation and in the following sections we explain the content design of our system including the virtual environment.

2 Content Creation

In videogames, the term content usually takes two different forms: assets and gameplay [13]. The term assets denotes every object that is presented to the user such as 3D models, objects, characters, animations, music and sound effects. Gameplay defines what the user does, the actions the user can do, and what the game has to do in response [13]. The creation of the assets involve graphic artists using diverse kinds of tools such as Maya, 3D Studio Max, Blender and POV-Ray to create 3D models and textures which also involve script writers and music composers [14]. Behaviors are the dynamic part of the game and define what players can do with the objects in the environment and how they react to the players' actions. Usually, behaviors are programmed by artificial intelligence (AI) programmers. Instructional content creation is similar to gameplay

content creation because it defines what the user can do and what the system has to do in response. In educational applications, teaching material is divided into a set of learning objects or instructional units. They are designed to be studied with no interruption. Learning objects contain some kind of text or image with an explanation. Also, there are different form of tests to assess the student.

Educational applications are sometimes called computer assisted instructions (CAI) systems, which require a domain expert cutting the domain knowledge into individual pieces to be presented to the students. A lot of tools can be used for building learning objects such as Power Point, Open Office, Macromedia AuthorWare or web based platforms. The simplicity and generality of these programs allows nearly all kind of domains to be represented. Unfortunately, the knowledge representation in these programs is inadequate and interaction capabilities is also very simple. Students learn reading and thinking instead of putting concepts into practice. This kinds of learning objects are not appropriate for procedural or problem solving skills, since the exponential demand of more content exceeds the capabilities of a single human author [18]. It occurs especially when the system should provide specific explanations for each common error. Some improvements to this model have been proposed to overcome these difficulties, such as "generative CAI". Another proposal is to use AI techniques for representing the instructional content, which is called Intelligent Tutoring Systems (ITS).

Using more sophisticated content representations allows us to create automatic explanations using natural language techniques, so user interactions can be richer, more open responses can be managed and analyzed, and the system will deduce user misconceptions or skills. More complex domains afford systems that follow the learning-by-doing approach [15], where students spend most of the time solving exercises in an environment with richer interaction. These kinds of systems have to decide the next exercise and infer how the student solves it, giving some feedback to help when mistakes are detected. Exercise creation and correction constitute the domain knowledge, so, systems must know the valid response and the more likely human mistakes and their causes. These systems also need the ability to propose exercises themselves. All the content material is highly dependent on the domain being taught and the representational techniques. This causes the creation of universal tools to be very difficult for ITS [13].

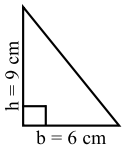
2.1 Content Design

In this paper we present a content design for an educational game system for teaching geometry. Problem solving is considered the most important part of math. However, teaching problem solving is not an easy task because there are lots of difficulties that students face when they try to solve problems [17]. In addition, when solving a problem there are a great variety of paths to obtain a correct solution causing a combinatorial explosion impossible to afford. Our system stores math problems in a database, which includes the step by step solution of every problem. Therefore, the system is able to lead the student through the process of problem solving.

Each problem statement in the database consists of text, images or 3D models, a time limit, difficulty level, and tips or recommendations, along with steps that lead the students to the solution of the problem. Each problem is associated with one of the

following geometry topics: triangles, quadrilaterals, circles, similarity, the Pythagorean theorem, trigonometry, perimeters, areas and volumes. Each step includes a multiple choice question and a recommendation if required. The possible answers and recommendations can be represented by text, figures, images or 3D models. The problems are designed following the question-led learning approach. Questions in this approach are focused to help students to reinforce their process of thinking, show them how to formulate precise statements and questions, bolster their habit of questioning, improve the quantity and quality of questions asked by them, and teach them different techniques and strategies for problem solving [11]. The next example shows a simple problem that could be stored in the database.

Problem: *Find the area of the following triangle*.



1. The formula you need to find the area is

   (a) $A = \frac{bxh}{2}$ (b) $A = bxh$ (c) $A = bxb$ (d) $A = \frac{bxh}{3}$

   tip: *remember that the triangle area is half of the area of a rectangle*
2. where b is

   (a) height (b) width (c) base (d) perimeter

3. and h is

   (a) base (b) width (c) height (d) perimeter

4. then compute the area using

   (a) $A = \frac{6x9}{3}\,cm^2$ (b) $A = \frac{6x9}{2}\,cm^2$ (c) $A = \frac{6x8}{2}\,cm^2$ (d) $A = \frac{9x6}{1}\,cm^2$

Each problem is linked to a specific object in a virtual environment and the object can be an image, a painting or a 3D model. An object can be associated to zero or more problems. A filename, description, location, scale and rotation angles are stored for each object. This information allows the system to display the objects in the virtual environment.

The database stores information about users, such as name, nickname, password, birthday and gender. Also, the number of solved problems, number of mistakes, scores, and number of sessions in the system is recorded. Figure (1) shows a diagram which includes three main entities: PROBLEMS, STEP and MUSEUMOBJECT, and their attributes.

Some advantages of this design are that the system will be able to update problems without modifying any programming code, store statistics about students performance, minimize dependencies between the domain expert and game programmer, use the same

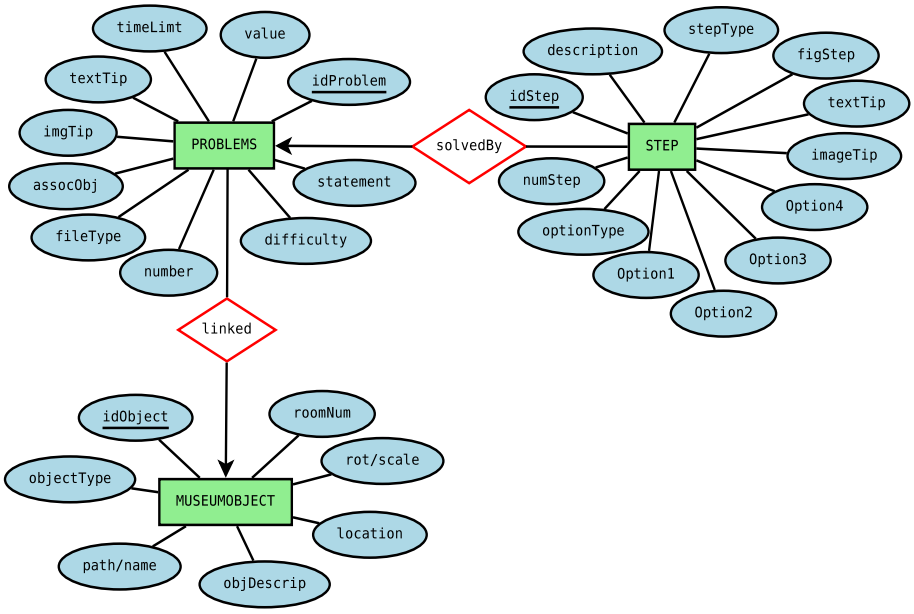


Fig. 1. ER Diagram of the database.

map to execute different exercises, use the same presentation to teach different contents, and to reduce the development cost by presenting different exercises on the same scenario.

The main strength of this approach is that domain experts can design problems without any programming skill. Also problems can have a high level of complexity and include as many steps as the domain expert wants. Another advantage to this approach is that almost any kind of domain can be represented.

3 Development of the Virtual Environment

The software is composed by two main modules, the graphic environment and the database control. Both modules are in constant communication while the system is running. Figure (2) shows the relationship between these two modules. The DB control module performs reading and writing transactions on the database. The graphic environment module creates the virtual world and displays the data for the geometric problems taken from the database. The graphic environment also reads and writes user information from and to the database by mean of the DB control.

3.1 Development of the Graphic Environment

The programming language we are using to develop the system is C++, since it is one of the most common programming languages to develop videogames. The 3D scenery

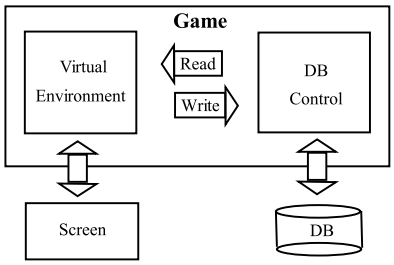


Fig. 2. System components and their interaction.

and graphics are displayed using the open source Irrlicht 3D Engine, which is written in C++ and compatible with NET languages. The irrKlang sound library is used to play sounds and music. Fonts and text are displayed using the FreeType library, which is a free, high-quality, and portable font engine. The database has been embedded using SQL Server Compact 3.5, a free easy-to-use database engine.

3.2 The Gameplay

Currently the museum has five functional rooms where the user can navigate and explore, with each room dedicated to a specific geometry domain. The domains of each room are the following: triangles, quadrilaterals, circles, areas, and solids. Each room has paintings on the walls of famous mathematicians and places. Each painting is associated with one or more problems. While the user navigates in the museum, a red light pointer is used to guide the direction of the movements and when the user points to a painting, a brief description is displayed at the top of the screen.

When users try the system for the first time, they must create an account which includes a nickname and password. All the user's statistical data are initially set to zero which includes the points, number of solved problems, and number of mistakes. These statistical data allows the system to keep track of student activities such as problems solved, mistakes committed and how long the system is used.

The aim of the game is to solve all the problems in the museum, which are designed using the question led learning approach [15]. When the user selects a painting by pointing and clicking on it, a problem is displayed on the screen and a chronometer starts running backward showing the time remaining to solve the problem. After several seconds, the system starts displaying the steps (multiple choice questions) one by one on the screen. When the user selects the correct answer the next step is shown, until the solution is found or the time expires. Figure (5) shows a screenshot of the system displaying a problem, question and four possible answers. In this version these answers are shown inside 3D objects which are floating like moving balloons. However, if the user makes a mistake, the system shows educational feedback and sends the user to answer several questions back, this is done as a kind of punishment. This make the user think twice before selecting a choice by chance. This process is shown in Figure (3) as a time dependent flow chart. When users solve a problem, the system gives them a certain number of points and additionally, depending on how fast the students solve the

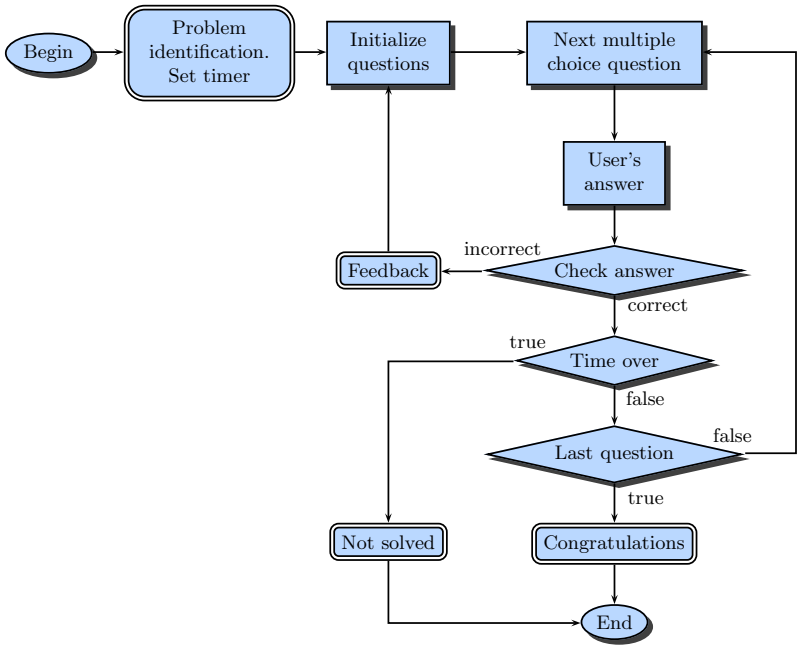


Fig. 3. Problem solving flowchart.

problem, they can earn extra points. The system stores the time required to solve the problems, number of mistakes and add points to their record. There is only one way for the students to fail when they are playing the game and it occurs when the time expires. In this case, the system stores the number of incorrect choices that the students made and their points do not decrease. This gives us information about of how the students tried to solve the problem. Additionally, we can infer if the students responded by guessing or if the time period was insufficient.

Once a problem is solved, it is deactivated and is not shown to the user anymore. That is why there must be many similar problems to assure the user understands the topic.

The users usually get feedback from the system, whether the user selects an object, responds a multiple choice question correctly, makes a mistake or solves a complete problem. The feedback could be recommendations, congratulations, sounds and images. When the user makes a mistake, the message ”Think about it and try again” appears. We propose to send only positive messages to the user.

3.3 Virtual Environment Components

We define the three principal components of the virtual environment as Scenery, Problem Manager, and Player. The Scenery component displays all the elements in the virtual environment, such as images, 3D models, textures, text and lights. The Problem Manager component reads the problems from the database, presents them to the user,

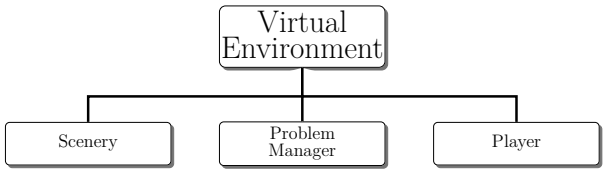


Fig. 4. Components of the virtual environment.

display the steps associated with the problems, evaluates user responses, and manages a chronometer to keep track the time limit for the problems. The Player component manages all the information about the user, such as scores and points, game sessions and solved problems. Figure (4) shows the main components of the virtual environment.

The screen is divided into four horizontal sections: user data, problem statement, virtual world and step statement. Figure (5) shows a screenshot of these sections. At the top of the screen, the user data is displayed in red, to the left is located the nickname (1), in the middle of the section users can see their score, and to the right, the number of solved problems in the current session. This information helps users track their progress in the game. The problem statement section is the area reserved to display the text statements of the problems in yellow and is located right under the user data section. The third section is the biggest area in the screen and is designated to display the virtual world where the user can navigate. This virtual world contains the interactive objects that display the problems, such as paintings, 3D models and possible answers to the multiple choice questions. The fourth section, at the bottom of the screen shows the questions or statements of problem steps in yellow. Figure (5) shows a view of the virtual world where the user's nickname is "Juanito", his score is 5 points and he has already solved two problems in the current session. Juanito is navigating in the triangle room and he has just started to solve a triangle geometric problem. The problem is displayed at the top of the screen and he is in the second step (question) of fourteen (7). His choices are in 3D transparent balloons (8,9,10), and at the top right corner is the chronometer showing the time remaining (6). In the scenario, there are some paintings on the walls, such as a famous mathematician (2) and a giant Olmec Colossal Head (3).

The system is in the initial phase to analyze its effectiveness as a learning tool, but teacher and student responses has been very positive. Teachers find the game useful for learning because it is based on videogames which have become part of the lives of this age group of students. Additionally, teachers like the fact that they can design their own material, set the difficulty level, and get statistical information about student performance. Students like the aspect of learning by combining 2D and 3D objects in a virtual environment. Students also asked if the level of difficulty could be increased by making the answers move faster.

3.4 Main Pedagogical Aspects

The main reasons we believe the system would be effective are the exciting and entertaining environment which keeps student attention on the screen. Curiosity stimulates

Fig. 5. Screenshot showing elements of the system.

learning and motivates a student, it can drive to seek out new and interesting information. The system can evoke a learner's curiosity by providing the environments that have an optimal level of informational complexity [12], since it can be designed by teachers and domain experts in a way that is neither too easy nor too complicated. Problems can gradually increase in difficulty and can be adapted or personalized by teachers. Problems can be designed using a pedagogical method such as Polya's four phased method [16] and include steps for critical reflexion. The problems in the system can be seen by the students as challenges, which can be overcame with feedback designed by teachers. This system also is focused on the learning by example approach since the system leads students step by step thorough the process of problem solving. An important advantage of the system is that teachers can easily identify statistical data to find students that need extra help to improve their scores.

4 Conclusions and Future Work

In this paper we have shown the development of a computer game system to teach mathematical problem-solving to students. First, we presented the design of the content of the system. The proposed system includes a virtual environment and a database. The graphics interface component controls the gameplay. The database stores information about the domain expert and users. This design permits updating the data domain expert without modifying the programming code. The Game-Based Learning approach was used to design the gameplay of the system that includes images, sounds and 3D

models. For the domain expert we focused on the Question-led Learning approach to design the problems stored in the database. The problems are presented in a sequence of steps that lead the students to find their solution. Each step includes a multiple choice question which is displayed using the virtual environment component. The clear separation between exercises and the virtual environment significantly reduces the cost of the educational game, since the virtual environment can be reused. The system is in the last development stage. The next stage is to apply usability tests and to distribute the system to schools and libraries.

Acknowledgments

We want to give many thanks to FOMIX-VERACRUZ and UNISTMO for supporting this work which is part of the project 95656. We also thank Kevin Mitchell for his contribution in the English translation.

References

1. Entertaiment Software Asociation, Game player Data, http://www.theesa.com/facts/pdfs/ESA_Essential_Facts_2010.PDF
2. Pérez Marqués Graells, Tecnología educativa, http://peremarques.pangea.org/videojue.htm
3. Paul G.: What Video Games Have to Teach Us About Leraning and Literacy, Palgrave Macmillan, First Edition (2003)
4. Thomas, C., Mark, S.: 'Using Game-Based eLearning Technologies in Overcoming Difficulties in Teaching Information Systems, Journal of Information Technology Education, Vol 5, pp. 459–476 (2006)
5. Eriic, A., C.: Software Review Virtual U: A Simulation of University System Management, Information Technology,Learning,and Performance Journal,Vol. 19, no. 1, Spring (2001)
6. BigBrainz : The ultimate multiplication tables tool, http://www.bigbrainz.com/
7. Hussain, T.S., Roberts, B., Bowers, C. Cannon-Bowers,J.A., Menaker, E.S., Coleman,S.L., Murphy,C., Pounds,K., Koenig,A., Wainess,R.,Lee,J.: Designing and Developing Effective Training Games for the US Navy, Interservice, Industry Training, Simulation, and Education Conference *(IITSEC)*,Paper No. 9477,(2009)
8. Turtiainen, E., Blignaut, S., Els, C., Laine, T. and Sutinen, E.: Story-based UFractions Mobile Game in South Africa: Contextualization Process and Multidimensional Playing Experiences. Proceedings of the Second Workshop of Story Telling and Educational Games, (STEG 2009)
9. Kickmeier-Rust, M. D., Albert, D., and Mattheiss, E.:An educational guide to planet earth: Adaptation and personalization in immersive educational games. In Y. Cao, A. Hannemann and B. F. Manjn (Eds.), proceedings of the 2nd International Workshop on Story-Telling and Educational Games, (STEG 2009)
10. Mc Farlane, A., Sparrowhawk A., Heald, Y. :Report on the educational use of games. In :Teem publications, www.teem.org.uk/publications/
11. Venustiano, S., Lucina A., María-Luisa C., Antonio L., Carmen M.: Un enfoque para la Enseñanza-Aprendizaje de solución de problemas de Matemáticas basado en preguntas, 11th Congreso Internacional y 14th Nacional de Material Didáctico Innovador, UAM, México, pp. 97–103,(2010)

12. Thomas, W., M.: What Makes Things Fun to Learn? A study of Intrinsically Motivating Computer Games. Cognitive and Instructional Sciences Series CIS-7 (SSL-80-11), (1980)
13. Marco, A. G., Pedro, P.G., Pedro A. G.: Games-Based Learning Advancements for Multi-Sensory Human Computer Interfaces. Information Science Reference (an imprint of IGI Global), Hershey New York (2009)
14. John, P. F., Omar, S.: Software Engineering for Game Developers. Thomson Course Technology PTR., Boston, MA (2005)
15. Marc, P.: Digital Game Based Learning. MacGraw Hill, USA (2004)
16. Polya, G.: How to Solve It: A New Aspect of Mathematical Method, Princeton University Press, (1989)
17. Venustiano, S., Antonio, L., Carmen, M., Lucina, T.: Leading Students to Solve Math Problems Using Question-led Learning. In: 4th European Conference on Games Based Learning, pp 368–376. Copenhagen, Denmark (2010)
18. Joshua, T., Karen, T., Magy-Seif E., Marek, H.: Authoring Tangible Interactive Narratives Using Cognitive Hyperlinks. In: INT3 '10 Proceedings of the Intelligent Narrative Technologies III Workshop ACM, New York, NY, USA (2010)

Adaptive Physics for Game-balancing in Videogames

Carlos Delgado-Mata[1], Jesús Ibáñez[2], and Alfredo Villalobos[1]

[1] IPIT (Instituto Panamericano de Investigación Tecnológica),
Escuela de Ingenieria, Universidad Panamericana campus Bonaterra,
Aguascalientes, México, C.P. 20290 cdelgado@up.edu.mx
[2] Grupo de Tecnologias Interactivas,
Departmento de Información y Tecnologias de Comunicación,
Universitat Pompeu Fabra,
Barcelona, España

Abstract. This article describes the work in progress of an adaptive video game that is based on the arcade game of air hockey. The video game uses adaptive physics to improve the experience between two human opponents. The physics of the game continually adapts to the ability of each player. That is, the game becomes more difficult for the skilful player; however, it becomes easier for the inexperienced player. This is achieved by using a behaviour based architecture; first used on mobile robots and later on artificial characters and computer driven video game opponents. The architecture affects the physical properties of some elements of the game, which are handled by the physics engine used in the project described herein.

Keywords: adaptive physics, physics engines, game balancing

1 Introduction

Video games have reached a near photo-realism graphical level. Therefore, virtual images can be confused with the images of reality. This creates a high expectation on video game players; however, the performance of some computer driven opponents does not correspond to the level of graphical realism. Thus, the user's perception is drastically affected. To resolve this problem, research has been carried-out to explore various techniques to improve the believability of the characters' behavior. In [8] a discussion on characters' believability in virtual environments is presented. Another example is that of the game Überpong [9], where the the behaviour of opponents are provided with a personality profile to make them more believable.

1.1 Adaptive Games

To improve the user's experience, research has also been carried out on video games that adapt to the players' ability, and thence to enrich their gaming experience. One such example is the so-called *game balancing*. That is, the game continually changes so that the player enjoys the gaming experience (and therefore plays longer). Moreover, he should not get frustrated if the game is too complicated and he should not get

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 37-46
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

bored if the game is too simple. An example of this research is to modify the complexity in case of one player [12]. Similarly, there is research on games that are modified (adapted) depending on the player's emotional state. A relevant example is presented in [20], where one game mechanic (opponent speed) is continuously adjusted depending on the player's emotional state. The data was used to predict the mode of play (boring, frustrating or enjoyable), and depending on it the game changes the speed of the opponent accordingly. Furthermore, preliminary results have been published on research on the modification of complexity in a game between two human opponents [15]; the objective of game is social rather than competitive. This type of game is appropriate where a player is experienced (for example a parent) and the other is novice (for example a preschooler). The work described herein uses adaptive physics to enhance the game experience for two human opponents.

1.2 Physics in video games.

As mentioned above, the level of graphical realism in the latest video-game generation is close to photo-realism. Because of this, the industrial and the academic community have decided to improve the objects' behaviour. Thus, an steep increase in the use of physics engines in video games has occurred. Moreover, some video games are predominantly a physical simulation, some standout examples are: The game produced by renowned filmmaker Steven Spielberg, Boom Blox, where objects interact according to their physical properties, such as mass, acceleration, deformation and gravity. Another game that uses a physics engine predominantly to manage the behaviour of the game's elements is The World of Goo; where the player creates physics based constructions (for example bridges) to move balls (Goo) from one place to another. Other games have been used as benchmarks for physics engines; one such example is Cellfactor, which was developed by the colombian company Immersion Games. This game was used as a demonstration of power of the Ageia PhysX chip. To add physics to video games it is best to use physics engines; some of them are commercial and some are open source.

The relevant commercial physics engines are:

1. Havok, an irish library [11], is used in hundreds of games developed for the console market leaders (Wii, PS3 and Xbox 360), as well as, games developed for the PC. Some of the standout games developed with Havok are: The excellent adventure game developed for PS3 and Xbox 360: Assassins Creed, and the games Super Smash Bros. Brawl and Boom Blox developed for the Wii.
2. PhysX is a physics library first acquired by Ageia and subsequently by Nvidia (the market leader in graphics processing units GPUs) [17]. PhysX is now included as a physics process unit (PPU) in an integrated circuit. PhysX is used in hundreds of video games, most notably are the acclaimed Batman: Arkam Asylum and the successful Unreal Tournament 3.

The relevant open-source physics engines are:

1. Box2D is a 2D physics library [4] developed by Erin Catto. Box2D's main advantage is that it is multi-platform. The library has been used in various languages and environments such as C++, Java, Flash, iPhone and Android.

Fig. 1. Air hockey table.

2. Bullet physics engine is a 3D engine [5], which was developed by Erwin Coumans. The engine has been used in video games and in movie effects. It is available for several platforms, such as the PlayStation 3, Xbox 360, Wii, Mac, iPhone, Linux and Windows. It has been used in various games and movies; notable examples are the successful video game Grand Theft Auto IV and the 2012 Hollywood film.
3. Chipmunk Physics Engine is a 2D physics engine [14] developed by Scott Lembcke. This library is similar to Box2D, but it is simpler. It is primarily used for iPhone applications.
4. Open Dynamics Engine is a 3D physics engine [18], which development began in 2001. It has been used in games like Bloodrayne 2.
5. Tokamak is an open source physics engine [13] developed by David Lam, no information was found on relevant video games using Tokamak.

2 Adaptive Physics

The video game described herein is based on the arcade game Air Hockey.

Air hockey is a game that consists of a table, two paddles and a puck (see Figure 1). The game is an abstraction of the popular canadian game of Hockey. The board consists of a flat surface surrounded by a railing to prevent the puck and the rackets from flying out of the table. The most recent table models include a device that provide an air-cushion to reduce friction. In the two smallest rail segments there are holes where the puck enters. That is, they act as goals.

The configurable items on the air hockey game (see Figure 2) proposed in our work are:

Fig. 2. Configurable items.

1. Player's 1 goal. The size of the goal of player 1 can be modified as follows:

$$goal1_{length}(t) = goal1_{length}(t-1) + \Delta goal1_{length} \tag{1}$$

That is, depending on the player's skill, the goal's size can increase or decrease.

2. Player's 1 paddle. The mass / volume of a player's 1 paddle could be modified as follows:

$$r1_{mass}(t) = r1_{mass}(t-1) + \Delta r1_{mass} \tag{2}$$

$$r1_{vol}(t) = r1_{vol}(t-1) + \Delta r1_{vol} \tag{3}$$

The density does not change, ie

$$\frac{r1_{mass}(t)}{r1_{vol}(t)} = \frac{r1_{mass}(t-1)}{r1_{vol}(t-1)} \tag{4}$$

3. Puck. Mass / volume of the "puck" could be modified as follows:

$$p_{mass}(t) = p_{mass}(t-1) + \Delta p_{mass} \tag{5}$$

$$p_{vol}(t) = p_{vol}(t-1) + \Delta p_{vol} \tag{6}$$

The density varies, ie

$$\frac{p_{mass}(t)}{p_{vol}(t)} \neq \frac{p_{mass}(t-1)}{p_{vol}(t-1)} \tag{7}$$

4. Player's 2 paddle. The mass / volume of the player's 2 paddle could be modified as follows:

$$r2_{mass}(t) = r2_{mass}(t-1) + \Delta r2_{mass} \tag{8}$$

$$r2_{vol}(t) = r2_{vol}(t-1) + \Delta r2_{vol} \tag{9}$$

The density does not change, ie

$$\frac{r2_{mass}(t)}{r2_{vol}(t)} = \frac{r2_{mass}(t-1)}{r2_{vol}(t-1)} \tag{10}$$

5. Player's 2 goal. The size of the goal of player 1 could be modified as follows:

$$goal2_{length}(t) = goal2_{length}(t-1) + \Delta goal2_{length} \tag{11}$$

That is, depending on the skill of the player's goal can be made larger or smaller.

6. Table's surface. The surface can be modified to make it rougher or smoother by affecting the air cushion which is provided by air hockey tables. The abstraction of this change is the modification of the friction constant.

$$\mu(t) = \mu(t-1) + \Delta\mu \tag{12}$$

7. Table configuration. That is, alter the length and width of the table.

$$table_{length}(t) = table_{length}(t-1) + \Delta table_{length} \tag{13}$$

$$table_{width}(t) = table_{width}(t-1) + \Delta table_{width} \tag{14}$$

Of the seven aforementioned configurable items, the table's surface and the puck's physical properties were selected. This decision was based on the feasibility to replicate the changes on a real air hockey table (like the one shown in figure 1).

The elements of a real table that can feasible be modified are:

- Table's surface. The friction between the table and the "puck" can be altered by changing the amount of air provided by the air pump.
- Player's 1 goal. The goal of player 1 can be modified with a device similar to that found in automatic sliding doors. That is, the goal can change it's length.
- Player's 2 goal. The goal of player 2 can also be modified with a device similar to that found in automatic sliding doors. That is, the goal can change it's length.

It is easier to change the size of the goals in a real table than it is to modify the properties of a real puck; never the less, the end result is similar: it make it easier or harder to score a goal.

3 Adaptive physics for video games

The ultimate goal of the work presented herein is that both players enjoy the game, regardless of the level of each player. That is, the game must be neither too complicated for the inexperienced player, because the player could get frustrated; or too simple for the skilled player, as the player could get bored.

To achieve this, the strategy described in [15] was used, where the following objectives have been pursued:

- The number of points per player must be balanced. That is, the performance of both players should be as close as possible so that both remain engaged with the game.
- The number of times each player hits the puck should not be too small nor too big. Few hits would make the game frustrating; whereas many hits would make the game boring.

To achieve these objectives, a set of parameters were selected.

3.1 Parameters

The configurable items (mentioned in section 2)

- Physical properties of the puck.
- Table surface.

To achieve the aforementioned objectives the following control variables were assigned.

- Goal difference. If the goal difference is large it implies that there is a significant skill difference between the players. The following formula indicates the case that if the goal difference is greater than a threshold the physics of the game must be adapted to make the game more entertaining.

$$abs(p1_{goals} - p2_{goals}) > goals\_treshold \tag{15}$$

- Number of hits to the puck before scoring a goal. If the number of hits to the puck is very large it implies that the game is boring; whereas if the number of hits the puck is small it means that the game is frustrating; in either case the physics of the game must adapt to make the game more entertaining.

$$p1_{hits} + p2_{hits} > max\_hits\_threshold \tag{16}$$

$$p1_{hits} + p2_{hits} < min\_hits\_threshold \tag{17}$$

If either threshold is exceeded, the physics of the game, must adapt to make the game more entertaining.

3.2 Artificial intelligence for generating adaptive physics.

Research on behaviour based robotics has shown that a character can perform complex and interesting behaviour [8][9] with the use of simple rules.

This research field was inspired by the work of Valentino Braitenberg [2]. Some of the robots that he proposed could be interpreted as more complex than the simple rules used to implement them. For example, he named the behaviour of some his proposed robots as: *love*, *fear* and *aggression*. Similar work is the horizontal robotic architecture put forward by Brooks [3]. This architecture, in turn, was the inspiration behind the BSA (Behavioural Synthesis Architecture); first used in cooperative robots [1]. This architecture was expanded to communicate emotions through artificial pheromones on virtual mammalian species [6]. This architecture was thence enhanced to affect the behaviour on groups (flocking) through an emotion based architecture [7]. Recently, a derivate of the above architecture was developed to provide opponents with personality profiles: aggressive, fearful, pathetic, bold, cautious, impulsive, predictable or analytical [10].

In this paper, we propose the use of a behaviour-based architecture to provide a video game (air hockey) with adaptive physics and it is described next:

3.3 Adaptive physics for the air hockey game.

Adaptive physics for the air hockey game is defined by the two parameters shown in figure 3 and described next.

1. The coefficient of friction for the table's surface material. This simulates the air cushion provided in the real air hockey table, like the one shown in figure 1. This parameter is affected by the following behaviours: avoid boredom and avoid frustration. These behaviours receive the stimulus: number of puck hits before scoring a goal.
2. The physical properties of the puck like volume and mass mentioned in section 2. This parameter is affected by the following behaviours: assist player 1 and assist player 2. These behaviours receive the stimulus: score difference.

4 Development

The air hockey game with adaptive physics is being developed in the programming language C++, the OGRE graphics library [19]; whereas, the physics engine used in the project is Bullet [5]. The models have been modeled in Blender [16]. Figure 4 shows the process to export a blender modeled geometry to the OGRE graphics library.

Figure 5 shows the rendering cycle for the adaptive physics game. The steps in the rendering cycle are:

- Read the input device
- Perform the video game logic (in this case for an air hockey game).
- Animation of the elements of the game.

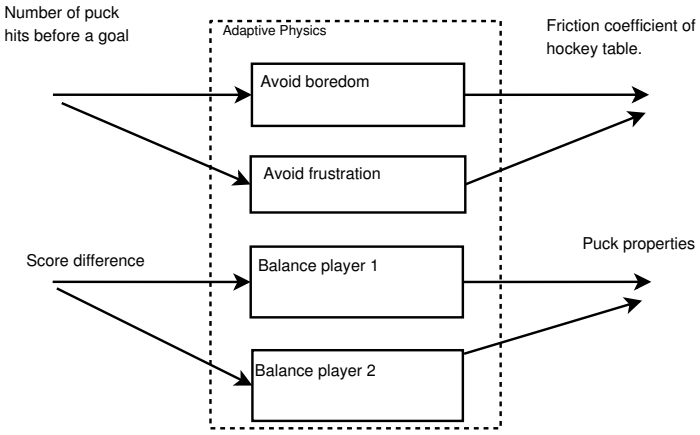


Fig. 3. Adaptive physics architecture.

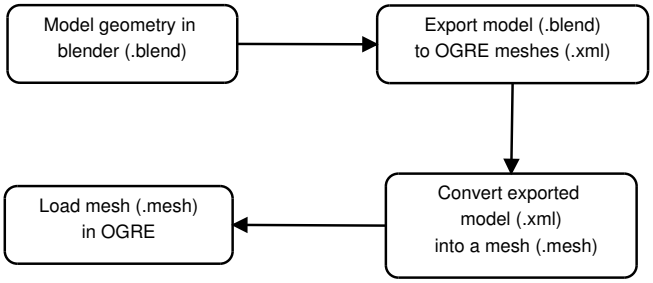


Fig. 4. Process from modeling in Blender to import a geometry in OGRE.

- Physics-based adaptive behaviours and described in the previous section.
- Run the physics engine (Bullet) simulation.
- Render the frame on the screen using OGRE.

The result is shown in figure 6.

5 Conclusions and Future Work.

The work in progress presented in this paper is an effort to improve the experience of players with different skill levels. This is accomplished by adjusting the physical parameters of some game elements, so that the game does not become boring for the skilled player or frustrating for the inexperienced player. For future work, experiments will be carried out to verify that the player's experience is significantly improved by adapting the game to the different player abilities. Furthermore, a real air hockey table will be developed and further experiments will be carried out therein. The results of both experiments will thence be cross-referenced.

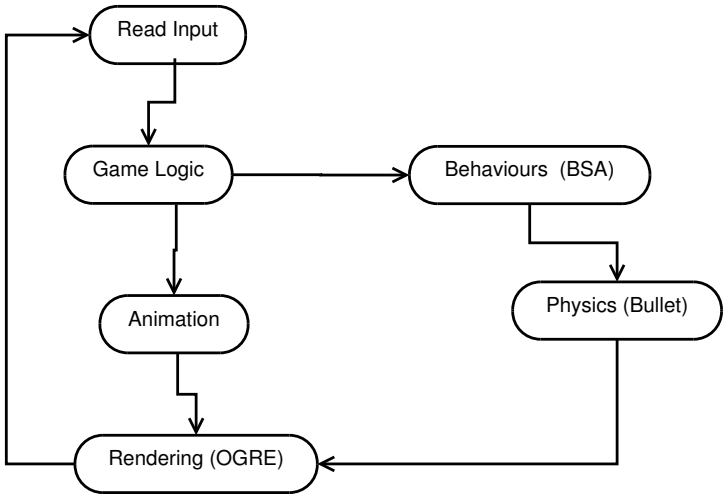


Fig. 5. Rendering cycle for the adaptive physics game.

Fig. 6. Screen taken from air hockey game with adaptive physics.

References

1. Barnes, D.P.: A behaviour synthesis architecture for cooperanting mobile robot. In: Gray, J.O., Caldwell, D.G. (eds.) Advanced Robotics and Intelligent Machines, pp. 295 – 314. IEE Control Engineering Series 51, United Kingdom (1996)

2. Braitenberg, V.: Vehicles: Experiments in Synthetic Psychology. Papermac, Oxford, United Kingdom (1984)
3. Brooks, R.A.: A robust layered control system for a mobile robot. IEEE Journal of Robotics and Automation 2(1), 14–23 (1986)
4. Catto, E.: Box 2d, http://www.box2d.org/
5. Coumans, E.: Bullet physics library, http://bulletphysics.org/wordpress/
6. Delgado-Mata, C., Aylett, R.: Communicating emotion in virtual environments through artificial scents. In: de Antonio, E., Aylett, R. (eds.) LNCS Intelligent Virtual Agents. Springer-Verlag, Madrid, Spain (2001)
7. Delgado-Mata, C., Ibanez, J., Aylett, R.: Let's run for it: Conspecific emotional flocking triggered via virtual pheromones. In: Butz, A., Krüger, A., Oliver, P. (eds.) LNCS Smart Graphics 2003. pp. 131–140. Springer-Verlag, Heidelberg, Germany (2003)
8. Delgado-Mata, C., Ibáñez-Martínez, J., Bee, S., Ruiz-Rodarte, R., Aylett, R.: On the use of virtual animals with artificial fear in virtual environments. New Generation Computing 25(2), 145–169 (2007)
9. Delgado-Mata, C., Ibáñez-Martínez, J., Gómez-Caballero, F., Guillén-Hernández, O.M.: Behavioural reactive agents to define personality traits in the videogame überpong. T. Edutainment 1, 135–149 (2008)
10. Delgado-Mata, C., Ruvalcaba-Manzano, R., Quezada-Patino, O., Gomez-Pimentel, D., Ibanez-Martinez, J.: Low cost video game technology to measure and improve motor skills in children. In: IEEE Africon. pp. 1–6. Nairobi, Kenya (September 2009)
11. Havok: Havok physics, http://www.havok.com/
12. Hunicke, R.: The case for dynamic difficulty adjustment in games. In: Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology. pp. 429–433. ACE '05, ACM, New York, NY, USA (2005), http://doi.acm.org/10.1145/1178477.1178573
13. Lam, D.: Tokamak physics engine, http://www.tokamakphysics.com/
14. Lembcke, S.: Chipmunk, http://code.google.com/p/chipmunk-physics/
15. Ibáñez Martínez, J., Delgado-Mata, C.: From competitive to social two-player videogames. In: Proceedings of the 2nd Workshop on Child, Computer and Interaction. pp. 18:1–18:5. WOCCI '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1640377.1640395
16. Mullen, T.: Mastering Blender. Sybex, USA (2009)
17. Nvidia: Nvidia physx, http://developer.nvidia.com/object/physx.html
18. Smith, R.: Open dynamics engine, http://www.ode.org/
19. Streeting, S.: Ogre: Open source 3d graphics engine, http://www.ogre3d.org/
20. Tijs, T., Brokken, D., Ijsselsteijn, W.: Creating an emotionally adaptive game. In: Proceedings of the 7th International Conference on Entertainment Computing. pp. 122–133. ICEC '08, Springer-Verlag, Berlin, Heidelberg (2009), http://dx.doi.org/10.1007/978-3-540-89222-9_14

Low Cost Technologies for Measuring the Psychomotor Coordination in Children from 7 to 11 Years

Ricardo Ruvalcaba Manzano[1], Carlos Delgado-Mata[1], Jesús Ibáñez[2]

[1] IPIT (Instituto Panamericano de Investigación Tecnológica)
Escuela de Ingeniería, Universidad Panamericana Campus Bonaterra,
Aguascalientes, México, C.P. 20290

[2] Grupo de Tecnologías Interactivas,
Departamento de Información y Tecnologías de Comunicación,
Universitat Pompeu Fabra,
Barcelona, España

ricardo.ruvalcabam@up.edu.mx, cdelgado@up.edu.mx

Abstract. Currently, there are psychometric instruments that are used to observe, evaluate and intervene on the children's motor coordination. These instruments are made with materials, that, occasionally rely on the use of technological elements. Therefore, the measurement process can be slow due to the amount of information that needs to be analyzed to get relevant results. The work presented herein describes a project to help improve the psychomotor coordination using a video game that unifies different multimodal technologies. Furthermore, this work assists in the measurement of the movements made by infants.

Keywords: children, motor coordination, video game, multimodal technologies.

1 Introduction

The main goal of the game development industry is focused on the entertainment of people. However recent research has shown that goals can go beyond entertainment. One such example, is the research, where videogame use was found to improve attention skills [1] . Other types of video games are focused on teaching or training of personnel in a specific area. Such is the case of the United States of America Army video game, America's Army, which was released in 2002 [2]. The game was used as pre-training and for future recruitment. Other games are used to assist the learning process [3].

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 47-55

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

1.1 Psychomotor coordination indicators

The environment evaluates and stimulates coordination, which is part of the dimensions of the psychological development [4]. Psychomotor coordination is defined as the combination of body movements that synchronizes and harmonizes all the internal forces that result in intended actions [5]. Table 1 shows coordination indicators, which are elements that help in the evaluation of coordinated acts. Each can be evaluated with a corresponding level of *"Adventures in the Jungle"*.

Psychomotor coordination indicators Description	Description
Efficacy	Fulfillment of the task.
Economy	Lower energy consumption.
Precision	Accuracy of the movement.
Force	Intensity with which the motion is carried.
Balance	Ability to control the body in space. (Static or dynamic).
Flexibilidad	Transition from static to dynamic and vice versa.
Peripheral Vision	Ability to see in an approximate range of 180 degrees.
Rhythm	Controlled or measured movements, sound or visual, usually produced by a sort of external stimuli.
Speed	Ability to quickly initiate and continue appropriate motor actions of short duration in response to a given signal.

Table 1. Psychomotor coordination indicators based on [6]

2 The Environment

The work presented in this document is focused on developing an assessment tool to measure psychomotor coordination in children aged 7 to 11 years. The research carried out is the construction, development and implementation of a tool to assist child experts (doctors, psychologists, physical therapists), so that the tool proposed herein might be helpful to meeting two significant needs:

- a) Assess and diagnose the motor coordination performance in children.
- b) Use the information stored in the database that this instrument collects for further research in the area of interest.

In order to solve this concern we developed: "The environment", which is a set of low-cost technology components, such as: A multi-touch screen, a *"Dance Dance*

Revolution" mat, a development platform (Framework) and a video game that unifies the above components "Adventures in The Jungle". Moreover, the environment was developed in order to facilitate the addition of new devices, it helps to capture the information of the actions taken by the child.

1. *Multi-touch screen:* It consists of a low-cost screen that captures finger movements, and whose primary function is to measure fine motor skills [7].
2. *Dance Dance Revolution:* It's a mat that captures the movements of the feet [8]. The aim of this device is to assess the child's gross motor activity.
3. *Development Platform (FrameWork):* Contains a set of routines needed for the development of applications [9] that require multimedia resources capable of responding to events generated by interfaces that work with the TUIO communication protocol.
4. *Video Game:* The first two technologies are unified by the videogame, which is composed of different levels. Each of them provides an indicator of psychomotor coordination. This information is stored in a database for later analysis.

2.1 Multi-touch screen

The multitouch screen is a device that is responsible for interacting with children finger movements [10]. Figure 1 shows an overview of the screen developed for the project described herein. The first component observed from left to right, is the computer that stores the application (video game) *"Adventures in the Jungle"*. This device sends a signal to the projector which then transmits the images to a mirror that is at the bottom. The mirror causes the image to scale up. The camera located on the side the mirror is responsible for capturing images or pointers generated by the user. As it can be seen, there is a final component a source of infrared light.

Having described touch screen components, it is worth mentioning that the disign of the screen cover is based on a jungle theme. The characteristics of the size and shape of the screen was based on average anthropological measures of children . This information was obtained from an investigation carried out on 1218 children assessed in the country of Costa Rica [11]. These measures may vary according to geographic location, diet and other factors. However, this information was used because of the anthropometric similarities amongst mexican and costarican nationals.

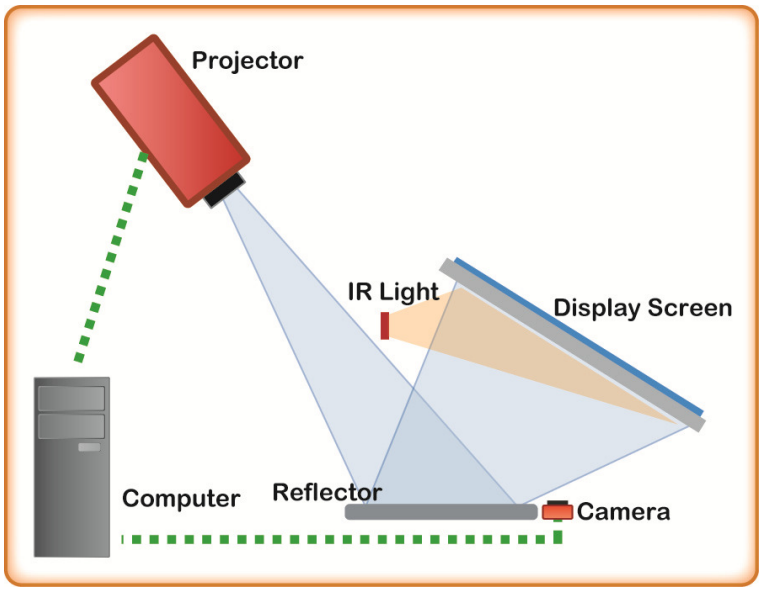


Fig. 1. General scheme of the screen based on [12]

Anthropological measures of children						
	Women			Men		
Measure	Percentile			Percentile		
	5	50	95	5	50	95
Height (m)	1,22	1,32	1,44	1,22	1,33	1,47
Weight (lb)	45,00	65,00	110,00	49,00	67,00	116,80
Popliteal height	31,50	34,63	37,92	30,46	34,90	38,30
Knee height	39,27	42,45	47,10	38,44	43,05	46,96
Thigh height	10,44	12,70	15,52	9,90	11,60	14,32
Scapula height	32,49	36,75	40,31	30,95	35,70	41,61
Elbow height	15,98	20,00	23,73	15,31	19,90	24,35
Long armrests	31,40	34,95	38,66	31,27	35,00	39,58
Distance buttock-popliteal	32,99	36,23	40,62	31,14	35,50	39,45
Wide hips	21,93	26,15	34,14	22,99	26,90	36,74
Shoulder Width	26,81	30,95	37,87	27,94	30,90	38,57
Buttock-knee distance	39,84	43,73	49,56	39,20	43,70	49,81

Table 2. Anthropological measures of children aged 9 based on [11]

Based on the information shown in the Table 2, a multitouch screen was designed and constructed. The resulting screen (shown in Figure 2) met the average measurements of infants (aged 7 to 11 years) that are the test subject for the project described in this document.

Fig. 2. Final design

2.2 Dance Dance Revolution

We made an adaptation for Dance Dance Revolution mat, thus so it was necessary to make sleeve that covered the original design of the table (Figure 3 shows a diagram of the sleeve adaptation). This wrapper was designed according to the theme of the environment, and it was later attached to the frame's base that is assembled with the multitouch screen.

Fig. 3. Diagram of the sleeve adaptation

The game was developed with Visual Studio 2010 (with C #) and XNA Game Studio 4.0. The application is composed of two main layers [13] (see Figure 4). The first layer has two libraries, the first and most important is AdventuresInTheJungle.Framework.dll, it contains a set of classes that form the skeleton of the game. The second library, that is in the same layer as the one described above, is responsible for managing access to the database (AdventuresInTheJungle.DataBase.dll). Finally, in the upper layer is the executable that contains the logic of the game.

Fig. 4. Sleeve coupling Video Game

2.3 Economy indicator level

This level is responsible for obtaining the number of moves that the child performs. It assesses the energy cost required to complete the test, which consists of finding a route from the place of departure (which is represented by a giraffe or a chimpanzee) to an end point (represented by a bush or a banana). The child can choose between different paths (See Figure 5), each of them are valid as long as next more is vertical or horizontal. Once the first part is completed, the scenarios are reversed, so that children is shown the scenario that the opponent has just completed. After the two activities are finished, the two children that were tested, should have completed the same tests.

When the level is completed, the information obtained is stored in a database for later analysis.

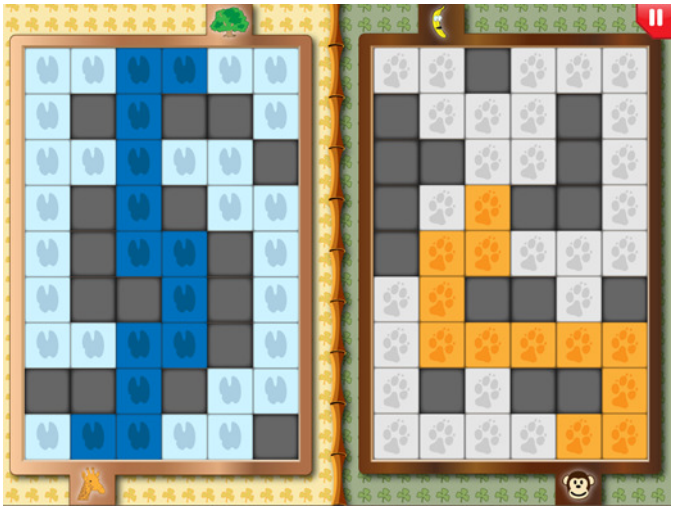


Fig. 5. Economy indicator level

3 Preliminary results and Conclusions

The environment was tested by a final run that evaluated a group of school children at Universidad Panamericana, campus Bonaterra in December 2010. In this version, it was observed that overall performance of the game, with the following results:

1. The setting met the required spatial dimensions to suit the physical characteristics of children.

2. Psychomotor coordination were encouraged

3. The content of the game scored measuring psychomotor coordination, specifically the area of the economy.

Children assessed, implemented correctly the level of economy, consisting of two panels that were made, each by different routes. The exercise involved traveling these roads, at the lowest possible energy cost. After completing the exercises, the system yielded information on the database environment. The results for this psychomotor coordination indicator are:

Name	Age	Test #1 (number of movements)	Test #2 (number of movements)	Test #1 (category)	Test #2 (category)
Carlos	9	15	16	High	Middle
Rowan	7	16	16	Middle	Low
Daniel	8	15	15	High	Middle

Table 3. Results obtained in the level of economy

These results are only preliminary. Later tests will be carried out with a significant group of children, those results are expected to be similar to the preliminary ones.

4 Future work

Future work will be focused on the development of standards covering the nine indicators of psychomotor coordination. This research can be used as an aid for interactive projects that integrate novel devices to help measure and encourage psychomotor coordination, one such device is Kinect, that can interpret an individual's body movements, without physically interacting with a device. These types of devices can help measure more sophisticated and complex movements.

5 References

1. G. West, S.: “Video game plating enhances practical attention skills,”. In : Journal of Vision, vol. 8 (2007)
2. Army, U.: “Americas army: The offcial army game,”. (Accessed 2002) Available at: http://www.americasarmy.com/
3. Natalia Padilla Zea, J.: Design of educational multiplayer videogames: A vision from collaborative learning. In : Advances in Engineering Software ACM, Oxford, UK, vol. 20 (2009)
4. Jeröme, L.: El movimiento en el desarrollo de la persona., Barcelona (1997)
5. Kiphard, E.: Insuficiencias de movimiento y coordinación en la Escuela Primaria., Buenos Aires (1976)
6. Vázquez, C.: MOTRICIDAD fundamentos y aplicaciones., Madrid, pp.355-361 (2004)
7. Delgado-Mata, C., Ruvalcaba-Manzano, R., Quezada-Patino, , Gómez-Pimentel, D., Ibáñez Martínez, J.: Low Cost Video Game Technology to

Measure and Improve Motor Skills in Children. In : AFRICON, Nairobi,Kenya, p.6 (2009)

8. Höysniemi, J.: International Survey on the Dance Dance. In : ACM Computers in Entertainment, Tampere, Finland, vol. 4 (2006)
9. Nuno Flores, A.: Patterns for Understanding Frameworks. In : Proceedings of the 15th ACM Conference on Pattern Languages of Programs, New York, NY, USA, p.11 (2008)
10. Han, J.: Low-cost multi-touch sensing through frustrated total internal reflection. In : UIST '05 Proceedings of the 18th annual ACM symposium on User interface software and technology, New York, NY, USA (2005)
11. Madriz Quirós, C., Ramírez Coretti, A., Serrano Montero, R.: Estudio antropométrico para el diseño de mobiliario para niños de edad escolar en Costa Rica. Tecnología en Marcha, 12 (2008)
12. Group, T.: TUIO. (Accessed January 12, 2011) Available at: http://www.tuio.org/
13. Shan, T. C.: Solution Architecture for N-Tier Applications. In : Services Computing, 2006. SCC '06. IEEE International Conference on, Chicago, IL (2006)

Development Process for Educational Games

Andrea Herrera, Lucina Torres*, Inti Reyes, William Zurita,
Venustiano Soancatl, María-Luisa Cruz, Luis Huerta, and
Antonio León.

Universidad del Istmo, Computer Science,
Cd. Universitaria s/n Ciudad Ixtepec, 70110 Oaxaca, México
{aismenee,inti,ezurita,
venus,mlcruz,luisdh2,leonborges}@bianni.edu.mx,{luz83_83}@hotmail.com*
http://www.unistmo.edu.mx

Abstract. Many students find mathematics a difficult subject. In the context of Mexico, the failure rate in mathematics has been very high. This paper discusses the creation of an educational computer-game system using a virtual environment. The objective of this application is to motivate and encourage students to practice mathematics. The videogame environment aims at making fun the learning process. Using User Centered Design, this game has to satisfy different students who may have very different likes. In this paper we present a first approach of a propposed model for this videogame development processes, considering that videogame development consists of much more than developing only software. This process is based on the *Game Unified Process* methodology and *User-Centered Design*. The processes are organized in phases and composed by more specific activities and also it uses artifacts to improve software development.

Keywords: Game-based learning, virtual environment, component based design, problem solving, development process.

1 Introduction

Last years, the use of computers has been essential on daily life as well as in education. Personal computers are being introduced on many levels of education: primary, secondary and high school [9]. Nowadays, the use of technology, especially computer games, provides a propitious environment for learning process, which takes place in a virtual world. Players create their own knowledge through the interaction with the virtual environment [5]. This kind of tools, used actually for teaching, are called: Educative Software (ES). The term Educative Software is denominated to a group of informatic resources confined to teaching and autonomous learning, and is used on teaching-learning process context [16]. The ES is created for the specific purpose of being used as a didactic media that facilitates teaching-learning processes. ES includes a wide variety of tools and resources, from learning oriented programs to complete operating systems. The use of technology makes easy the learning process by giving high interactivity, chance to practice many times, fault tolerance, and increase imagination and achievement sense [3], moreover, they become more popular since they are an important part of children and youth' lifestyle [17, 18]. Educative Videogames create fun, motivation

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 56-66

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

and interactive learning virtual environments, so they are a resource that gets students attention and invites to use it, providing teaching-learning processes, and attacking one of the biggest problems of traditional teaching: keeping motivated and engaged students to continue learning and applying knowledge [5].

This work consists on the design and development of an educative videogame called "Museo Virtual de Geometría" for teaching-learning process for third grade of secondary school. With the objective of answering to the noticed problematic on national and international studies: Exámenes de la Calidad y el Logro Educativos (EXCALE), Evaluación Nacional de Logro Académico en Centros Escolares (ENLACE), and Programme for International Student Assessment (PISA) which have demonstrated the educative lag presented by mathematics students at a world-wide level [7, 8]. Actually the mathematics problems solution is considered the most important part of mathematics teaching, and it requires the students to have necessary abilities to solve them. However, there are many troubles the students have to face when they try to solve problems; some of the troubles are: the students don't understand the problem due to plenty of phrases, lack of linguistics keys that help students to select appropriate arithmetic operations, absence of practice applications, poor lecture and understanding skills, lack of motivation, limited mathematic problems solution experience, hardness on converting textual problems into appropriate numeric formats, lack of skills, tools and resources access needed, to mention some of them.

By the mentioned above, it's evident the need of the students to improve their mathematics performance. The proposed educative videogame motivates the students to learn methematics, through fun, responding to world continuous social transformations for the teaching-learning process. Thus, educative videogames have an enormous potential to improve mathematics teaching; also, these videogames can be used to reinforce student' study and help on mathematical concepts development [10].

Videogame development implies a complex duty that requires specialized knowledge on areas like graphics processing, programming, animation, sound, etc [14]. This causes the development to be interdisciplinary and parallel. Moreover, there is little research in educational videogames development field [4]; is well known that developers implement their methods, but their use is not organized, so it takes to a less structured software developing for this kind of applications. All this makes us to require software engineering techniques and tools for the design and development. Actually, the Software Engineering process on videogames development is not defined clearly, which makes hard the practices and processes development less reliable on this field [2]. A videogame could fail because of poor management of the entire process implied in the development elaboration, so videogames developers need very different skills, compared to previous designers, to lead this process with caution and efficiency [6], according to the needs of society.

2 Development Process of "Museo Virtual de Geometría"

In most current game projects, the development process is the area that needs the most attention. In many recent projects, it can be seen that hasn't been used a real develoment model. With the integration of software engineering for developing the game, we

get some benefits like: everyone involved in the game development reduces his effort, becasuse it allows software developer to identify and apply solid practices. Thus, the system interdisciplinary development is taken into account and propitiates a good team communication of development by the established guidelines and the making of the project documentation on each phase. The goal of Geometry Virtual Museum (MVG) Project is to develop a software tool to support the geometry teaching on secondary school level.

For design and development of the MVG it was proposed a software development process according (See Table (1)) to the educative videogame needs, incorporating the most important characteristics and elements of User Based Design (UBD) and Game Unified Process (GUP). The proposed development process defines its phases, activities and artifacts, by mentioning *what to use?*, *how to use it?* and *when to use it?* The phases of the development process are: Planning, Requirement specification, Game Design, Design Development, Test and Relase.

Table (1) shows four of the seven phases of the process with the activities for phase and artifacts. As for the MVG software development, it was used the approach "Component Based Design". The component based design basic steps are: requirements gathering, component partitioning, inner design, component evolution and component interconnection. Moreover, the component based design lets the reuse of code "pieces" pre-elaborated to make a variety of tasks, leading to different benefits like the quality improvements, development cycle reduction and a greater return of investment. Among the advantages of using this design approach we found: software reuse, tests simplification, system maintenance simplification, better quality, and shorter development cycles. With the purpose of establishing development team tasks, the MVG has divided into units, as show on Fig. (1).

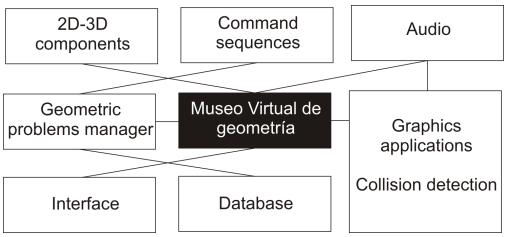


Fig. 1. Geometry Virtual Museum Units.

Based on this division, there was specified the set of activities to do, to guide the developers during the whole development cycle. The MVG planning started with requirements elaboration that defines the project. The planning includes the estimation of aspects related with the products and tasks to develop, the determination of the necessary resources to build the system, and the project risks identification and analysis. The project planning is checked according to project advance to deal with the requirements and the obligations changes, inexact estimations, corrective actions, and process changes. Fig. (2) presents project plan designed for planning and requirement specification phases.

Once the project planning was developed, it was chosen the development strategy, to have a guide for developing the product. This implies the identification of the software components to be reused and also includes decisions about the architecture characteristics. The project work plan was organized around the product and its software components.

Table 1. Identified activities for the development process of the MVG.

PHASE	DESCRIPTION
Planning	Guide to stablish and mantain the plans that define the project activities
Activities	To develop a project plan Developing time management
Artifacts	Project plan
Requirement specification	Gives neccesary activities for requirements developing and management. Education, analysis, needings validation, expectatives and game restrictions
Activities	Analisys and understanding of learning complexity Identifying interface requirements Requirements analysis and validation
Artifacts	Developing strategy, user case diagram, sequence diagram
Game Design	Involves decisions about game characteristics
Activities	To choose the platform and tools to use Possible graphics from user perspective Graphics and art styles to use
Game design	Refers to a collection of operations, classes, and components that embody the functionality of system
Activities	Game script elaboration Didactic goals identification Characters choice Navigation diagrams definition

The activities structure allows a first sight over the following elements: Possible risks by activity and mitigation tasks, Identification of tasks for user value products and Tasks for additional plan development (such as Configuration Management and Process Quality Assurance, Product and Validation). Fig.(3) shows the software components on which the project is divided and its functionality. For development team it was important to do this step because it lets us to identify the component limits to implement.

The MVG was made into an iterative process, so, during the whole project life the product needings are identified and understood, which may causes requirements to change. The designed development process uses the User Case Diagrams (UCD) to capture the system functions and the entities outside the system (actors). With the UCD, the system functional requirements specification is obtained and the interactions between the system and external factors are shown. The use case is a way to specify:

Table 2. Identified activities for the development process of the MVG.

Template for Project Planning (PROJ)
Software system to reinforce geometry learning-teaching process, on medium basic level on Veracruz Date:September 2010

Weekly data	Planned	Actual
Project's hours for week	_____	_____
Project's hours for cycle to current date	_____	_____
Earned value by week	_____	_____
Earned value by phase	_____	_____

Data for Team members	Planned hours	Actual hours	Planned value	Earned value
Team leader	_____	_____	_____	_____
Project manager	_____	_____	_____	_____
Planning leader	_____	_____	_____	_____
Quality leader	_____	_____	_____	_____
Support leader	_____	_____	_____	_____
Total	_____	_____	_____	_____

what the system has to do and to guide the construction of the MVG. On Fig. (2) is presented the MVG User Case Diagram general diagram. To the analysis and understanding of the difficulty of learning, it is proposed the methodology presented on Fig. (3) for geometric problem teaching-learning. Thus, on the information treatment axis there were designed activities to grow on the students, the ability to solve geometric problems through analysis and selection of information established into a step set, as shown on Fig. (4).

The student obtains information to solve the problem through the steps; if he is not able to solve the step, he will get an advice as feedback. The proposed methodology for developing geometric problems is integrated by eight stages, with the objective of guiding the student to find the problem solution. So, the student skill to solve geometry problems is increased trough analysis and information selection proposed into steps, as shown on the next example (see Fig. (4)). The student, trough the steps, gathers information to solve the proposed problem and, if he is not able to complete the step, he will get a recommendation or tip to solve it.

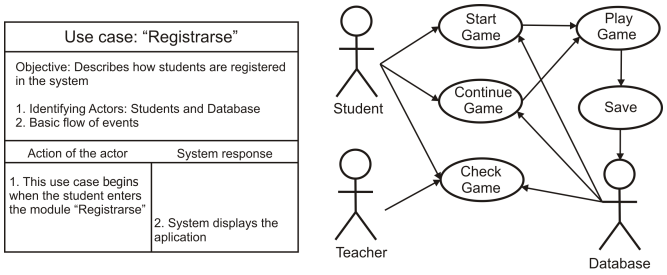


Fig. 2. "Museo Virtual de Geometría" General Diagram.

Table 3. Identified activities for the development process of the MVG.

Project name:	Software system to reinforce geometry learning-teaching process, on medium-basic level on Veracruz
Date:	September 2010

General description of functionality.	
The ”Museo Virtual de Geometría” is an educational software tool designed for high school geometry teaching-learning. The MVG is a possible solution that can assists students with mathematics and also, it is a tool for to motivate students by fun and interesting games that encourage them to play and holds their attention.	
Component or system modules	
Store statistics	Solved problems are stored into a local database This component is responsible of: Error count: the component will count the noumber of tries when solving a problem Solution time: to solve a problem, it is, a limit time given by this component Points (score): each solved problem has asigned an amount of points which will be registered in this component, to know the student performance
Database	Manage the Museum information, so, is in charge of: Locate and obtain object characteristics that will be used in the Virtual Museum Read related problems with an object, as well as properties for specific object related information storage Read steps related to a problem

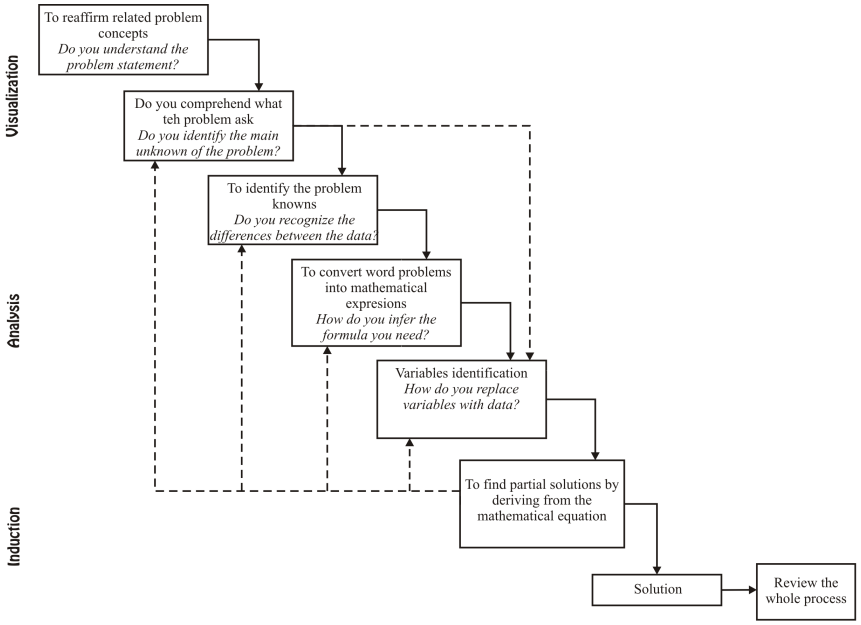
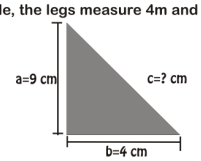


Fig. 3. Steps for the analysis and understanding of geometry learning difficulty.

The methodology starts with the reaffirmation of problem associated concepts. The objective of this point is that the student understands what the problem is asking for. On Fig. (4), steps one, two and three fulfill with this stage (stage one of visualization phase). On a particular way and based on the shown problem, it is clearly visible that the concepts the student must have to solve the problem are associated with the concepts of: triangle, right triangle, legs and hypotenuse. Once the student can answer to the question of "Do you understand everything said on the problem?" the methodology stimulates the student to understand what the problem is asking for and so, he can answer the question "Do you know what the goal is?" Step four and five from Fig. (4) belong to the second stage of the visualization phase. The last visualization phase helps the student to identify the data he has to solve the problem; step six is an example of this. The next phase proposed by the methodology is the analysis phase, integrated by the stages of "converting verbal expressions to mathematical ones" and "identifying the problem variables". With this, the student can answer "How will you express the mathematical formula to use?" and "What numeric data must be used?" On step seven and eight the first phase of analysis is fulfilled; here, the formula of the problem variable is found. The last stage of the analysis phase is shown on step nine and ten, where we identify the values of the variables to be used. The last stage proposed by the methodology is the induction, with the aim of consolidating knowledge produced trough the steps.

On the shown problem, this stage phase is represented with the steps eleven, twelve and thirteen; with this steps the student find partial solutions that makes easier the answer for the problem. So, the methodology concludes with the solution stage, where the student answers the initial proposed problem. Step fourteen exemplifies the way to find the problem solution. It is important to remark that the methodology is iterative, so it let

Problem. In a right triangle, the legs measure 4m and 9m respectively. Calculate the hypotenuse.



Step 1. What name receives the triangle which has one angle of 90° and its two other angles are acute?

☐ Obtuse ☐ Scalene ☐ Isosceles ☐ Right

Step 2. What is a leg?

☐ One of the two shortest sides ☐ Lower side ☐ Longest side ☐ Left side

Step 3. What is the hypotenuse?

☐ One of the two shortest sides ☐ Lower side ☐ Longest side ☐ Left side

Step 4. What do we need to obtain?

☐ Hypotenuse ☐ Height ☐ Leg ☐ Bisector

Step 5. Which Theorem you must know to solve the problem?

☐ Pythagorean Theorem ☐ Tales Theorem ☐ Venus Theorem ☐ Apu´s Theorem

Step 6. Taking into account the data about the problem's triangle, which data do you know to solve it? (Focusing on the triangle problem data, which data do you know to solve it?)

☐ The legs value ☐ The hypotenuse ☐ The height ☐ The legs and hypotenuse values

Step 7. To applying the Pytagorean Theorem, which relation you must use if c is the hypotenuse and a and b the legs?

☐ $c=\sqrt{a^2+b^2}$ ☐ $c=\sqrt{a^3+b^3}$ ☐ $c=\sqrt{abc^2+b^2}$ ☐ $c=\sqrt{a^2+ab^2}$

Step 8. To obtain the leg a value, I must use the next phrase:

☐ $a=\sqrt{b^2+c^2}$ ☐ $a=\sqrt{b^3+c^3}$ ☐ $a=\sqrt{abc^2+b^2}$ ☐ $a=\sqrt{a^2+ab^2}$

Step 9. How much is leg a?

☐ 9 ☐ 6 ☐ 5 ☐ 4

Step 10. How much is leg b?

☐ 4 ☐ 16 ☐ 5 ☐ 10

Step 11. How much is the squared leg a?

☐ 81 ☐ 20 ☐ 10 ☐ 100

Step 12. How much is the squared leg b?

☐ 16 ☐ 20 ☐ 10 ☐ 100

Step 13. How much is the sum of legs a=16 and b=81?

☐ 97 ☐ 30 ☐ 5 ☐ 100

Step 14. How much is the hypotenuse from the problem's triangle?

☐ $c=\sqrt{16+81}$ ☐ $c=\sqrt{10+15}$ ☐ $c=\sqrt{100+10}$ ☐ $c=\sqrt{5+5}$

Fig. 4. Steps for the analysis and understanding of geometry problem solving.

us to improve the problem development and its elaboration to be flexible and adaptable to the problem needings. On Fig. (5) we can see a screenshot showing an example of a problem using the proposed metodology.

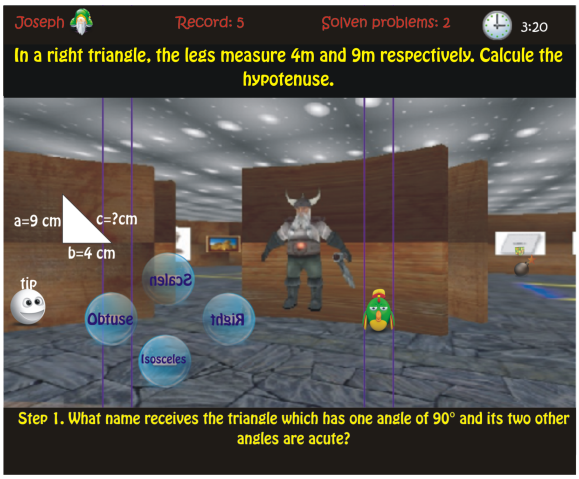


Fig. 5. Screenshot showing how is presented a problem into the game.

The "Museo Virtual de Geometría" was designed in this way because videogames with labyrinth environments are very popular between teenagers. For the design we consider the following points: Interface, Rules, Levels, GamePlay and Features. The MVG has been created using Irrlicht (an open source 3D engine written in C++). In MVG there are three difficulty levels: basic, intermediate and advanced. To obtain a right scenarios design it is necessary to take into account the likes and preferences for colors and textures of the final users. To achieve this, there was designed and added different scenarios (shown on Fig. 6) and contextual surveys applied on January 2011 to students from the secondary school: Secundaria Técnica No. 27 , in Ciudad Ixtepec, to know likeness and preferences of the final user.

3 Conclusions and Future Work

The structuration of a developing process for game design is a very important fact to coordinate the developing team and minimize time and effort.

The MVG is very interesting among secondary school students and has the potential to relax evaluations, introduce chaos, create conditions to improve sharing of tacit knowledge, and increase the bonds between organizational members. It can support the platform for knowledge creation and organizational learning. With the MVG, the students learned and reaffirmed mathematic concepts by playing. Answering the problem by steps, the students felt that they dominate a complex game, because they solve a problem using steps that manage its complexity. The application was an interesting tool for students and teachers from secondary schools, as they have fun and learn through

Fig. 6. Different MVG scenarios.

activities based on curricula and methodologies for teaching proposed by the Secretaría de Educación Pública. It is convenient to conduct a more extensive usability study that allows the feedback to improve the application. On the information treatment, we consider that is suitable to feedback with problems proposed by teachers of the subject, so, we require to make an interface on which the teacher can incorporate problems of the issues they consider are complicated to learn.

Acknowledgments

We want to give many thanks to CONACYT, FOMIX-VERACRUZ and UNISTMO for supporting this work which is part of the project 95656.

References

1. Morsi, R., Jackson, E.: Playing and Learning? Educational Gaming for Engineering Education. 37 th ASEE/IEEE Frontiers in Education Conference (2007).
2. Callele, D., Neufeld, E., Schneider, K.: Requirements Engineering and the Creative Process in the Video Game Industry. Proceedings of the 2005 13th IEEE International Conference on Requirements Engineering (RE'05) (2005).
3. Gee J. P., What Video Games Have to Teach us About Learning and Literacy. Palgrave Macmillan (2007).
4. Hun, S., Hee, G., Hoon, H., Heon, D., Yul, S.: An Empirical Model of the Game Software Development Processes. Proceedings of the Fourth International Conference on Software Engineering Research, Management and Applications (SERA'06) (2006).
5. Taran, G.: Using Games in Software Engineering Education to Teach Risk Management. 20th Conference on Software Engineering Education & Training (CSEET'07) (2007).
6. Fullerton, T.: Play-Centric Games Education. IEEE Computer Society (2006).
7. Soancatl, V., Torres, L., Cruz, M., León, J., Martínez, C.: Un enfoque Para la Enseñ anza-Aprendizaje de Solución Problemas de Matemáticas Basado en Preguntas. 11 congreso Internacional y 14o Nacional de Material Didactico inovador, nuevas tecnologías educativas (2010).

8. Cruz, M., León, J., Soancatl, V., Torres, L.: Aprendizaje basado en juegos: un enfoque alternativo para el desarrollo del software didáctico, XII Congreso Nacional de Ingenieria Electromecanica y de Sistemas (2010).
9. Bandoh, H., Nemoto, H., Sawada, S., Indurkhya, B., Nakagawa, M.: Development of educational software for whiteboard environment in a classroom, Advanced Learning Technologies (2000).
10. Moselhy, H.: Support Mathematical Instruction in Web-based Learning System using Object-Oriented Approach. International Conference on Advanced Computer Theory and Engineering (2008).
11. Wastiau, P., Kearney, C., Van den Berghe, W.: How are digital games used in school? Complete results of the study. Final report. Brussel : European Schoolnet (2009).
12. Futurelab. Teaching with Game. Final report. `http://www.futurelab.org.uk/resources/docng_with_games/TWG_report.pdf` (2007).
13. Nummenmaa, T., Berki, E., Mikkonen, T.: Exploring Games as Formal Models. 2009 Fourth South-East European Workshop on Formal Methods (2009).
14. Moreno, P., Martnez, I.,Sierra, J., Fernndez, B.: A Content-Centric Development Process Model. IEEE Computer Society (2008).
15. Siang, A.C.; Radha Krishna Rao; , Theories of learning: a computer game perspective, Multimedia Software Engineering (2003).
16. Haiguang Fang; , Modeling and Analysis for Educational Software Quality Hierarchy Triangle, Web-based Learning (2008).
17. Jianhua Ren; Shuang Liu; Hui Ding; Yeli Deng; , Design and implementation of children's art educational software "Happy baby drawing board". Information Management and Engineering (ICIME) (2010).
18. Graven, Olaf Hallan; MacKinnon, Lachlan; , Exploitation of games and virtual environments for e-learning, Information Technology Based Higher Education and Training (2006).

Is M-marketing an Accessible Technology for Consumers with Visual Disabilities? – A Preliminary Study

Rafael Martínez-Peláez[1], Alejandro Bautista Mendoza[1], Daniel Cortes Morales[1], Cristina Satizábal[2], Oswaldo Ávila Barrón[1], Darío Barragán López[1]

[1] Universidad de la Sierra Sur, Licenciatura en Informática, C/ Guillermo Rojas Mijangos S/N, Ciudad Universitaria, Miahuatlán de Porfirio Díaz, 70800, Oaxaca, México
{rpelaez, li2005060048, li2004060023, oavila, dbarragan}@unsis.edu.mx

[2] Universidad de la Sierra Sur, División de Estudios de Postgrado, C/ Guillermo Rojas Mijangos S/N, Ciudad Universitaria, Miahuatlán de Porfirio Díaz, 70800, Oaxaca, México
isatizabal@unsis.edu.mx

Abstract. Marketing is becoming more and more important in our daily life. However, the communication channel and technology used is not accessible, in many cases, for consumers with visual disabilities. In this paper, we evaluate the current state of the art, which includes t-marketing, e-marketing and m-marketing, with the goal of reviewing suitable techniques and tools for consumers with visual disabilities. Our study demonstrates that m-marketing provides personalization, interactivity and comfort to consumers with visual disabilities, making it more accessible and flexible than e-marketing. Moreover, we found that m-marketing is less intrusive than t-marketing and e-marketing in maintaining consumers' privacy.

1 Introduction

Recently, we have lived in a boom of wireless technologies which are applied in different areas of our daily life, creating mobile services [1], [2], [3], [4], [5]. Mobile services refer to any electronic transaction using mobile devices and wireless technologies, where the end user is the beneficiary of the service. In this new scenario, the end user can get access to mobile services whenever and wherever she wants by means of her private and personal mobile device. For that reason, mobile services are changing the way in which people interact among each other.

M-marketing (mobile-marketing) is another successful mobile service which provides new options to consumers and organizations [6], [7], [8]. We define m-marketing as the process of promotion of ideas, goods and services using wireless technologies and mobile devices with the goal of creating consumer satisfaction.

Unfortunately, m-marketing has inherited problems from e-marketing (electronic-marketing), like spam and the absence of trust [9], which must be resolved before this service grows to replace t-marketing (traditional-marketing). Another aspect which must be resolved is the accessibility to promotions of special sales and offers to

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 67-76

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

consumers with visual disabilities – people who are blind, who have little or no functional vision, and people who have low vision - in order to contribute to create a fair society.

This study aims to analyze the technological environment perception in relation to m-marketing to focus on people with visual disabilities. According to [10] people with visual disabilities in the Worldwide exceed 314 million and 45 million of these are blind. In Mexico people with visual disabilities are around 500,000 [11].

The M-marketing focus on people with visual disabilities is interesting to study because even though, on the one hand, it has had a relatively developed technology environment – communication and mobile devices – for many years; on the other hand, its application to consumers with visual disabilities is recent. Highlighting the challenges in introducing new technology in this field can provide useful insights that can relate to other contexts.

The research methodology consists of literature reviews of theoretical material related to marketing evolution – t-marketing, e-marketing and m-marketing – which focuses on consumers with visual disabilities and human-computer interaction (HCI). In addition, this study tries to understand the opinion and perception of a consumer with visual disabilities about the correct use and application of techniques and tools in marketing.

This paper is organized as follows. In section 2, we describe the techniques and tools which can be used by organizations in their marketing campaign. Section 3 describes the research method which includes human-computer interaction and a marketing focus on consumers with visual disabilities. We present our results in section 4. In section 5, we give conclusions and plans for future work.

2 Marketing

According to [6], marketing is defined as the process of planning and executing the conception, pricing, promotion, and distribution of goods, services and ideas to create exchanges that can satisfy individual and organizational goals. In this way, marketing for consumers with visual disabilities is an alternative field which requires special attention. The use of new technologies and approaches in marketing management focusing on consumers with visual disabilities brings opportunities and challenges for academic and industrial communities.

In this section, we describe how technology can be applied with a marketing management focus on consumers with visual disabilities through audio-media and tactile-presentation, making human-computer interaction accessible for anyone.

2.1 T-marketing

T-marketing is the most common way to promote goods or services via print-format, television-advertising and radio-advertising [12]. This type of marketing is very well known by organizations and consumers. In order to capture the attention of consumers with visual disabilities, academic and industrial communities have designed the following options:

2.1.1 Audio-media

Audio gives to consumers with visual disabilities information about the promotion of ideas, goods and services that they might not be able to see. We found the following options:

- Audio file with a voice talking about an idea, good or service [13], [14].
- Speech system can be used to – automatically – read aloud text [13].

2.1.2 Tactile-presentation

Organizations can use Braille codes [15] for representing documents. These codes consist of characters of either six or eight dots in columns of three or four dots. By means of Braille codes, it is possible to translate material such as promotion or special sales into sequential strings of similar Braille characters. Marketing focus on consumers with visual disabilities can use the following techniques and tools:

- Microcapsule paper is a special paper onto which millions of thermally-foamed microcapsules have been uniformly coated [16]. These microcapsules of wheat-flour will instantly expand to hundreds of times the size of the original volume upon absorbing the energy of light or heat. By means of this technology any material such as a drawing, map or photograph, whether hand-written or printed, can be 3D.
- Flexi paper must be passed through a machine called "tactile image enhancer" which creates a sensory image [16], [17]. By means of the flexi paper and the machine, it is possible raise any image on the paper.
- Tooling consists of impressing points on the reverse side of Braille paper or foil using special tools [16]. By means of this technique if any is necessary to prepare the Braille impression on the pages before adding the raised areas.

In this way, blind users can use their hands to identify the message printed on the paper. This means that blind people used and enhanced their tactile sense.

2.2 E-marketing

E-marketing appeared thanks to the constant growth of internet users. E-marketing is the process by which a product or service is promoted via the internet and electronic media. E-marketing is a view in where the organization's website creates relationships with internet users creating direct access to personally relevant news. In this perspective, organization's Web-site is an autonomous system which not only provides information, but also remembers consumer's information about their interests. According to [18], e-marketing increases consumer satisfaction resulting in consumer loyalty but how we can give access to promotions of special sales and offers via internet to consumers with visual disabilities? In order to answer the question, we have reviewed the literature and we found the following techniques and tools which can be used to promote ideas, goods and services focused on consumers with visual disabilities [19], [20], [21]:

- Screen-reader software speaks all the text information which comes on the screen as well as the text which is typed on the keyboard. Another advantage of this software is that it provides to blind users the support to identify the content of a

web page, such as windows, buttons and media player controls, giving as a result a better interaction between people with visual disabilities and computers.

- Braille display is hardware which presents the information on a Braille code. These displays are made of crystalline material points, which through electrical pulses, are able to raise their level on the display, so that characters are represented in Braille. The information about the content to display is sent from the operating system to screen-reader software, which through a series of algorithms translate the information into Braille to be represented on the Braille display. By means of, consumers with visual disabilities can read with their fingers.
- Multimedia presentation includes the combination of text, graphics, video, and audio which can be used for consumers with visual disabilities.

2.3 M-marketing

M-marketing is the next step in the evolution of marketing. It appears thanks to the adoption of mobile devices in developing and non-developing countries, advances in wireless technology, higher connection speed, larger screens and new functionalities in mobile devices.

In this case, organizations have several channels to promote ideas, goods and services because m-marketing requires wireless technology to establish a contact with consumers. Wireless technology can be classified into groups: short-range and long-range [22], [23], [24].

Short-range wireless communication includes the following:

- Bluetooth is low-power consumption and robustness. Bluetooth consists of an radio-frequency transceiver, base-band and protocol stack and offers services that enable the connection of devices and the exchange of a variety of data (audio, video, text, and images) between them.
- Near field communication (NFC) is a high frequency wireless communication technology which enables the exchange data over about a 10cm distance. NFC combines the interface of a smartcard and a reader into a single device.
- Radio frequency identification (RFID) is a wireless technology which detects and identifies objects using radio signals. RFID system comprises two basic components – tags and interrogators –. Tag has a unique identification number (ID) and memory to store the data (e.g. manufacturer name) that enables the entire system to identify items. On the other hand, the interrogators can read and/or write data from/to the tag.
- Wireless-USB is a high-bandwidth wireless radio communication protocol. The goal of this technology is intended to replace the cables from USB based PC peripherals.

Long-range wireless communication includes the following:

- Wi-Fi has a data rate up to 11 Mbps within a range between 60m to 100m. It is standardized by IEEE 802.11b (2.4 GHz).
- WiMax was designed as a wireless alternative to low-cost broadband via standard interface to the public networks operating over around 50Km at 70Mbps.
- 3G or 4G offer high speed wireless connectivity.

Due to the mobile device's popularizations, companies have developed devices with better features (e.g. different wireless communication technologies) and more accessible for anyone; so that, the inter-connection among different devices can be automatic, making easier the use of mobile devices for people with visual disabilities.

By means of wireless communication and mobile devices, consumers with visual disabilities can access promotions via screen reader software, voice recognition and multimedia [25], [26].

3 Research Method

After a review of many techniques and tools proposed in the literature, we talked with the only student who is blind in the *Universidad de la Sierra Sur*. We wanted to know his opinion and perspective about the advantages and disadvantages of the techniques and tools described in section 2.

For this study we considered HCI concepts [27], [28], [29] that provide understanding and knowledge of the way in which consumers perceive and process information, so it is possible to improve the relationship between consumers and computer systems, and the way in which they perform tasks.

Mobile devices and informatics systems being the medium by which m-marketing offers advertising, those systems must adapt to blind users in terms of accessibility, comfort, convenience and interactivity, which fall into a discipline called Human-Computer Interaction (HCI). That's why we considered in the interview the following concepts: HCI and marketing.

3.1 Human-Computer Interaction

- Accessibility involves two points: 1) how users with disabilities (e.g. visual) access electronic information and 2) how content is presented to people with disabilities. The challenges affect both sides – consumers and organizations. On one hand, consumers need to identify the best technique or tool that can provide the most convenience access to information. On the other hand, organizations must identify those obstacles which prevent accessibility to information in order to design the best tools and techniques for people with disabilities.
- Comfort: represents the satisfaction from people with disabilities with the technology and tools used to give access to information and/or services.
- Convenience: evaluates the quality of being suitable technology and tools for people with disabilities.
- Interactivity: indicates a measure of the effectiveness with which people with disabilities can achieve specified goals by means of technology or tools defined.

3.2 Marketing

- Convenience: evaluates the ease with which blind people can access advertising.

- Intrusiveness: represents the aggressiveness of advertising practices which interrupt the activities of consumers.
- Localization: presents promotions based on physical location of the consumer.
- Personalization: promotion of ideas, goods and services oriented to a specific group of consumers.

4 Results

Table 1 shows our preliminary results in the first phase of the research. We compared the techniques and tools used in t-marketing, e-marketing and m-marketing focus on consumers with visual disabilities. The comparison is carrying out based on the information found in the literature and our perception.

Table 1. General perception of accessibility to promotions via t-, e- and m-marketing

	Characteristics	**t- marketing**	**e- marketing**	**m-marketing**
HCI	Accessibility	High	Medium	Medium
	Comfort	High	Low	Medium
	Convenience	High	Low	Medium
	Interactivity	High	Low	Medium
Marketing	Convenience	Medium	Low	High
	Intrusiveness	Medium	High	Low
	Localization	High	Low	High
	Personalization	Low	Medium	High

Table 1 demonstrates that t-marketing is the best option for organizations to prepare management campaigns focused on consumers with visual disabilities because it provides more advantages in terms of HCI. According with section 2, people with visual disabilities can perceive information by means of ear-sense and tact-sense, thanks to the tools and mechanisms used by companies (e.g. bottles and medicine boxes with Braille text printed on them and audio files), consumers can get information anywhere and anytime, while in e-marketing blind users need use computers and internet connection. Moreover, computers must have installed special software like screen-readers, which have the following drawbacks:

- High rate of speech.
- Don’t respect dots and commas.
- Don’t describe images.

In terms of comfort and interactivity, t-marketing offers the possibility to consumers with visual disabilities to be more independent, making decisions by themselves to purchase a product or not.

However, t-marketing has disadvantages in terms of Marketing. Table 1 demonstrates that t-marketing is poor in personalization because it is focused on capturing many consumers as possible. On the other hand, m-marketing makes use of mobile devices, which can be customized in accordance with the needs and characteristics of its owner, from which we can get a user profile and design advertising based on this profile. Plus, these devices some kind of have built-in

communication technology, giving the facility to the consumer to connect to any network, so they can access the advertising anywhere at any time.

Also, Table 1 shows that m-marketing is the best option for companies to provide their advertising to consumers with visual disabilities. This is convenient because using mobile devices and wireless technologies, they can design advertising using multimedia files, so the cost of it will be lower, giving as a result that a blind user can be part of advertising culture. Moreover, m-marketing offers the advantage that consumers get advertising according to where they are, automatically. Regarding personalization, in m-marketing, blind users can get advertising according to their interests and profiles, eliminating the biggest problem of e-marketing, known as spam, making m-marketing less intrusive.

In order to understand much better the implications of these techniques and tools in the daily life of a blind person, we extended our study with the goal of knowing the experience and perspective of a blind person – a student from Universidad de la Sierra Sur. We designed a questionnaire for knowing his perception about the use of tools and techniques for blind people in marketing (Appendix 1 shows the questionnaire and his answers which explain his perception). It is very clear that few organizations take advantage of the techniques and tools, described in section 2, to prepare their marketing campaign.

Today, the cost of mobile devices is becoming less, offering more services and ease of use, being more accessible for anyone, even for people with visual disabilities. People with visual disabilities have more mobility, comfort and privacy using such devices. Therefore, organizations should use m-marketing as a first option for marketing campaigns oriented toward consumers with visual disabilities.

At the same time, m-marketing is less intrusive than t-marketing and e-marketing because handheld devices are more personal devices than computers, which results in higher privacy for consumers.

5 Conclusions and Future Work

We have presented the first results of our work where the results are clear. We have found several techniques and tools focused on people with visual disabilities. These have been designed and developed by universities, organizations and governments with the goal of giving equal opportunities to participate in a fair society and to have the benefit of the correct use of computer resources. Unfortunately, organizations in Mexico have not taken advantage of such a mechanism in their marketing campaign.

Our study emphasizes the fact that, in order to successfully use techniques and tools for consumers with visual disabilities, organizations must be directly involved in the correct design and development of such technology. It also emphasized that if technology is adapted to consumers with visual disabilities, everybody can benefit, even consumers with no disability.

When we look at consumers who have visual disabilities, we must think about how we can apply technology in their benefit. Several efforts have been described in this paper which can be used to improve their quality of life.

All the techniques and tools described in this paper were developed for people with visual disabilities, but these technologies are not used by marketing or advertising.

Our paper presented the perception of blind person about the correct use of such technology in his daily life. The results described in Appendix 1 are clear; he knows about the technology but he has used it few times in his life.

In order to understand better the use and adoption of techniques and tools for people with visual disabilities, we want to extend our study to all of the country by means of the internet and organizations for blind people.

The concept of the correct use of techniques and tools in marketing focused on consumers with visual disabilities should be considered by organizations. Thanks to researchers and designers, we have technology for blind people; now, organizations must take advantage of them.

Acknowledgments

We would like to thank the reviewers for very constructive and detailed comments which contribute to improve the quality of the paper in the final version. We also would like to thank the following colleagues for their contribution and help: Fernando Martínez Aguilar and Graig Alexander Hilts. Finally, we would like to thank Universidad de la Sierra Sur for its support.

References

[1] Sadeh N., *M-commerce: technologies, services, and business models*, 1st ed. Boston: John Wiley & Sons, Inc., April, 2002.

[2] Siau K. and Shen Z., "Building customer trust in mobile commerce," *Communications of the ACM*, vol. 46, pp. 91-94, 2003.

[3] Balasubramanian S., Peterson R. A. and Jarvenpaa S. L., "Exploring the implications of m-commerce for markets and marketing," *Journal of the Academy of Marketing Science*, vol. 30, pp. 348-361, 2002.

[4] Choi Y. B., Crowgey R. L., Price J. M., and VanPelt J. S., "The state-of-the-art of mobile payment architecture and emerging issues," *International Journal of Electronic Finance*, vol. 1, pp. 94-103, 2006.

[5] Martínez-Peláez R., Rico-Novella F. and Satizabal C., "TOMIN: trustworthy mobile cash with expiration-date attached," *Journal of Software*, vol. 5, pp. 579-584, 2010.

[6] Dickinger A., Haghirian P., Murphy J., and Scharl A., "An investigation and conceptual model of SMS marketing," presented at 37th Hawaii International Conference on System Sciences, 2004.

[7] Tripathi A. K. and Nair S. K., "Mobile advertising in capacitated wireless networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, pp. 1284-1296, 2006.

[8] Huang R. Y. and Symonds J., "Mobile marketing evolution: systematic literature review on multi-channel communication and multi-characteristics campaign," presented at Enterprise Distributed Object Computing Conference, 2009.

[9] Radin T. J., Calkins M. and Predmore C., "New challenges to old problems: building trust in e-marketing," *Business and Society Review*, vol. 112, pp. 73-98, 2007.

[10] Mehra D., Singh V., Paul R., Garg A., and Agarwal C., "Using triggered bus identification and homing system: making public transport accessible for the visually challenged," Social Research in Transport (SORT) Clearinghouse| 2010.

[11] "www.inegi.org.mx, accessed in November 2010."

[12] Kotler P., "A generic concept of marketing," *Journal of Marketing*, vol. 36, pp. 46-54, 1972.
[13] "National library service for the blind and physically handicapped: digital plans and progress, www.loc.gov/nls, accessed in November 2010."
[14] Nielson G. and Harvey G., "Interactive talking books for the blind on CD-room," presented at Johns Hopkins National Search for Computing Applications to Assist Persons with Disabilities, 1992.
[15] Braille L., "Method of writings words, music, and plains songs b means of dots, for use by the blind and arranged for them," 1829.
[16] Way T. P. and Barner K. E., "Automatic visual to tactile translation-part I: human factors, access methods, and image manipulation," *IEEE Transactions on Rehabilitation Engineering*, vol. 5, pp. 81-94, 1997.
[17] "Tactile graphics starter kit. American Printing House for the Blind," 2005.
[18] Clifton R., "Brands and our times," *Journal of Brand Management*, vol. 9, pp. 157-161, 2002.
[19] Asakawa C. and Itoh T., "User interface of a hoe pager reader," presented at International ACM Assistive Technologies, 1998.
[20] Chalamandaris A., Raptis S., Tsiakoulis P., and Karabetsos S., "Enhancing accessibility of Web content for the print-impaired and blind people," in *Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computers Society on HCI and Usability for e-Inclusion*, vol. LNCS 5889, 2009, pp. 249-263.
[21] Talib M. N., Cai S. and Hu M., "Voice browsing approach to e-business access: a blind's perspective," *Computer and Information Science*, vol. 3, pp. 17-23, 2010.
[22] Ng H. S., Sim M. L. and Tan C. M., "Wireless technologies for telemedicine," *BT Technology Journal*, vol. 24, pp. 130-137, 2006.
[23] Varshney U., "Using wireless technologies in healthcare," *International Journal of Mobile Communications*, vol. 12, pp. 113-127, 2007.
[24] Rashvand V. and al. e., "Ubiquitous wireless telemedicine," *IET Communications*, vol. 2, pp. 237-254, 2008.
[25] Nicholson J., Kulyukin V. and Coster D., "ShopTalk: indepent blind shopping through verbal route directions and barcode scans," *The Open Rehabilitation Journal*, vol. 2, pp. 11-23, 2009.
[26] Gharpure C. P. and Kulyukin V. A., "Robot-assited shopping for the blind: issues in spatial cognition and product selection," *Intelligent Service Robotics*, vol. 1, pp. 237-251, 2008.
[27] Sun S. Y., Su C. F. and Ju T., "A study of consumer value-added services in mobile commerce-focusing on domestic cellular phone companies in Taiwan, China," presented at International Conference on Entertainment Computing, 2005.
[28] Yawei T. and Li Z., "Anatomic model on web customer satisfaction based on customer behavior," presented at International Conference on Interaction Sciences, 2009.
[29] Jidong H., C. D. and S. Y., "What affect the acceptance of mobile marketing?," presented at International conference on networking and digital society, 2010.

Appendix 1. Point of view from a consumer with visual disabilities

Questions	Answers
Do you have computer at home?	Yes
Do you have a laptop?	Yes

Do you have a handheld?	Yes
Which device is more comfortable?	Mobile phone
Have you received promotions by t-marketing?, specially for blind people	Yes
If your answer is yes, describe your experience	I could read an advertising, written in Braille code, in a bottle of water
Have you received promotions by e-marketing?, specially for blind people	Not yet, but I can listen some promotions and advertising
If your answer is yes, describe your experience	By means of the screen reader software, I can understand some advertising
Have you received promotions by m-marketing?, specially for blind people	No yet. but I can listen some promotions and advertising
If your answer is yes, describe your experience	By means of multimedia messaging system (MMS), I can listen some promotions and advertising
From your point of view, which marketing can be more useful for blind people?	T-marketing's techniques and tools are good options but unfortunately these techniques and tools are not feasible in my city. On the other hand, e-marketing requires a computer and be online all the time to know the last promotions and advertisings send to my email, making it less attractive. Finally, m-marketing can be a good solution for consumers with visual disabilities because we can receive MMS and promotions that focus on blind people.

Setting the iOS mobile platform for visual servoing applications

Marco Perez-Cisneros[1], Sabrina Vega-Maldonado[2], Erik Cuevas[1], Daniel Zaldivar[1], Patricia Sanchez-Rosario[1]

[1] División de Electrónica y Computación, CUCEI, Universidad de Guadalajara,
{marco.perez, erik.cuevas, daniel.zaldivar, patricia.sanchez}@cucei.udg.mx
[2] Maestría en Sistemas de Información, CUCEA, Universidad de Guadalajara,
vms842@alumnos.cucea.udg.mx

Abstract. This work aims to contribute towards developing a novel platform for deploying Visual Servoing (VS) applications over mobile processing units. As several operating systems and programming environments are available within a quite volatile market, one platform must be selected in order to explore its actual capacities for deploying a visually guided robotic application. This paper presents the on-going evolution of one project targeting the development of a real-time VS control toolkit under the iOS platform and its related tools. The discussion includes first results along with some conclusions and future work.

Keywords: iOS programming, visual servoing, objective-C language.

1 Introduction

Recent technology developments have enriched the consumer market by providing powerful processing platforms ranging from mobile telephone to gaming consoles. In particular, the technology around the last generation of mobile phones, largely known as smartphones, have ubiquitously integrated several devices which encompass novel user interfaces, powerful processors and a handy sensor set with relevant telecommunications ports under a unified architecture.

A comprehensive consumer market analysis from 2006 [1], estimated that there were more than 2.5 billion mobile phones worldwide at the time, with numbers at 2009 reaching 175 million of smartphones alone [2]. In the last four years, the development has been accelerated as many companies are investing largely on developing new hardware platforms and software applications. In 2010, the mobile market widened with experienced companies re-launching their efforts either to relocate their technology surpassing competitors in the segment or to introduce more powerful processing devices.

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 77-86

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

Under such context, mobile processing in general is likely to drive computing subjects for the foreseeable future as new applications and services are becoming available worldwide [3]. Subjects such as distributed computing and pervasive mobile processing are rapidly evolving and becoming attractive to several subjects on science and engineering.
On the other hand, Visual Servoing (VS) is a research area regarding the commanding of robotic devices from visual information. It naturally shares common issues with robotic control, real-time systems and in particular with computer vision as several visual algorithms are demanded by VS, including active vision, visual pose estimation and dynamic vision.

This work aims to contribute towards developing a novel platform for deploying VS application over mobile processing units. As several operating systems and programming environments are available within a quite volatile market, one platform must be selected in order to explore its actual capacities for deploying a visually guided robotic application. This paper presents the on-going evolution of one project targeting the development of a real-time VS control toolkit under the iOS platform and its related tools.

This paper organized as follows: Section 2 provides information about PDA processing platforms while Section 3 depicts a particular discussion on the iOS platform including its application development environment. Section 4 presents a quick overview of VS while Section 5 presents the target VS application. Section 6 offers details about some simulation results and Section 7 discusses on some conclusions and the work ahead.

2 Mobile processing platforms

Following a wide heterogeneity on the mobile processing market, developing applications for a given selected platform still possess a clear challenge for any programmer. Cross-platform programming is still at a very early development stage and knowing more than two mobile programming languages up to an expert level, still demands a breath-taking exercise. However, under such circumstances, exploring the use of an integrated processing unit has become extremely attractive to robotics and its control requirements.

A quick overview on the available hardware platforms and its correspondent software libraries is presented by Table 1 as an overview of the information recently presented in [2]. Such table includes a reduced notation as follows: IDE = Integrated development environment, SDK =software development kit, ADT = Android development tools, JDE = Java development environment, PDK = Palm development kit and NDK = native-code development kit.

Table 1. Available programming platforms to date and their correspondent compatible hardware

OS	Runs on	Related to	Development languages	IDEs, Libraries, frameworks
Android	Open Handset Alliance	Linux	Java	Android SDK and NDK; ADT plug-in for Eclipse
BlackBerry OS	BlackBerry	Unix (BSD &NeXTstep)	Java	BlackBerry JDE
iPhone OS	iPhone, iPad, iPod Touch	Mac OS X	Objective-C / C++	iPhone SDK
PALM WEB OS	Palm Pre, Pixi	Linux	HTML, CSS, Java Script; C / C++	PDK; WebOS plug-in DK; Project Ares (Web based)
Symbian OS	ARM processors	Psion EPOC	C++, Java, other	
Windows Mobile	Windows Mobile smartphones	Windows CE	Visual C++	Windows Phone SDK (works with Visual Studio)

Table 2 shows a handy comparison between commercially available Smartphone platforms. It is easy to browse through their capabilities and services [4].

Table 2. Available smartphone platforms: one hardware comparison array

Platform	iOS 4	Android 2.1 with Sense	Symbian^3	WebOS	Windows Mobile 7
Processor	Apple A4	1GHz Qualcomm Snapdragon	680MHz ARM_{11}-based	600MHz TI $OMAP_{3430}$	1GHz Qualcomm Snapdragon
Storage	16GB / 32GB internal	440MB internal, microSDHC expansion	16GB internal, microSDHC expansion	16GB	Approx. 200MB internal,
Cellular	Quadband GSM, pentaband HSPA	CDMA, EV-DO Rev. A, WiMAX	Quadband GSM, pentaband HSPA	CDMA / EV-DO Rev. A	Quadband GSM, dualband HSPA
WiFi	802.11 b/g/n	802.11 b/g	802.11 b/g/n	802.11b/g	802.11 b/g[1]
Display size	3.5 inches	4.3 inches	3.5 inches	3.1 inches	4.3 inches
Display resol	960 x 649	800 x 480	640 x 360	480 x 320	800 x 480
Display tech.	IPS LCD	LCD	AMOLED	LCD	LCD
Primary camera	5 megapixel AF, LED flash	8 megapixel AF, LED flash	12 megapixel AF, xenon flash	3 megapixel, LED flash	5 megapixel AF, LED flash
Video record	720p at 30fps	720p at 24fps	720p at 25fps	VGA 30fps	VGA at 30fps
Location / orientation sensors	AGPS, compass, accelerometer, gyroscope	AGPS, compass, accelerometer	AGPS, compass, accelerometer	AGPS, accelero-meter	AGPS, compass, accelero-meter

The iPhone platform has been selected as primary hardware set thanks to its mature pervasive computing model which includes well consolidated hardware sensors, communication ports and processing units under one integrated operating system. Remarkably for robotics, most devices include image acquisition and video processing units despite a non-trivial programming language is required. The energy management hardware has been impressively redesigned with sufficiently long

service availability which reinforces its usage for robotic applications. Poor cross-platform code transportation still represent its main drawback as the iOS platform has developed its own programming language and developer kit, making practically impossible to transport code function from other programming platforms.

3 Apple's iOS platform

Apple mobile operating system, also known as iOs, has evolved to become the third most popular system accessing the Internet worldwide. It is also the third regarding any existing operating systems, recently surpassing Linux and lying just behind the popular MS Windows for desktop computers and its own relative, the desktop version of Apple's Mac OS [5].

The iOS platform works over an Apple-Samsung hardware set which includes a GPU, built-in RAM memory and a central ARM processor running over several velocities, depending on the hardware version. For instance, the iPhone© 3G employs an ARM processor at 412Mhz, while its predecessor 3GS includes an ARM Cortex A8 at 600Mhz. Apple more advanced iPhone©, version 4, includes an ARM processor running at 1GHz holding a double core architecture which is considered for future device versions over the Cortex A15 processor. A quick overview of A15 architecture is sketched below at Fig. 1, following the official manufacturer's data sheet [5].

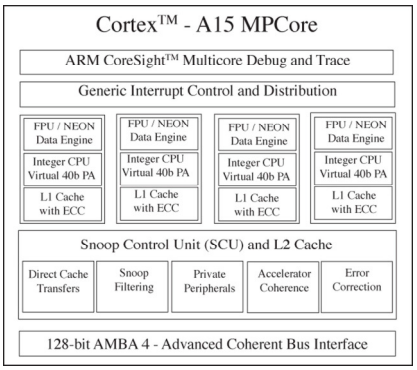


Fig. 1: Cortex A15 ARM microprocessor architecture [5].

The screen size is another important feature to be considered in vision-based robotics applications. The iPhone© platform includes a 480 x 320 pixel screen which has been regarded, until date, as the best voted graphic user interface, including file management and data base access from the same environment.

4 iOS Application Development

The development under the iOS standard follows the popular convention of Model-View-Controller which organizes the overall code by separating data regarding the user interface definition from the information controlling the application. The first component, known as the Model, validates access to all required data and their functionality. A second component, known as the View is basically concerned to define the user interface while the third component, known as the controller, replies to common user events that may eventually modify the goal-achievement management of the program.

The integrated development environment which natively serves the iOS programming is the Xcode©. The environment is a friendly-user interface that makes intensive use of graphic tools and manages different platforms under the same environment. Depending on the final target for the application, Xcode provides Navigation-based applications, OpenGL libraries, utility applications and view-based applications, among others. Fig. 2 shows a classical view of the Xcode integrated development environment containing one the seminal applications for this paper: gray-scale image conversion.

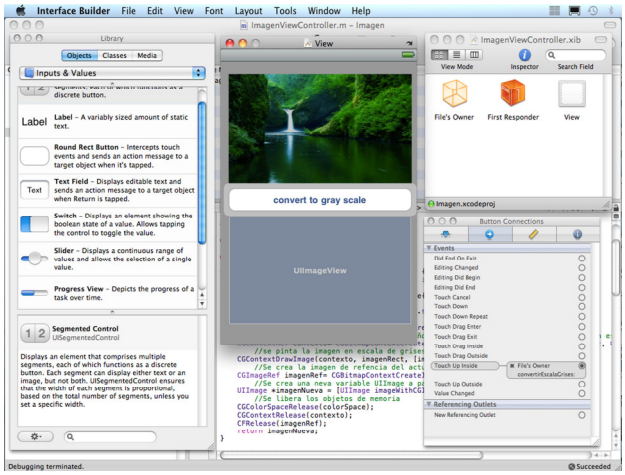


Fig. 2: Interface builder and related windows.

The Interface Builder is the code development hosting the designing of the user's visual interface and its final user interface. The builder includes a link to the code behind each control in the window.

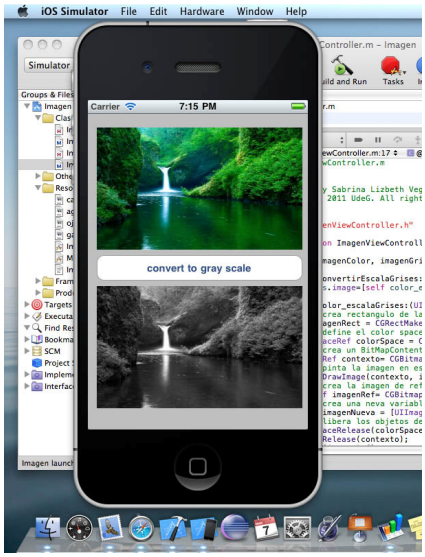


Fig. 3: Gray-level image converted working on the iOS device simulator provided by Xcode

Another important component on the iOS environment is the Device Simulator. Xcode includes a full device view for quick interaction to operating controls and output fields in the software. Fig. 3 shows the view for the gray-level program previously shown in Fig. 2.

4.1 Objective-C programming Program Code

The overall iOS programming environment requires its own programming language known as Objective-C which is a C-language-based meta-set. The Application Program Interface (API) includes high level functions and waste intelligent recollection features. The overall control structure is similar to C language, except by the fact that all objects are created and maintained over a dynamically appointed memory. Despite Objective-C is the main language for the platform, it may welcome other languages such as C#. One of the most important steps towards using the iPhone device for visual servoing applications is its ability to process several computer vision algorithms. A first important step is to generate the gray-level image which can be subsequently used by other processing algorithms [10]. An example showing the programming code for gray-level conversion is presented below.

```
//Gray-scale conversión.
-(UIImage *)convertGrayScale: (UIImage *) image {
   //One rectangle from the image is taken.
   CGRect imgRect = CGRectMake(0,0,image.size.width,image.size.height)
   //the color to be converted into gray-scales is chosen.
   CGColorSpaceRef gray = CGColorSpaceCreateDeviceGray();
```

```
    //A bitmap is created following
    CGContextRef context= CGBitmapContextCreate(nil,image.size.width,
image.size.height,8,0,gray,kCGImageAlphaNone);
    //image is converted into a gray-scale
    CGContextDrawImage(context, imgRect, [image CGImage]);
    //One reference image is created
    CGImageRef imgRef= CGBitmapContextCreateImage(context);
    //A new image is created from the new reference image.
    UIImage *imgNew = [UIImage imageWithCGImage:imgRef];
    //memory is freed
    CGColorSpaceRelease(gray);
    CGContextRelease(context);
    CFRelease(imgRef);
    return imgNew;
}
```

5 Target VS application

Visual Servoing (VS) is a mature research area which shares common issues with robotic control, real-time systems and in special with computer vision, involving subjects often employed in VS schemes such as active vision, visual pose estimation and dynamic vision [6]. Different implementations of VS have been traditionally identified within two main classifications: position-based and image-based visual servoing. An introductory overview of both classes is presented in the now classic VS tutorial in [7]. Basically, in the image-based visual servoing (IBVS) the error signal is computed in the image plane and the regulation commands are generated with respect to such error by means of a visual Jacobian. On the other hand, in the position-based schemes, the image features are used to estimate an object-workspace characterization in such a way that the error can be computed in the Cartesian space and used in the control loop.

Traditionally image-based systems have been regarded to possess a good robustness to calibration errors [7], even in the absence of the object and workspace model. However image-based VS schemes also exhibit some weaknesses such as singularities in the visual Jacobian which may lead to conflicts in the control loop. Other major drawback resides in the fact that an image-based system does not control the robot's end-effector in the Cartesian space, sometimes resulting in complicated or even unrealistic joint configurations being demanded to the robot. New IBVS schemes have been proposed to avoid these problems. Some of the new schemes combine 2-D and 3-D information and pose estimation to create a more complete visual servoing algorithm, for instance the 2-1/2 Visual servoing [7]. However these servoing schemes often require the description of object or its visual features.

In this paper, a low-profile anthropomorphic planar manipulator is equipped with a commercial CCD camera attached to its end-effector. An IBVS algorithm is used to achieve low-speed tracking of a moving object, which is model-free in the sense that no object model is provided [8]. Like other tracking problems, the aim is to keep the camera in a plane parallel to the tracked object. The objective is to investigate the use of mobile processing units to implement computational vision task in a first stage and the overall servoing control scheme in a second stage.

5.1 Experimental Set

The VS platform in this paper, requires a special but easy to conform hardware setup. The mobile telephone unit and one computer are required to implement the visual servoing system. Basically the smartphone performs all the artificial vision processing and computer controls the robot. The first one is known as the Eye Processor (EP) whereas the second one as the Visual Servoing Controller (VSC). Their tasks can be simple enumerated as follows: the EP provides a high level interface with camera drivers and other user interfaces to two visual tracking algorithms, the color-based and the optic-flow-based trackers. Also it provides the access point to the network communication subprocess which is in charge of establishing the high-speed link to the VSC computer through a reliable filtered link which includes a set of digital filters to overcome noise in the image and in the network link. On the other hand, the VSC computer provides the server for the network link, with an optional input filter to cancel out noise in the network hardware. Most VS schemes do not assume significant mechanical problems, low latency and high resolution on the driven robot. This is not always the case as this paper employs one low-cost TQ MA2000 6-DOF manipulator, designed primarily for teaching purposes. It has limited repeatability, low accuracy and sometimes high mechanical backlash. The planar manipulator is equipped with the iPhone 3GS attached to its end-effector, allowing the phone's camera to see a wide perspective of the robot's workspace. The iPhone's camera can provide high-quality (720p) and low-noise images but at the low rate of 30fps.

5.2 Camera Calibration

The intrinsic and extrinsic camera parameters should be experimentally calculated by using the camera model and calibration method [9]. The iPhone camera model is thus compacted into the classical C matrix:

$$C = \begin{bmatrix} f_u & \alpha_c f_u & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1724 & 0 & 696 \\ 0 & 1720 & 722 \\ 0 & 0 & 0 \end{bmatrix} \tag{1.1}$$

With with $f_u$and $f_v$being the focal distance in pixel units whereas α_c=0, f_c, u_0 and v_0 represent the camera intrinsic parameters.

5.3 Experiment simulation

The control task can be described as follows: the TQ MA2000 Robot is resting over a flat surface. A circular miniature railway is placed within the robot's workspace, 0.10 meters away from its base. The camera registers four features of one replica train running over the track in order to perform the tracking. The train motion is simulated as a simple circular motion with a constant angular velocity as follows:

$$\begin{bmatrix} x_{train} \\ y_{train} \\ z_{train} \end{bmatrix} = \begin{bmatrix} 0.44\cos(-0.698t - 3/4\,\pi) + 0.54 \\ 0.44\sin(-0.698t - 3/4\,\pi) \\ 0.04 \end{bmatrix} \tag{1.2}$$

with r = 0.44 meters and the railway circle centered at (0.54, 0, 0). The phase value of $3/4\pi$ locates the initial position of the train in the right corner of the robot's workspace. The negative sign in the expression produces a clock-wise rotation. Notice that the train speed is chosen to be ω=0.112 m/sec. Using this value the train is able to ride through the arc of an angle of 45 degrees in 13.99 seconds.

6 Simulation results

The system is simulated through 13.99 seconds, which is the time required for the train to ride all the way in front of the robot. Fig. 4 shows the trajectory of the features in the image plane using a PI regulator with K_p = -2.0 and K_i = -0.1. Fig. 4 shows the difference between the target and current features in the image.

7 Conclusions and future work

This paper has presented the on-going research on developing a full set of VS libraries for mobile processing units such as smartphones or similar devices. To date, the work has focused on exploring few available platforms in the market, in particular the conveniences of the iPhone mobile set. However, much remains for future implementation such as the real-time operation and the cross-platform implementation of final VS libraries for mobile processing units.

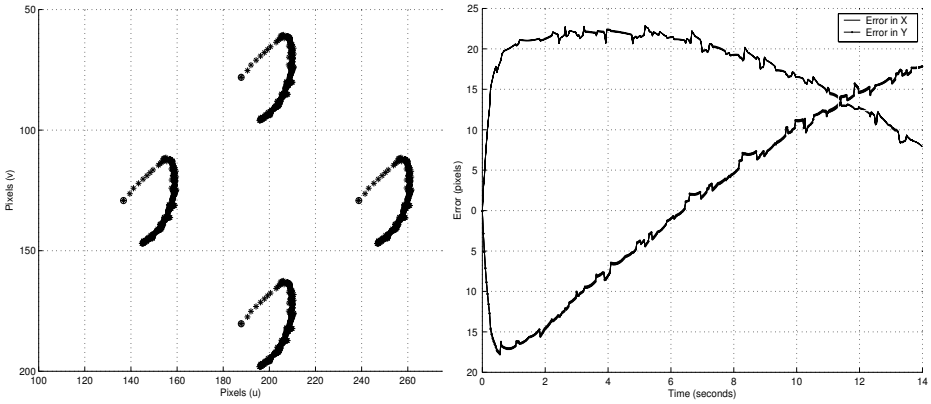


Fig. 4: Feature trajectory on the image plane (left) and the difference between the target and the current feature set in the image plane (right).

References

1. Lars Kulik, "Mobile Computing Systems Programming: A Graduate Distributed Computing Course", IEEE Distributed Systems Online, vol. 8, no. 5, 2007.
2. Daniel Derns.: Tools & Toys, Writing Small. In: IEEE Spectrum. June 2010
3. Gregory D. Abowd, Georgia Teach, Liviu Iftode, Helena Mitchell: The Smart Phone: A first Platform for Pervasive Computing. In: IEEE CS and IEEE ComSoc April-June 2005.
4. "Smartphone comparative table" http://www.iphonefanatic.net/apple/esto-si-es-una-tabla-comparativa-entre-smartphones", 2010
5. Cortex-A15 Processor, http://www.arm.com/products/processors/cortex-a/cortex-a15.php.
6. M A Perez-Cisneros & P A Cook, 2004, Open Platform for Real-Time Robotic Visual Servoing**,** Procs of 10th IASTED Int Conf on Robotics & Applications Hawaii, August, pp.142-147
7. Francois Chaumette, Seth Hutchinson, "Visual Servo Control Part I: Basic Approaches". IEEE Robotics and Automation Magazine, USA 13:4 (2006), 82-90.
8. C. Collewet and , F. Chaumette. "Positioning a camera with respect to planar objects of unknown shape by coupling 2-D visual servoing and 3-D estimations". IEEE Transactions on Robotics and Automation, 18(3):322-333, 2002.
9. Heikkilä, Janne and Silven, Olli. "A Four-Step Camera Calibration Procedure with Implicit Image Correction", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1997, San Juan, Puerto Rico, pp. 1106-1112
10. Cuevas, E., Zaldivar, D. and Perez-Cisneros, M., "Procesamiento Digital de Imágenes con Matlab y Simulink", Book, written on Spanish, ISBN 9785478979738, RA-MA, SPAIN

Developing visual servoing applications over the Windows7 mobile platform

Marco Perez-Cisneros, Laura Lopez-Lopez, Erik Cuevas, Daniel Zaldivar, Marco Cedano-Olvera

[1] División de Electrónica y Computación, CUCEI, Universidad de Guadalajara,
{marco.perez, laural.lopez, erik.cuevas, daniel.zaldivar,
marco.cedano}@cucei.udg.mx,

Abstract. This work aims to contribute towards developing a novel platform for deploying Visual Servoing (VS) applications over mobile processing units. As several operating systems and programming environments are available within a quite volatile market, one platform must be selected in order to explore its actual capacities for deploying a visually guided robotic application. This paper presents the on-going evolution of one project targeting the development of a real-time VS control toolkit under the Windows7 mobile platform and its related tools. The discussion includes first results along with some conclusions and future work.

Keywords: Windows7 Mobile Programming, Visual Servoing, C# language.

1 Introduction

Recent technology developments have enriched the consumer market by providing powerful processing platforms ranging from mobile telephone to gaming consoles. In particular, the technology around the last generation of mobile phones, largely known as smartphones, have ubiquitously integrated several devices which encompass novel user interfaces, powerful processors and a handy sensor set with relevant telecommunications ports under a unified architecture.

A comprehensive consumer market analysis from 2006 [1], estimated that there were more than 2.5 billion mobile phones worldwide at the time, with numbers at 2009 reaching 175 million of smartphones alone [2]. In the last four years, the development has been accelerated as many companies are investing largely on developing new hardware platforms and software applications. In 2010, the mobile market widened with experienced companies re-launching their efforts either to relocate their technology surpassing competitors in the segment or to introduce more powerful processing devices.

Under such context, mobile processing in general is likely to drive computing subjects for the foreseeable future as new applications and services are becoming available worldwide [3]. Subjects such as distributed computing and pervasive mobile

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 87-96

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

processing are rapidly evolving and becoming attractive to several subjects on science and engineering.

On the other hand, Visual Servoing (VS) is a research area regarding the commanding of robotic devices from visual information. It naturally shares common issues with robotic control, real-time systems and in particular with computer visión as several visual algorithms are demanded by VS, including active vision, visual pose estimation and dynamic vision.

This work aims to contribute towards developing a novel platform for deploying VS application over mobile processing units. As several operating systems and programming environments are available within a quite volatile market, one platform must be selected in order to explore its actual capacities for deploying a visually guided robotic application. This paper presents the on-going evolution of one project targeting the development of a real-time VS control toolkit under the Windows7 mobile platform and its related tools.

This paper organized as follows: Section 2 provides information about PDA processing platforms while Section 3 depicts a particular discussion on the Windows7 mobile platform including its application development environment. Section 4 presents a quick overview of VS while Section 5 presents the target VS application. Section 6 offers details about some simulation results and Section 7 discusses on some conclusions and the work ahead.

2 Mobile processing platforms

Following a wide heterogeneity on the mobile processing market, developing applications for a given selected platform still possess a clear challenge for any programmer. First of all, cross-platform programming is still at a very early development stage and knowing more than two mobile programming languages up to an expert level, still demands a breath-taking exercise. However, under such circumstances, exploring the use of an integrated processing unit has become extremely attractive to robotics and its control requirements.

A quick overview on the available hardware platforms and its correspondent software libraries is presented by Table 1 as an overview of the information recently presented in [2]. Such table includes a reduced notation as follows: IDE = Integrated development environment, SDK =software development kit, ADT = Android development tools, JDE = Java development environment, PDK = Palm development kit and NDK = native-code development kit.

In general, the Windows phone7 platform has been selected thanks to its mature programming platform based on the Microsoft Visual Studio. Our research group has done a wide effort to develop new libraries and side applications for robotics research on this platform. Just recently, new potential applications are emerging from the Intel's OpenCV library for Windows Mobile 7 which promises to speed up the

application development regarding computational vision programming on mobile platforms.

Table 1. Available smartphone platforms: one hardware comparison array

Platform	iOS 4	Android 2.1 with Sense	Symbian^3	WebOS	Windows Mobile 7
Processor	Apple A4	1GHz Qualcomm Snapdragon	680MHz ARM_{11}-based	600MHz TI $OMAP_{3430}$	1GHz Qualcomm Snapdragon
Storage	16GB / 32GB internal	440MB internal, microSDHC expansion	16GB internal, microSDHC expansion	16GB	Approx. 200MB internal,
Cellular	Quadband GSM, pentaband HSPA	CDMA, EV-DO Rev. A, WiMAX	Quadband GSM, pentaband HSPA	CDMA / EV-DO Rev. A	Quadband GSM, dualband HSPA
WiFi	802.11 b/g/n	802.11 b/g	802.11 b/g/n	802.11b/g	802.11 b/g[1]
Display size	3.5 inches	4.3 inches	3.5 inches	3.1 inches	4.3 inches
Display resol	960 x 649	800 x 480	640 x 360	480 x 320	800 x 480
Display tech.	IPS LCD	LCD	AMOLED	LCD	LCD
Primary camera	5 megapixel AF, LED flash	8 megapixel AF, LED flash	12 megapixel AF, xenon flash	3 megapixel, LED flash	5 megapixel AF, LED flash
Video record	720p at 30fps	720p at 24fps	720p at 25fps	VGA 30fps	VGA at 30fps
Location / orientation sensors	AGPS, compass, accelerometer, gyroscope	AGPS, compass, accelerometer	AGPS, compass, accelerometer	AGPS, accelero-meter	AGPS, compass, accelero-meter

From a hardware perspective, accessing the camera image and transferring the information into the work memory has been greatly simplified through the Silverlight© library, allowing new robotics applications to be envisioned at short distance. The energy management hardware has been impressively redesigned with sufficiently long service availability which reinforces its usage for robotic applications. Poor cross-platform code transportation still represent its main drawback as the Windows 7 platform focus entirely on C# programming language, making practically impossible to transport code from other programming platforms. Table 2 shows a handy comparison between commercially available Smartphone platforms. It is easy to browse through their capabilities and services.

3 Microsoft's Windows7 mobile platform

The novel Windows7 mobile operating system, also known as Win7, has evolved to become the second most popular system accessing the Internet worldwide. Its overall usage lies far away from the desktop version of MS Windows, its parental relative.

The Windows7 platform works over an Qualcomm Snapdragon QSD8650 hardware set which includes a GPU, built-in RAM memory and a central processor running over 1 GHz. A hard convention has been agreed regarding the screen size and resolution, considering 800 x 480 pixels.

Table 2. Available programming platforms to date and their correspondent compatible hardware.

OS	Runs on	Related to	Development languages	IDEs, Libraries, frameworks
Android	Open Handset Alliance	Linux	Java	Android SDK and NDK; ADT plug-in for Eclipse
BlackBerry OS	BlackBerry	Unix (BSD and NeXTstep)	Java	BlackBerry JDE
iPhone OS	iPhone, iPad, iPod Touch	Mac OS X	Objective-C / C++	iPhone SDK
PALM WEB OS	Palm Pre, Pixi	Linux	HTML, CSS, Java Script; C / C++ (vía PDK)	PDK; WebOS plug-in DK; Project Ares (Web based)
Symbian OS	ARM processors	Psion EPOC	C++, Java, others	
Windows Mobile	Windows Mobile smartphones	Windows CE	Visual C++	Windows Phone SDK (works with Visual Studio)

4 Windows7 Mobile Application Development

The development under the Win7 standard follows the popular convention of Model-View-Controller which organizes the overall code by separating data from the user interface definition and the required logic control sentences. The first component, known as the Model, aims to validate access to all required data and their functionality. A second component, known as the View, is basically concerned to the user interface while the third component is called the controller because it must reply to common user events that may eventually modify the goal-achievement management of the program.

The integrated development environment which natively serves the Windows7 mobile programming is the MS Visual Studio©. The environment is a friendly-user interface that makes intensive use of graphic tools and manages different platforms under the same environment. A quick view of its classical Integrated Developer Environment can be seen at Fig. 1. The Interface Builder is the code development hosting the visual interface designing and the correspondent user interface. The builder includes a link to the code behind each control in the window. Another important component on the Windows7 mobile environment is the device simulator which is included into the Visual Studio, providing a quick interaction with operating controls and output fields in the software.

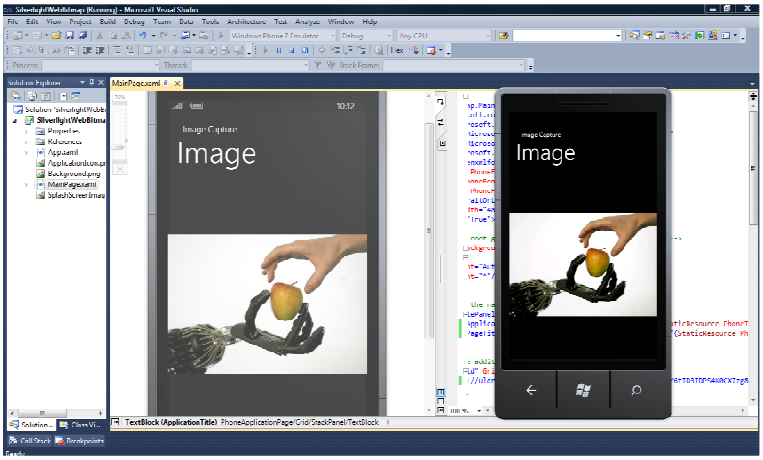


Fig. 1: Visual Studio interface builder and related windows.

4.1 Windows 7 programming code

The overall Windows7 programming (WP7) environment has also defined its own development based on the C# language which is built as a C-language meta-set. The Application Program Interface (API) includes high level functions and waste intelligent recollection features. On the other hand, the overall control structure is similar to C language, despite all objects are always created and maintained over a dynamically appointed memory.

WP7 offers transparent interoperability to other Windows-based OS. Therefore, code reusability is assured as the new technology Windows Presentation Foundation is available to assist the transport of desktop applications into the mobile platforms. Despite C# is the main language in the platform WP7 provides an easy bridge to welcome other languages. A code sample for deploying a basic camera application is explained below giving a quick insight into the language structure.

The System.Windows.Media.Imaging library is required in order to access all video-coding functions. The code example below refers to most common camera manipulation functionalities as follows:

```
void btnCapturar_Click(object sender, RoutedEventArgs e)
{
   //Validates if video capture is allowed.
   if (!CaptureDeviceConfiguration.AllowedDeviceAccess )
   {
      CaptureDeviceConfiguration.RequestDeviceAccess();
   }
```

```
        //If such validation is true, it is possible to assign the handler variable.
        //Video declaration (_cs)
        if (CaptureDeviceConfiguration.RequestDeviceAccess())
        {
           System.Windows.Media.VideoCaptureDevice videodev;
           videodev = CaptureDeviceConfiguration.GetDefaultVideoCaptureDevice();
           if (videodev !=null)
           {
              VideoBrush vb = new VideoBrush();
              vb.SetSource(_cs);
              _cs.Start();
              grid1.Background = vb;
           }
        }
    }
```

5 Target VS application

Visual Servoing (VS) is a mature research area which shares common issues with robotic control, real-time systems and in special with computer vision, involving subjects often employed in VS schemes such as active vision, visual pose estimation and dynamic vision [4]. Different implementations of VS have been traditionally identified within two main classifications: position-based and image-based visual servoing. An introductory overview of both classes is presented in the now classic VS tutorial in [5]. Basically, in the image-based visual servoing (IBVS) the error signal is computed in the image plane and the regulation commands are generated with respect to such error by means of a visual Jacobian. On the other hand, in the position-based schemes, the image features are used to estimate an object-workspace characterization in such a way that the error can be computed in the Cartesian space and used in the control loop.

Traditionally image-based systems have been regarded to possess a good robustness to calibration errors [6], even in the absence of the object and workspace model. However image-based VS schemes also exhibit some weaknesses such as singularities in the visual Jacobian which may lead to conflicts in the control loop. Other major drawback resides in the fact that an image-based system does not control the robot's end-effector in the Cartesian space, sometimes resulting in complicated or even unrealistic joint configurations being demanded to the robot. New IBVS schemes have been proposed to avoid these problems. Some of the new schemes combine 2-D and 3-D information and pose estimation to create a more complete visual servoing algorithm, for instance the 2-1/2 Visual Servoing [6]. However these servoing schemes often require the description of object or its visual features.

In this paper, a low-profile anthropomorphic planar manipulator is equipped with a commercial CCD camera attached to its end-effector. An IBVS algorithm is used to achieve low-speed tracking of a moving object, which is model-free in the sense that no object model is provided [7]. Like other tracking problems, the aim is to keep the camera in a plane parallel to the tracked object. The objective is to investigate the use of mobile processing units to implement computational vision task in a first stage and the overall servoing control scheme in a second stage.

5.1 Experimental Set

The VS platform in this paper, requires a special but easy to conform hardware setup. The mobile telephone unit and one computer are required to implement the visual servoing system. Basically the smartphone performs all the artificial vision processing and computer controls the robot. The first one is known as the Eye Processor (EP) whereas the second one as the Visual Servoing Controller (VSC). Their tasks can be simple enumerated as follows: the EP provides a high level interface with camera drivers and other user interfaces to two visual tracking algorithms, the colour-based and the optic-flow-based trackers [8]. Also it provides the access point to the network communication sub-process which is in charge of establishing the high-speed link to the VSC computer through a reliable filtered link which includes a set of digital filters to overcome noise in the image and in the network link. On the other hand, the VSC computer provides the server for the network link, with an optional input filter to cancel out noise in the network hardware.

The experimental setup employs the low-cost TQ MA2000 6-DOF robot manipulator which has been designed primarily for teaching purposes. It has limited repeatability, low accuracy and sometimes high mechanical backlash. This planar manipulator is equipped with an Win7 mobile phone attached to its end-effector. The W7 phone's built-in camera provides high-quality and low-noise images at a low rate of 30fps. Actually, the W7 mobile phone processor implements the EP duties. As it is discussed later in this paper, it has been envisioned that future experiment with this platform would seek to implement VSC tasks on the same Win7 mobile phone processing unit.

5.2 Camera Calibration

The intrinsic and extrinsic camera parameters for the built-in Win7 mobile phone camera should be experimentally calculated by using the camera model and calibration method from [7]. The Win7 camera model is thus compacted into the classical C matrix:

$$C = \begin{bmatrix} f_u & \alpha_c f_u & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 3198 & 0 & 1218 \\ 0 & 3278 & 767 \\ 0 & 0 & 0 \end{bmatrix} \tag{1.1}$$

With f_u and f_v being the focal distance in pixel units whereas α_c=0, f_c, u_0 and v_0 represent the camera intrinsic parameters. Notice that unfortunately the camera in the mobile handset holds a considerable focal distance which seriously limits its actual capability to focus on closer objects.

5.3 Experiment simulation

The control task can be described as follows: the TQ MA2000 Robot is resting over a flat surface. A circular miniature railway is placed within the robot's workspace, 0.10

meters away from its base. The camera registers four features of one replica train running over the track in order to perform the tracking. The train motion is simulated as a simple circular motion with a constant angular velocity as follows:

$$\begin{bmatrix} x_{train} \\ y_{train} \\ z_{train} \end{bmatrix} = \begin{bmatrix} 0.44\cos(-0.698t - 3/4\,\pi) + 0.54 \\ 0.44\sin(-0.698t - 3/4\,\pi) \\ 0.04 \end{bmatrix} \tag{1.2}$$

with r = 0.44 meters and the railway circle centred at (0.54, 0, 0). The phase value of $3/4\pi$ locates the initial position of the train in the right corner of the robot's workspace. The negative sign in the expression produces a clock-wise rotation. After the initial design and training stage, the VS system is tested on simulation over the MA2000 robot. The performance is analysed through the step-response of the VS tracking system. Initially the robot and the camera are located over the moving object following the teaching-by showing technique. The object is a small replica train. The step response experiment begins with the train moving from its stationary position at a constant velocity of 3.65 cm/sec, following a straight trajectory. As the robot starts tracking, the object stops after 2 seconds allowing the robot to completely catch the object as its features return to the target location in the image space.

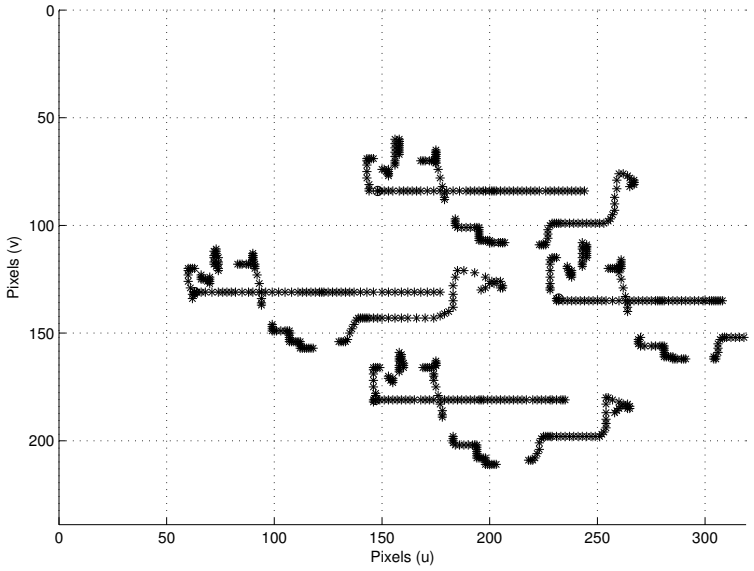


Fig. 2: Trajectory of the features in the image plane with respect to their target location in the step response experiment using linear controllers.

6 Simulation results

Results from the simulation are used by the designer to improve the real-time system performance. The system is simulated for a step response with an operative interval of 2 seconds, which is the time required for the robot to catch the train's movement. Fig. 2 shows the trajectory of the features in the image plane using a PI regulator with K_p = -2.0 and K_i = -0.1 for a step response. Fig. 3 shows the trajectory for each link in the MA2000 robot. Notice how all links go to an stable response after 6 seconds approximately.

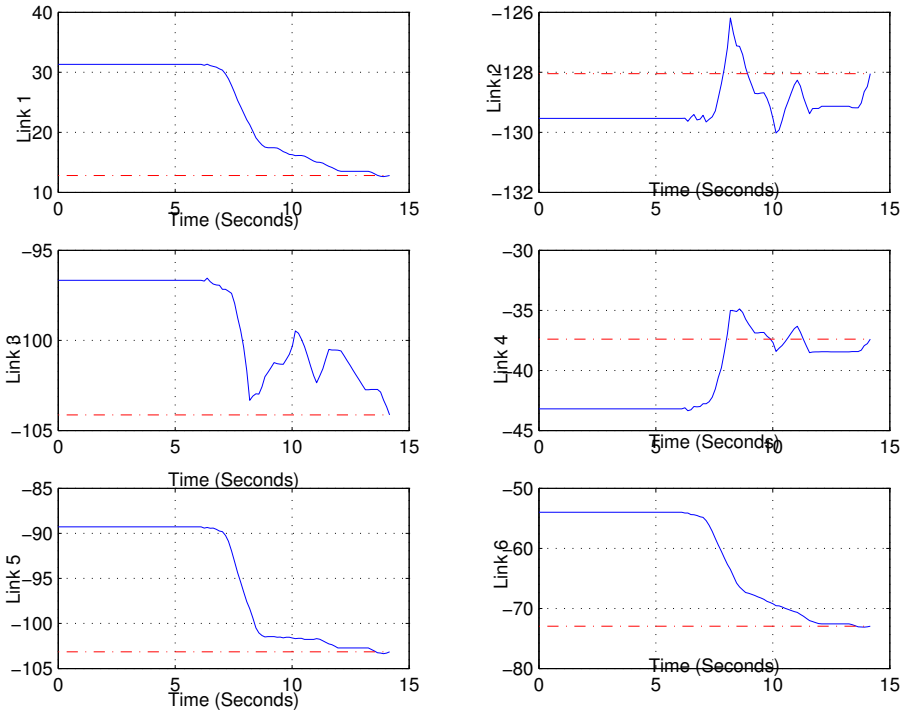


Fig. 3: Robot states during the step response simulation.

7 Conclusions and future work

This paper has presented the on-going research on developing a full set of VS libraries for mobile processing units such as smartphones or similar devices. To date, the work has focused on exploring each available platform in the market, seeking for robust development environment. However, much remains for future implementation such as the real-time operation and the cross-platform implementation of final VS libraries over mobile processing units.

References

1. Lars Kulik, "Mobile Computing Systems Programming: A Graduate Distributed Computing Course", IEEE Distributed Systems Online, vol. 8, no. 5, 2007.
2. Daniel Derns.: Tools & Toys, Writing Small. In: IEEE Spectrum. June 2010.
3. Gregory D. Abowd, Georgia Teach, Liviu Iftode, Helena Mitchell: The Smart Phone: A first Platform for Pervasive Computing. IEEE CS April-June 2005.
4. M A Perez-Cisneros & P A Cook, 2004, Open Platform for Real-Time Robotic Visual Servoing**,** Procs of 10th IASTED Int Conf on Robotics & Applications Hawaii, August, pp.142-147
5. Francois Chaumette, Seth Hutchinson, "Visual Servo Control Part I: Basic Approaches". IEEE Robotics and Automation Magazine, 13:4 (2006), 82-90.
6. C. Collewet and , F. Chaumette. "Positioning a camera with respect to planar objects of unknown shape by coupling 2-D visual servoing and 3-D estimations". IEEE Transactions on Robotics and Automation, 18(3):322-333, 2002.
7. Heikkilä, Janne and Silven, Olli. "A Four-Step Camera Calibration Procedure with Implicit Image Correction", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1997, San Juan, Puerto Rico, pp. 1106-1112
8. Cuevas, E., Zaldivar, D. and Perez-Cisneros, M., "Procesamiento Digital de Imágenes con Matlab y Simulink", Book, written on Spanish, ISBN 9785478979738, RA-MA, SPAIN

Ambient Intelligence Application of Bayesian Networks in a Home Security System

Gustavo López, Ramón Brena

ITESM, Campus Monterrey
Av. Eugenio Garza Sada 2501 Sur, Monterrey, N.L., México
gustavo.alh@gmail.com, ramon.brena@itesm.mx

Abstract. The abundance of sensors in daily life infrastructures and mobile devices can allow to determine what the users are doing, which is the situation of the environment they are in, and therefore what needs they can have and take action accordingly. Artificial Intelligence techniques are applied in order to give the users the functionality that best suits their needs. This is what is called "context-aware computing". The term "Ambient Intelligence" refers to this technology and emphasizes the incorporation of local intelligence to computing components. Ambient Intelligence is a huge field that goes from the acquisition of data from the environment, to fusioning the gathered information and data, to extracting situation characteristics, and to finally selecting and providing adequate information and services based on the extracted context. There are many applications of this technology. In this research paper, we focus in the use of bayesian networks techniques to intelligently use the information about the location and situation of persons and their surrounding environment, registered by the use of sensors, in order to deliver appropriate actions in a home security system and improve the adequacy of this kind of commercial systems.

Key words: ambient intelligence, ami, context awareness, bayesian networks, home security systems

1 Introduction

Advances in the miniaturization of electronics are allowing technological devices with various capabilities and interfaces to become part of our daily life. Sensors, actuators, and computing components can now be purchased at very affordable prices. This technology can be networked and used with the coordination of highly intelligent software to understand the events and relevant context of a specific environment and to take sensible decisions in real-time or *a posteriori*. [1]

This is related to several current technological trends. For example, Home Automation has been around for years, and applications include simple things like turning lights on and off depending on the user's location in the house, but too rigid systems have limited the possibilities of this area. The term "Ambient Intelligence", or AmI, refers to a situation where devices work in concert in order to relate to human needs. In "Ubiquitous Computing", computing devices "disappear" and integrate to the world people live in.

As it is stated in the Ubiquitous Computing manifesto [2]:

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 97-106

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

"The emergence of powerful digital infrastructures, wireless networks and mobile devices has already started to move computing away from the desktop and embed it in the public spaces, architectures, furniture and personal fabric of everyday life."

The Ubiquitous Computing paradigm and, most recently, the Ambient Intelligence paradigm, are the visions in which technology becomes invisible, embedded, present whenever we need it, enabled by simple interactions, attuned to all our senses and adaptive to users and contexts [3]. From a technological point of view, Ambient Intelligence builds on early innovation concepts such as Ubiquitous Computing [4] and Pervasive Computing [5]. The major new things in Ambient Intelligence are the user involvement and situation modeling.

The algorithmic techniques and methods that apply to design for intelligence in Ambient Intelligent systems are rooted in the field of Artificial Intelligence. AI is the scientific and technological pursuit that aims at designing and analyzing algorithms that upon execution give electronic systems intelligent behavior. [6]

One of such techniques are the Bayesian Networks [7]. They are graphical structures for representing the probabilistic relationships among a large number of variables, present in real-world phenomena, and doing probabilistic inference with those variables.

A relevant real life problem for the application of AmI and Situation Modeling theory is that of a Home Security System. Common security systems are rigid, stimulus-response systems. There is an area of opportunity for implementing AmI systems, procedures and techniques in this real scenario.

The basic idea behind Home Security Systems based on Ambient Intelligence is to integrate current technological infrastructure (sensors, actuators, etc) with recent research in AmI's field focused toward personal security. This brings systems with intelligence that take situations into account (e.g. implicated users' position, habits, interpretation of activities, pertinence of notifications, etc) and act based on them.

This kind of intelligent security systems integrate the following technological components:

- Intelligent software (Identification of situations)
- Sensors (Doors, Windows, Movement, Smoke, etc)
- Communication networks
- Actuators (Notifications, Alarms, Lights, etc)
- Portable devices

Ambient Intelligence is a huge field that goes from the sensor acquisition of data from the environment, to fusioning the gathered information and data, to extracting situation characteristics, and to finally selecting and providing adequate information and services based on the extracted context. Each of these areas, or components of the process, have underlying theory and concepts of their own.

The focus of this research work, is in the Intelligent Software that implements *Situation Modeling in Ambient Intelligence* taking information from the system's sensors

and providing services using the system's actuators. The problem of this Case Study is approached using the Bayesian Networks Artificial Intelligence technique. With the use of Bayesian Networks, we examine how the information provided by the security sensors along with other parameters in the users' environment can allow to determine which is the state of the environment, which is the situation they are in, and therefore perform the corresponding action with the actuators.

2 Proposed Solution Model

In this work several sensors are distributed in the important locations of a house (possible access locations, proper measurement locations, etc). These sensors are considered to be forming a sensor network connected to a central intelligent system that takes appropriate actions. Each sensor pass its information to the system in real-time. The intelligent system makes use of a Bayesian Network to take these input parameters and decide the most pertinent action to take.

A Bayesian Network was chosen to address the problem due to its properties. Its most important advantages are the following [8]:

- By exploiting conditional independences entailed by influence chains, we are able to represent a large instance in a Bayesian network using little space.
- We are often able to perform probabilistic inference among the features in an acceptable amount of time.
- The graphical nature of Bayesian networks gives us a much better intuitive grasp of the relationships among the features.

A typical distribution of security sensors and actuators in a house is used in this case study. The schematic for the configuration of the sensors and actuators is as in Figure 1. The symbols in the figure represent the following:

- D1: Door sensor
- W1,W2,W3,W4: Window Glassbreak sensors
- M1,M2,M3,M4: Motion detectors
- S1,S2: Smoke detectors
- G1: Carbon Monoxide detector
- L1: Luminance sensor
- T1: Temperature sensor
- C1,C2: Cameras
- SI1: Siren
- LI1: Light alarm
- AC1: Central Air Conditioner

2.1 Bayesian Network Design and Implementation

A Bayesian Network was used as the solution model for this case study. It provides an important way to solve the problem and implement the AmI paradigm. The solution model, its design and implementation using this technique is shown in this section.

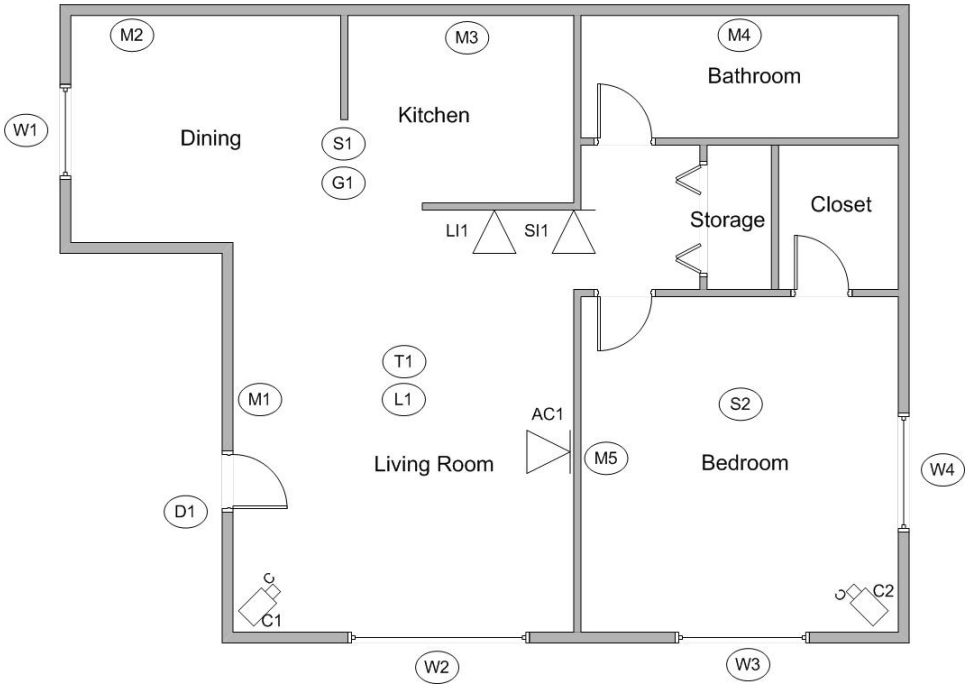


Fig. 1. Sample configuration of sensors in home

Concepts from the Ambient Intelligence and Situation Modeling framework are implemented.

For the design of the network, the AmI awareness concepts are applied. The five types of context-awareness of this application are addressed as follows:

1. *Identity-awareness (IAw)*: a simple user module is implemented to hold what kind of user(s) are in the system. In this case, four kinds of user were to be modeled: House residents, Guests, Intruders, and Animals.
2. *Location-awareness (LAw)*: the different sensors provide the location of the users in the house. In this scenario, five locations inside the house were to be modeled: Kitchen, Living Room, Bathroom, Bedroom, and Dining Room.
3. *Time-awareness (TAw)*: the time is taken directly from the system and provides valuable information. For this implementation, seven time periods represent hours with common activities: 1-5am, 5-8am, 8-1pm, 1-3pm, 3-7pm, 7-10pm, 10-1am.
4. *Activity-awareness (AAw)*: the task which the user carries out is inferred from the sensors. The activities to be modeled are: Having a Meeting, Sleeping, Cooking, Eating, and Watching TV.
5. *Objective-Awareness (OAw)*: this type of awareness is not considered in this implementation due to the nature of the problem. Most of the security system's functionality is obtained through the previous types of awareness.

All these types of awareness answer the five basic questions ("Who", "Where", "What", "When" and "Why") which provide the guidelines for context modeling. This kind of information allows us to adapt or build the needed technology to disperse throughout the environment and to model the human behavioral support. [9]

As a result from the previous guidelines, we obtain what are the middle layers of the Bayesian Network, which answer the important main questions of an AmI system. Now, the next step is to introduce the physical configuration of the home security system into the Bayesian model. This is done by introducing each one of the sensors as the input, first layer, to the system and each one of the actuators as the output, final layer, that provides a service. By doing this, we get the required nodes for the desired functionality of the network.

What follows is to establish the relationships between the nodes. There are three main design points that are relevant to mention, those are:

- One important design point in doing this is to establish the relationships so as to minimize the number of connections, and by that, minimizing the size of the conditional probability tables of the Bayesian Network.
- Another important point is to consider real-world causal relationships (i.e. cause-effect relationships between the variables in the nodes), as this usually minimizes the size of the conditional probability tables.
- A recommendation point is to exploit conditional independences entailed by influence chains, as this allows to represent a large instance in a Bayesian Network using little space.

For this particular problem, the relationships between the nodes are established following a cause and effect pattern. This is implemented as follows:

1. The activation of the *sensors* is an effect of the different *kind of users* being in an specific *location* in the house. This relationship generates causal arrows going from the users and location nodes to the sensors.
2. Different *hours of the day* cause the *users* to be in the house or not, and also to be in a different *location* of the house. This relationship generates causal arrows going from the time node to the location and users' nodes.
3. The fact that a certain *kind of users* are in a certain *location* of the house implies they are doing a certain *activity*. This relationship generates causal arrows going from the users and location nodes to the activity nodes.
4. Risk *situations* happen when specific *events* occur implying a specific kind of *user* in a specific *location* of the house. This relationship generates causal arrows going from specific combinations of sensors, location and users nodes to the risk situation nodes.
5. The different *activities* and *situations* in the house make more pertinent for certain *actions* to be carried out. This relationship generates causal arrows going from activity and risk situation nodes to the actuators.

As a result from the previous node relationship implementation, we obtain the interconnected network representing the model of the problem. Figure 2 shows this Bayesian Network. The different types of awareness are highlighted. This solution model takes

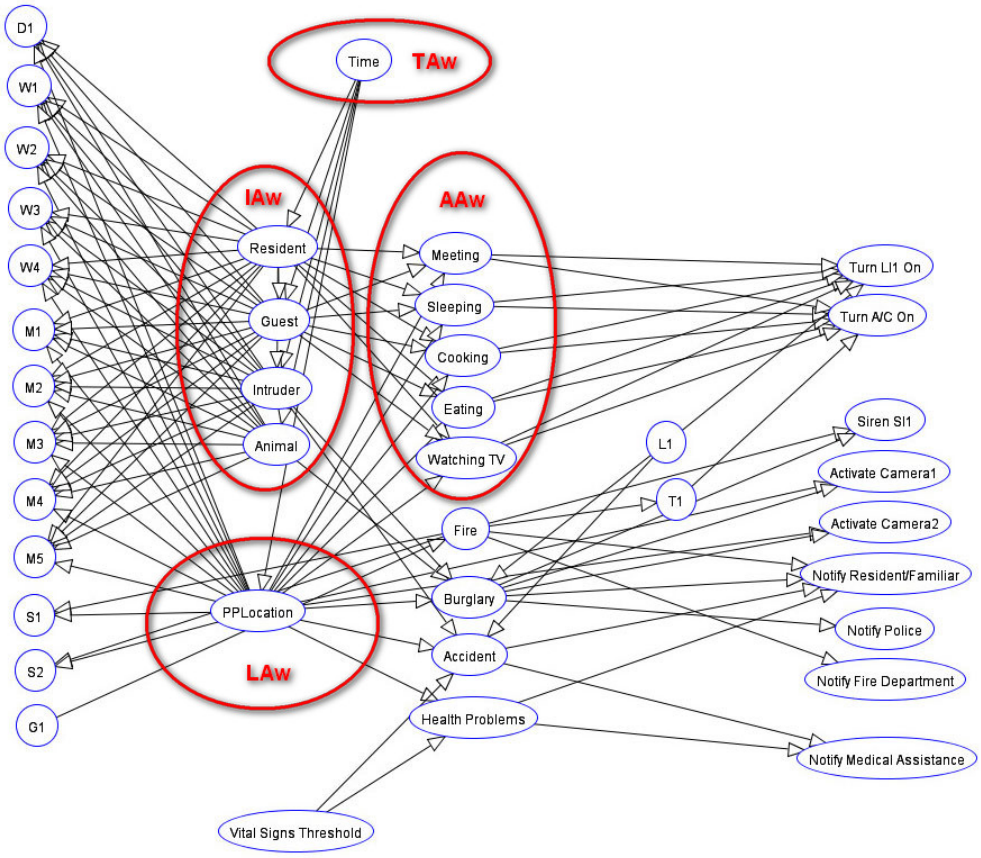


Fig. 2. Bayesian Network for sample home configuration

into account the configuration of sensors and actuators in the house as well as the previous mentioned types of awareness.

Finally, to complete the Bayesian Network so that it can be used as a proper solution model, conditional probability tables have to be introduced. For each of the nodes, values to the probability tables are filled out according to how the relationships between them behave in the real world. The size and complexity of the table for a particular node depend on the number of arrows entering the node. In general, the proposed network is neither very complex nor very simple.

The description of the solution model presented in this section gives an overview of the concepts from the AmI theoretical framework as well as the procedures implemented to make it work. The result is a Bayesian Network model of a Home Security System that works with the Ambient Intelligence paradigm at its core.

3 Experimentation and Results

In this section the functionality and performance of the model is evaluated. There can be several performance measures, but the ones of interest at this moment are the values of the outputs of the security system given a certain configuration of the input sensors. Experiments are carried out to test this. In these experiments, different values of the main sensor parameters are tested. This is done in order to see if the model provides the desired functionality, based on its design and expected behavior.

The experiments are set up in this way: a scenario of a state of the house is supposed and the state of the sensors under this scenario is fed to the system, the system then calculates the value probabilities for the output actuators. These output values are analyzed according to each scenario. The following four scenarios are considered:

- **Scenario 1**: The first scenario is a normal night at 2 am, with the resident at home in his bedroom activating sensor M5. The state of the system is as shown in Figure 3 a).
- **Scenario 2**: The second scenario is minutes afterwards with the resident at home in his bedroom when the W1 sensor detects the dinner room glass breaking. The state of the system is as shown in Figure 3 b).
- **Scenario 3**: The third scenario is the same as Scenario 2, but with the clock pointing at 8pm. The state of the system is as shown in Figure 3 c).
- **Scenario 4**: The fourth scenario continues from scenario 3, but guests arrives at home and are in the dinning room. The state of the system is as shown in Figure 3 d).

The outputs shown in Figure 3 present the changes in the probabilities of the actuators according to each scenario. These outputs are a result of the Bayesian Network probabilities propagation [8] given the inputs stated in each particular scenario. The main changes are the following:

- From Scenario 1 to Scenario 2, it can be seen how the probabilities of activating the alarm and send notifications to the resident and police increase due to the overall configuration of the situation (2 am hour of the day, dinner room glassbreak sensor activated with no residents around, resident located at bedroom, etc).
- Scenario 3 shows the effects of a subtle change in the hour of the day. The same things happening in Scenario 2, while happening at another time produce different probabilities because the prior information of the system points that burglary is more probable at certain hours than at others. Nonetheless, if a certain important event configuration of the sensors occurs, the hour could possibly not have as much effect in the probabilities.
- Scenario 4 gives another pointer. If the system detects there are users in the room where a certain sensor becomes activated, it is more probable to have been activated by accident, and therefore output probabilities change accordingly.

Thus, it can be seen from the results generated by the model that following the AmI paradigm proposed in the solution model shows to be beneficial for the Home Security System application. Clearly, the information provided by the sensors give us

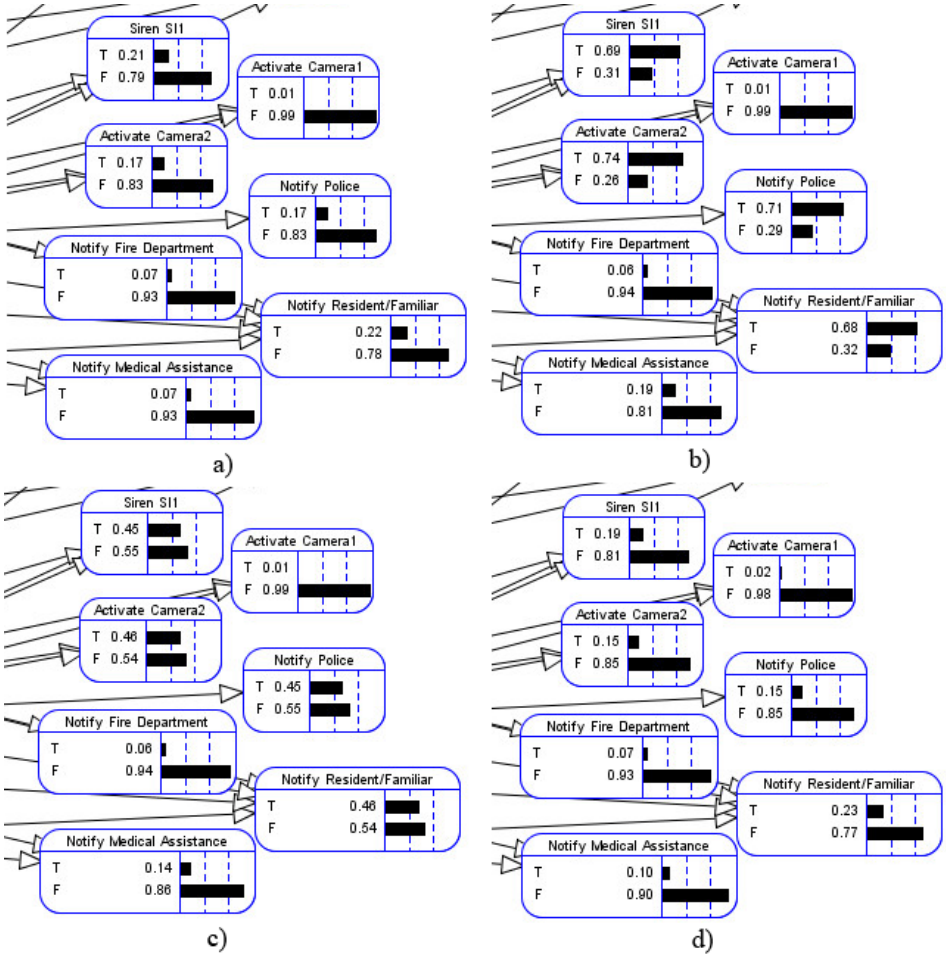


Fig. 3. Output for the four Scenarios configuration. **T** = true and **F** = false

important data to make inferences, obtain conclusions, and act accordingly. Also, the presented results show the power of only a particular implementation of the problem by using Bayesian Networks, which is very interesting as it can be explored using other Artificial Intelligence techniques such as the presented in [7] and [10].

4 Summary and Discussion

The overall results of the experiments carried out allow to actually make conclusions about the quality of the developed model as a solution to the intelligent home security system problem. The following is a summary of the obtained results:

- It can be seen from the experiments that following the Ambient Intelligence paradigm proposed in the solution model can help in identifying potential risky situations for the persons.
- The experiments demonstrated that the Bayesian Network solution model presented the potential for modeling security situations in a house.
- The information gathered by the sensors can be processed intelligently in order to produce useful inferences about the context.
- Results can get very accurate depending on the actual accuracy of the situation modeling, i.e. the better the model represents the actual system, the better the accuracy of the results.
- The experiments also show the flexibility of introducing new features to the network but also the complexity that is added by doing so.

5 Conclusions and Future Work

The solution model presented here took advantage of a whole variety of concepts such as: Context-Awareness, Probability Reasoning, Situation Modeling, Stochastic Processes, Security Systems, among others. It interconnects all those concepts into an integrated model that receives environmental input and produces results according the interpreted situation. The result is a model of an intelligent home security system that works with the AmI paradigm at its core.

Future work includes introducing the concept of *sequential behaviors*, expanding the model to recognize and interpret more complex situations, introducing important features such as a *user profile* and *activity history*, and implementing other Artificial Intelligence techniques. All of these would contribute to taking the AmI problem solution to a more useful level.

References

1. Nakashima, H., Aghajan, H., Augusto, J. C., Eds.: Handbook of Ambient Intelligence and Smart Environments. Springer, New York (2010)
2. Chalmers, D., Chalmers, M., Crowcroft, J., Kwiatkowska, M., Milner, R., O'neill, E., Rodden, T., Sassone, V., Sloman, M: Ubiquitous computing: Experience, design and science. Tech. rep. (2006)
3. Dey, A. K., Abowd, G. D.: Towards a better understanding of context and context awareness. In Proceedings of the Workshop on the What, Who, Where, When and How of Context-Awareness, affiliated with the CHI 2000 Conference on Human Factors in Computer Systems (2000)
4. Weiser, M.: The computer for the twenty-first century. Scientific American, vol. 265(3), pp. 94104. (1991)
5. Satyanarayanan, M.: Pervasive computing, vision and challenges. IEEE Personal Communications (August 2001), pp. 10-17. (2001)
6. Weber, W., Rabaey, J., Aarts, E., Eds.: Ambient Intelligence. Springer, New York (2005)
7. Hopgood, A. A.: Intelligent Systems for Engineers and Scientists. CRC Press (2001)
8. Neapolitan, R. E.: Learning Bayesian Networks. Prentice Hall (2003)

9. Mikulecky, P., Liskova, T., Cech, P., Bures, V.: Ambient Intelligence Perspectives: Selected Papers from the first International Ambient Intelligence Forum 2008 - Volume 1 Ambient Intelligence and Smart Environments. IOS Press, Amsterdam, The Netherlands, The Netherlands (2008)
10. Russell S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall (2002)

Remote Authentication Scheme Using Smart Cards and Fingerprint Technology

Rafael Martínez-Peláez[1], Cristina Satizábal[2], Darío Barragán López[1], Edgar Aquino García[1], Oswaldo Ávila Barrón[1]

[1] Universidad de la Sierra Sur, Licenciatura en Informática, C/ Guillermo Rojas Mijangos S/N, Ciudad Universitaria, Miahuatlán de Porfirio Díaz, 70800, Oaxaca, México
{rpelaez, dbarragan, eaquino, oavila}@unsis.edu.mx
[2] Universidad de la Sierra Sur, División de Estudios de Postgrado, C/ Guillermo Rojas Mijangos S/N, Ciudad Universitaria, Miahuatlán de Porfirio Díaz, 70800, Oaxaca, México
isatizabal@unsis.edu.mx

Abstract. A remote user authentication scheme is a mechanism which identifies legal users and allows access to network services over an open network. However, several authentication schemes proposed in the literature cannot prove the user's physical participation during the login phase, making them vulnerable to different attacks. We propose a new remote user authentication scheme using nonce, smart cards and fingerprint technology for electronic activities. Our scheme provides mutual authentication and session key establishment between the user and the server. The server verifies the identity of the user by means of her fingerprint-template. In order to reduce security risks, the server does not maintain a verification table. The server computes a BioHash and stores it encrypted in user's smart card. Security analysis shows that our scheme provides strong security because the scheme resists common well-know attacks, making it suitable for practical implementation.

1 Introduction

A remote user authentication scheme allows two or more entities to establish a session key which can be used for creating a secure channel over an open network. A session key agreement protocol is used to establish the session key between two entities where each entity contributes with some information to derive that key. Moreover, a remote user authentication scheme permits to identify the legal from the illegal user.

In the literature, we can find several remote user authentication schemes [1-15]. Some of these schemes are based on smart cards [16], [17] and a few of them take advantage of biometrics [18], [19]. In this paper, we look at the remote user authentication schemes which combine smart cards and fingerprint biometric-based [19], [20].

Remote user authentication schemes based on password authentication require that a remote server maintains a verification table, making it vulnerable to steal information [4]. Also, an attacker can use social engineering attacks or brute force

© C. Delgado, C. Gutiérrez, R. Velázquez, H. Sossa (Eds.)
Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 107-113

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

attacks to obtain users' password. For that reasons, password authentication schemes are known as weak security mechanisms.

On the other hand, remote user authentication schemes based on public key cryptography [6], [9], [14] provide strong security in network communication, making them useful in several transactions. However, public key cryptography does not resolve the question Is really who claims to be?. Although public key cryptography offers significant security benefits, it does not resolve the question.

However, biometric technology can answer the previous question. For many years, fingerprint recognition has been used for identification and authentication [18], [19], [20]. The use of fingerprint recognition is justified by four reasons: 1) it is accepted as a valid method for personal identification; 2) the price of a fingerprint reader device is cheap; 3) the processing cost is low; and 4) it is easy to use; for those reasons, fingerprint is the most popular biometric technology around the world.

In 2002, Lee et al. [9] proposed a fingerprint-based remote user authentication scheme. However, Hsieh et al. [11] demonstrated that Lee et al.'s scheme is vulnerable to impersonate attacks.

In this paper, we take advantage of BioHash [21] to avoid storing the fingerprint-template in clear. In order to enhance the security of the proposed scheme, the BioHash is encrypted and the result is stored in user's smart card. Moreover, the scheme requires low-computational cost because user's smart card performs or-exclusive operations and one-way hash functions [22], [23] to create the session key.

The paper is organized as follows. We present the four phases of our scheme in section 2. Section 3 presents the security analysis of the proposed scheme. Finally, conclusions are given in section 4.

2 Proposed Scheme

Our scheme is based on low-computational cryptography which does not require high computational power but provides strong security. The notations used in this paper are described in Table 1.

Table 1. Notations

U	User
S	Server
T	Fingerprint-Template of *U*
ID	Identity of *U*
PW	Password of *U*
K_{pri}/K_{pub}	Private/public key of *U* or *S*
SK	Session key between *U* and *S*
Z	Secret key of *S*
Y	Secret value of *S*
$E_K(.)$	Encryption function using *K*
$D_K(E_K(.))$	Decryption function using *K*
H(.)	One-way hash function
$\|\|$	Concatenate operator
$\oplus$	OR-exclusive operation

The proposed scheme contains four phases:

1. Registration phase, in where *U* will obtain the security parameters to be a legal member.
2. Login phase, in where *U* will be identified by her smart card and initialize the authentication process.
3. Mutual authentication and session key establishment phase, in where *S* will verify the identity of *U* and share common information to create *SK*.
4. Password change phase, in where *U* will have the possibility to change her *PW* without contact *S*.

2.1 Registration Phase

In this phase, *U* is registered by *S*. The process is as follows:

U shares her *ID* with *S* and imprints her fingerprint biometric impression at the sensor to obtain her template *T* which is extracted by the method described in [24]. Then, *S* performs the following operations:

```
Chooses randomly a PW
Computes A = H(H(ID || PW) || Y)
Computes B = H(T)
Computes C = F(B, Z)
Computes G = E_Kpub(C)
Computes I = H(PW) ⊕ A ⊕ H(ID)
Computes J = H (A || H(ID || PW))
```

S stores (*G*, *I*, *J*) in *U*'s smart card. Then, *S* delivers *U*'s smart card and *PW*, through a secure channel.

2.2 Login Phase

In this phase, *U* is identified by her smart card. *U* keys her *ID* and *PW*. Then, *U* carries out the following steps:

```
Computes A' = H(PW) ⊕ I ⊕ H(ID)
Computes J' = H(A' || H(ID || PW))
Verifies J' ?= J
Generates randomly N
Computes L = A ⊕ N
Computes M = H(H(ID || PW) || N)
```

U sends (*L*, *M*, *H*(*ID* || *PW*)) to *S*.

2.3 Mutual Authentication and Session Key Establishment Phase

In this phase, S verifies the identity of U. Moreover, U and S establish $SK_{U\text{-}S}$. The process is as follows:

```
Computes A′ = H(H(ID || PW) || Y)
Computes N′ = A′ ⊕ L
Computes M′ = H(H(ID || PW) || N)
Verifies M′ ?= M
```

If M' and M does not hold, S rejects it. Otherwise, S performs the following operations:

```
Computes N2
Computes P = H(A || N)
Computes Q = P ⊕ N2
Computes SK = H(N2 || N || A)
Computes R = ESK(N+1, Request_hash_value_T)
```

S sends (Q, R) to U.

Upon receiving Q and R, U performs the following operations:

```
Computes P′ = H(A || N)
Computes N2′ = P ⊕ Q
Computes SK = H(N2′ || N || A)
Computes N+1′
Computes DSK(ESK(R)) = N+1, Request_hash_value_T
Verifies N+1′ ?= N+1
```

If $N+1'$ and $N+1$ does not hold, U rejects it. Otherwise, U imprints her fingerprint-biometric impression at the sensor and performs the following operations:

```
Computes H(T)
Computes V = ESK(H(T), G)
```

U sends (V) to S.

Upon receiving V, S performs the following operations:

```
Computes DSK(ESK(V))= H(T), G
Computes DKpriv(EKpub(G)) = C
Computes C′ = F(H(T), Z)
Verifies C′ ?= C
```

If C' and C are equal, the identity of U is assured. Otherwise, S rejects U's request.

2.4 Password Change Phase

Whenever U wants to change or update her PW for another one new, she must perform the following operations:

Computes $A' = H(PW) \oplus I \oplus H(ID)$
Computes $J' = H(A' \,||\, H(ID \,||\, PW))$
Verifies $J' \stackrel{?}{=} J$
Request new password = PW_{new}
Computes $I_{new} = H(PW_{new}) \oplus A \oplus H(ID)$
Computes $J_{new} = H(A \,||\, H(ID \,||\, PW_{new}))$

Finally, the smart card stores I_{new} and J_{new} replacing the old I and J. Now, the new password is successfully updated.

3 Security Analysis

In this section, we prove that our proposed scheme is secure.

Lemma 1: The proposed scheme authenticates the source of the message.

Proof: In fact, $SK = H(N_2 \,||\, N \,||\, A)$ is known only by U and S. S is the unique entity who can compute $H(H(ID \,||\, PW) \,||\, Y)$ and U can recover A from I because she knows the correct ID and PW. Hence, U is sure that she and S share a session key SK. Even though an attacker can capture message V, she cannot recover $H(T)$ and G without know the correct key SK. Moreover, if an attacker can capture message L, she cannot recover A from L without know N and she cannot extract N without know A. In addition, if the attacker intercepts M, she cannot extract $H(ID \,||\, PW)$ and N from M because is computationally infeasible invert a one-way hash function.

Lemma 2: The proposed scheme can resist impersonate attack.

Proof: Suppose that an attacker wants to impersonate U. Assuming that the attacker obtains U's smart card and extracts G, I and J by means of [25], she cannot recover A, PW and ID using any type of combination of G, I and J.

Lemma 3: The proposed scheme can resist server spoofing attack.

Proof: If an attacker has the possibility to intercept messages L, M and $H(ID \,||\, PW)$, she cannot compute A without know Y, she cannot extract N without have A, she cannot compute P without know A and N, and she cannot compute $SK = H(N_2 \,||\, N \,||\, A)$, giving as a result that she cannot compute a valid messages Q and R.

Lemma 4: The proposed scheme can resist user spoofing attack.

Proof: If an attacker has the possibility to intercept messages Q and R, she cannot compute P without know A and N, she cannot recover N_2 from Q without have P and she cannot computes SK for decrypting R.

Lemma 5: The proposed scheme protects U's template.

Proof: In this scheme, S computes a one-way hash function over U's template creating a BioHash. Moreover, S encrypts the BioHash using its public key K_{pub}.

Furthermore, S stores U's BioHash encrypted in U's smart card without store any type of information in a private/public database. If an attacker obtains $G = E_{Kpub}(H(T))$, she cannot recover $H(T)$ without know K_{priv}.

Lemma 6: The proposed scheme withstands leak of password attack.

Proof: In an attacker obtains U's smart card, she cannot recover U's ID, PW and T by using G, I and J or by any type of combination among them.

4 Conclusions

We have proposed a remote user authentication scheme, based on nonce, smart-cards and fingerprint technology, which does not require a verification table. The scheme is based on two key concepts: 1) ID-based, which is used to create the session key between the user and the server; and 2) fingerprint verification, which is used to verify the identity of the user. Thus, the scheme requires the three authentication categories – something she knows, something she has and something she is –, it can resist well-known attacks. Security analysis demonstrated that the proposed scheme is secure against impersonate, server spoofing, user spoofing, and leak of password attacks. Moreover, the scheme protects the user's template creating a BioHash. The scheme can be used in a system which requires high security, such as e-banking.

Acknowledgments

We would like to thank the reviewers for very constructive and detailed comments which contribute to improve the quality of the paper in the final version. We also would like to thank Universidad de la Sierra Sur for its support.

References

[1] Rivest R., Shamir A. and Adleman L., "A method for obtaining digital signatures and public key cryptosystems," *Communications of the ACM*, vol. 21, pp. 120-126, 1978.

[2] Shamir A., "How to share a secret," *Communications of the ACM*, vol. 22, pp. 612-613, 1979.

[3] Lamport L., "Password authentication with insecure communication," *Communications of the ACM*, vol. 24, pp. 770-772, 1981.

[4] Hwang T., Chen Y. and Laih C. S., "Non-interactive password authentication without password tables," presented at IEEE Region 10 Conference on Computer and Communication System, 1990.

[5] Chang C. C. and Wu T. C., "Remote password authentication with smart cards," *IEE Proceedings-E*, vol. 138, pp. 165-168, 1991.

[6] Hwang M. S. and Li L. H., "A new remote user authentication scheme using smart cards," *IEEE Transactions on Consumer Electronics*, vol. 46, pp. 28-30, 2000.

[7] Chien H. Y., Jan J. K. and Tseng Y. M., "An Efficient and practical solution to remote authentication: smart card," *Computers & Security*, vol. 21, pp. 372-375, 2002.

[8] Lee C. C., Li L. H. and Hwang M. S., "A remote user authentication scheme using hash functions," *ACM SIGOPS Operating Systems Review*, vol. 36, pp. 23-29, 2002.
[9] Lee J. K., Ryu S. R. and Yoo K. Y., "Fingerprint-based remote user authentication scheme using smart cards," *IEE Electronic Letters*, vol. 38, pp. 554-555, 2002.
[10] Lin C. W., Shen J. J. and Hwang M. S., "Security enhancement for optimal strong-password authentication protocol," *Operating Systems Review*, vol. 37, pp. 12-16, 2003.
[11] Hsieh B. T., Yeh H. T., Sun H. M., and Lin C. T., "Cryptanalysis of a fingerprint-based remote user authentication scheme using smart cards," presented at IEEE 37th International Carnahan Conference on Security Technology, 2003.
[12] Carlisle A. and Lloyd S., *Understanding PKI: concepts, standards, and deployment considerations*, 2nd ed: Addison-Wesley, Cop., November, 2003.
[13] Das M. L., Saxena A. and Gulati V. P., "A Dynamic ID-based remote user authentication scheme," *IEEE Transactions on Consumer Electronics*, vol. 50, pp. 629-631, 2004.
[14] Badra M. and Urien P., "Introducing smart cards to remote authenticate passwords using Public Key Encryption," presented at IEEE/Sarnoff Symposium on Advances in Wired and Wireless Communication, 2004.
[15] Martínez-Peláez R., Rico-Novella F., Satizabal C., and Pomykala J., "Strong remote user authentication scheme using smart cards," presented at Eighth International Network Conference, 2010.
[16] Vedder K. and Weikmann F., "Smart cards - requirements, properties, and applications," in *State of the Art in Applied Cryptography, Course on Computer Security and Industrial Cryptography*, vol. LNCS 1528, 1998, pp. 307-331.
[17] Trask N. T. and Meyerstein M. V., "Smart Cards in Electronic Commerce," *BT Technology Journal*, vol. 17, pp. 57-66, 1999.
[18] Pankanti S., Bolle R. M. and Jain A., "Biometrics: the future of identification," *Computer*, vol. 33, pp. 46-49, 2000.
[19] Jain A. K., Ross A. and Prabhakar S., "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 4- 20, 2004.
[20] Putte T. V. and Keuning J., "Biometrical fingerprint recognition: don't get your fingers burned," presented at IFIP TC8/WG8.8 Fourth Working Conference on Smart Card Research and Advanced Applications, 2000.
[21] Jain A. K., Nandakumar K. and Nagar A., "Biometric Template Security," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 1 - 20, 2008.
[22] Rivest R., "RFC 1321 - the MD5 message-disgest algorithm," IETF Working Group 1992.
[23] NIST, "Secure Hash Standard (SHA), FIPS PUB 180-1," National Institute of Standards and Technology 1995.
[24] Bae I. G. and al. e., "Online fingerprint verification system using direct minutia extraction," presented at International Conference on Computer Applications in Industry and Engineering, 2000.
[25] Messerges T. S., Dabbish E. A. and Sloan R. H., "Examining smart-card security under the threat of power analysis attacks," *IEEE Transactions on Computers*, vol. 51, pp. 541-552, 2002.

Modeling PMx Trends Contaminants by using Support Vector Machines

Artemio Sotomayor-Olmedo[1], Marco Aceves-Fernandez[1], Efren Gorrostieta-Hurtado[1], Carlos Pedraza-Ortega[1], Emilio Vargas-Soto[1], Ruth Rico-Hernandez[1]

[1]UniversidadAutonoma de Queretaro –Facultad de Informatica, Av. De las Ciencias. S/N, 76230 Queretaro, Mexico
{artemio.sotomayor, marco.aceves, efren.gorrostieta, carlos.pedraza,rico}@uaq.mx

Abstract. Monitoring, modeling and forecasting of air quality parameters are important topics in environmental and health research due to their impact caused by exposing to airborne particles in urban environments. The aim of this article is to show that forecast of daily airborne pollution using support vector machines (SVM) and Kernel functions such as Polynomial, Gaussian, and Spline are feasible. Results are presented using data measurements of Particulate Matter of aerodynamical size on the order of 10 and 2.5 micrograms (PMx) in London-Bloomsbury at south England.

Keywords: Particulate matter, Support Vector Machines, Kernel techniques, PMx, airborne pollution, forecast.

1 Introduction

In recent times, urban air pollution has been a growing problem especially for urban communities. Size, shape and chemical properties govern the lifetime of particles in the atmosphere and the site of deposition within the respiratory tract. Health effects differ upon the size of airborne particulates. In this contribution, PM10 (particles less or equal than 10 micrometers) and PM2.5 (particles less or equal than 2.5 micrometers) are considered due to its effect on human health, according to several authors [1-6]. This is the primary reason this research has been done; to monitor, and model the levels and spread of PMx in urban environments. In previous contributions, it has been shown that forecast of concentration levels of PM10 may be possible by using other techniques such as neural networks and various fuzzy clustering algorithms [7-8]. However, even though these works have shown that is feasible to accurately model the non-linear behavior of the system, a more robust model is needed with an enhanced method to reduce the error between the raw data and the model. For this reason, support vector machines (SVM) are chosen for this work. In this appraisal, the modeling will be carried out using support vector machines working in regression mode. Support vector machines are a recent statistical learning technique, based on machine learning and generalization theories, it implies an idea and could be considered as a method to minimize the risk [9]. Also, a

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 114-122

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

generalization capability makes possible their application to modeling dynamical and non-linear data sets.

2 Support Vector Machines

2.1 Support Vector Machines Background using Regression Mode

The support vector machines (SVM) theory, was developed by Vapnik in 1995, and is applied in many machine-learning applications such as object classification, time series prediction, regression analysis and pattern recognition. Support vector machines (SVM) are based on the principle of structured risk minimization (SRM) [10-11].

In the analysis using SVM, the main idea is to map the original data $\boldsymbol{x}$ into a feature space $\boldsymbol{F}$ with higher dimensionality via non-linear mapping function $\boldsymbol{\phi}$, which is generally unknown, and then carry on linear regression in the feature space [10]. Thus, the regression approximation addresses a problem of estimating function based on a given data set (where x_i represent the input vectors, $d_i$are the desired values), which is produced from the ϕ function. SVM method approximates the function by:

$$y = \sum_{i=1}^{m} w_i \phi_i(x) + b = w\phi(x) + b \tag{1}$$

where $w = [w_1, \ldots, w_m]$ represent the weights vector, b are the bias coefficients and $\phi(x)=[\phi_1(x),\ldots,\phi_m(x)]$ the basis function vector.

The learning task is transformed to the weights of the network at minimum. The error function is defined through the ε-insensitive loss function, $L_\varepsilon(d,y(x))$ and is given by:

$$L_\varepsilon(d, y(x)) = \begin{cases} |d - y(x)| - \varepsilon & |d - y(x)| \geq \varepsilon \\ 0 & others \end{cases} \tag{2}$$

The solution of the so defined optimization problem is solved by the introduction of the Lagrange multipliers α_i, α_i^* (where i=1,2,…,k) responsible for the functional constraints defined in Eq. 2. The minimization of the Lagrange function has been changed to the dual problem [9]:

$$\phi(\alpha, \alpha^*) = \left[\sum_{i=1}^{k} d_i(\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=1}^{k} (\alpha_i - \alpha_i^*) - \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} (\alpha_i, \alpha_i^*)(\alpha_j, \alpha_j^*) K(x_i, x_j) \right] \tag{3}$$

With constraints:

$$\sum_{i=1}^{k} (\alpha_i, \alpha_i^*) = 0, \tag{4}$$

$$0 \le \alpha_i \le C, 0 \le \alpha_i^* \le C$$

Where C is a regularized constant that determines the trade-off between the training risk and the model uniformity.

According to the nature of quadratic programming, only those data corresponding to non-zero $(\alpha_i - \alpha_i^*)$ pairs can be referred to support vectors (nsv). In Eq. 3 $K(x_i, x_j) = \phi(x_i) * \phi(x_j)$ is the inner product kernel which satisfy Mercer's condition [13] that is required for the generation of kernel functions given by:

$$K(x_i, x_j) = \left\langle \phi(x_i), \phi(x_j) \right\rangle \tag{5}$$

Thus the support vectors associates with the desired outputs $y(x)$ and with the input training data x can be defined by:

$$y(x) = \sum_{i=1}^{Nsv} (\alpha_i, \alpha_i^*) K(x, x_i) + b \tag{6}$$

Where x_i are learning vectors. This leads to a SVM architecture (Fig. 1) and are also founded in [9][10][14].

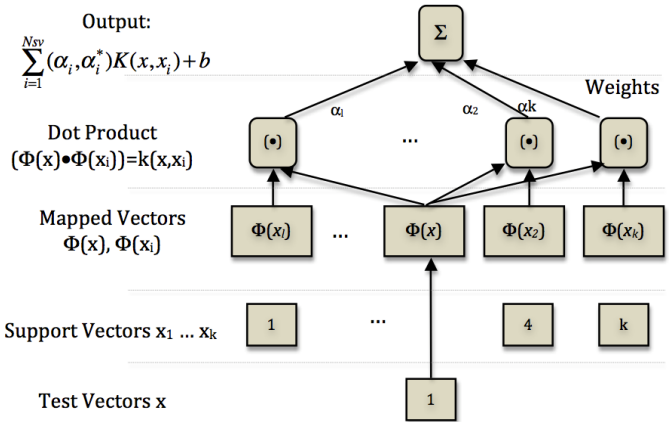


Fig. 1. Support Vector Machine Architecture.

2.2 Kernel functions.

The use of an appropriate kernel is the key feature in support vector applications, since it provide the capability of mapping non-linear data into "feature" spaces that in essence are linear, then an optimization process can be applied as in the linear case. This provides a means to dimensionality the problem properly, but still the results depends of the good selection of a set of training datasets.

The Gaussian kernel function is defined in [11-13] Eq. 7.

$$K(x_i, x_j) = \exp\left(\frac{-\left\| x_i - x_j \right\|^2}{2\sigma^2} \right) \tag{7}$$

The Gaussian kernel process delivers an estimate for the reliability of the prediction in the form of the variance of the predictive distribution and the analysis can be used to estimate the evidence in favor of a particular choice of covariance function. The covariance or kernel function can be seen as a model of the data, thus providing a principled method for model selection [13-15].

A polynomial mapping is a widely used method for non-linear modeling [13][15], defined by:

$$K(x_i, x_j) = \left\langle x_i, x_j \right\rangle^d \tag{8}$$

Unless the used of equation 8 implies an inherit problem, some Support Vector Machines become zeros, therefore is preferable to rewrite the expression as:

$$K(x_i, x_j) = \left(\left\langle x_i, x_j \right\rangle + 1\right)^d \tag{9}$$

In this survey, a Spline kernel is presented as a choice for modeling due to their flexibility. A spline, of order with N knots located at τ_s is expressed by:

$$K(x_i, x_j) = \sum_{r=0}^{\kappa} x_i^r x_j^r + \sum_{s=1}^{\kappa} (x_i - \tau_s)^k + (x_i - \tau_s)^k \tag{10}$$

If κ=1 and the Spline function is defined as

$$K(x_i, x_j) = 1 + \left\langle x_i, x_j \right\rangle + \frac{1}{2}\left\langle x_i, x_j \right\rangle \min(x_i, x_j) - \frac{1}{6}\left\langle x_i, x_j \right\rangle \min(x_i, x_j)^3 \tag{11}$$

Where the solution is a piecewise cubic.

2.3 General Considerations.

Bias Analysis

The inclusion of a bias within the kernel function generally leads to a more efficient implementation and a slightly better accuracy model. Conversely, the solution achieved with an implicit or explicit bias are not the same. This dichotomy emphasizes the difficulties with the interpretation of generalization in high dimensional feature spaces. In this work the explicit bias approach is used.

Free Parameters

Other important issues in support vector applications are the selection of free parameters such as the coefficient of *C,* the value of error ε it determine the margin within which error is neglected and in the Gaussian kernel function the value of variances σ[15-17].

The Quadratic Programming Problem

The SVM training works flawlessly for not too large data points. However, when the number of data points is large (e.g, over 2,000 data points), the Quadratic Programming (QP), problem becomes extremely difficult to solve with standard QP solvers and methods [13-17]. In the study case of this survey, the number of data points is 365, where each data point represents the daily average ofPMx concentration. Therefore the analysis and solving of the QP problem is not considered in the scope of this survey.

According to [11-13], the insensitive loss function is equal to these slack variables, where the ε-insensitive loss function is defined in equation 2, similarly the quadratic ε-insensitive loss function is defined by

$$L_\varepsilon(d, y(x)) = |d - y(x)|_\varepsilon^2 \tag{12}$$

Figures 2a y 2b, show the form of the linear and quadratic ε-insensitive loss functions. The ε-insensitive loss function is attractive because unlike the quadratic cost function, where all data points will be support vectors, then the solution can be sparse. The quadratic loss function produces a solution which is equivalent to ridge regression.

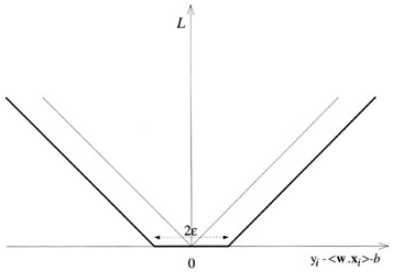


Fig. 2a. The linear ε-insensitive loss for zero and non-zero ε

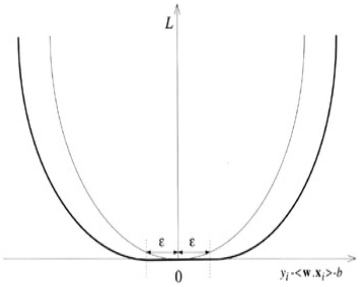
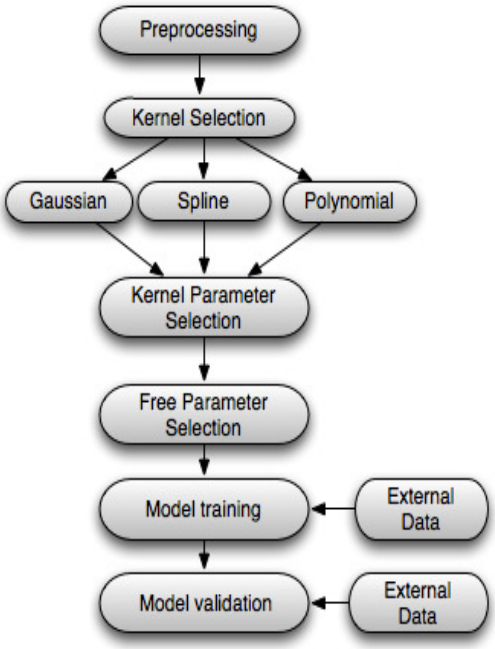


Fig. 2b. The quadratic ε-insensitive loss for zero and non-zero ε

Fig. 2. The ε-insensitive loss functions for zero and non-zero ε

3 Methodology

The proposed Methodology have been taken from [3-4], such this works provides the general steps to make pollutants modeling and predictions by using SVM working in regression mode. In this survey Gaussian, Polynomial and Spline kernel functions are used [3][4][13][14][15], a Gaussian distribution provides a natural representation of the system behavior [13][15]. The aim of this survey is to show the relations between kernel Gaussian, Polynomial and Spline kernels and the obtained SVM models. In order to perform an appropriate design, train, and testing of SVM this article describes a generic methodology based in a review of [3-4]. See Fig 3.



(a) Preprocessing of the input data by selecting the most relevant features, scaling the data in the range [−1, 1], and checking for possible outliers.

(b) Selecting an appropriate kernel function that determines the hypothesis space of the decision and regression function.

(c) Selecting the parameters of the kernel function, in polynomial kernels the degree for polynomials and the variances of the Gaussian kernels respectively.

(d) Choosing the penalty factor C and the desired accuracy by defining the ε-insensitive loss function.

(e) If required, solving the QP problem in l for classification problem and 2l variables in the case of regression problems.

(f) Validating the model obtained on some previously, during the training, unseen test data, and if not pleased iterate between steps (c) (or, eventually b) and (e).

Fig 3. Diagram of the proposed Model Solution

The fundamental reason for considering SVM working in regression mode as an approach for PMx modeling is the non-linear aspect of the application. There is no predetermined heuristic for the choice of free parameters and design for the SVM, many applications appear to be specific, in order to improve the SVM performance trough the automatic adjustment of free parameters. Using SVM on real time applications appear to be rather complex since of the computational demands of the deriving results.

4 Experimental Results

The Support Vector methodology can be applied in the case of regression, maintaining all the main features that characterize the system behavior. A Support Vector Machine in a kernel-induced feature space learns a non-linear function while the capability of the system is controlled by a set of parameters that do not depend on the dimensionality of the space. In this section, a set of results and simulations is presented. This is carried out by using the proposed regression SVM model approach with Gaussian, Polynomial and Spline kernel functions and standard nonlinear data sets of PMx. During 2009, simulations were carried out using the proposed SVM model. The σ values were modified to 1 and 2. Likewise, the ε values were modified to 7, 11 and 13. For every case study, the normalized value C remained content to a value of 100.Also is observed that the error rate of standard SVM varies wildly depending on different values of SVM free-parameters and kernel parameters. Figure 4 shows a summary of the results with the Support vector machine (in red circles), the raw data (black cross) and the behavior of the data (solid black line). These results show that the best results are obtained with σ of 2 and an ε of 13 (figures 4a and 4b) due to the small number of SVMs and small error rate, whilst the worst-case scenario is obtained with a σ of 1 and an ε of 13 (figure 4d), since a large number of SVMs are obtained.

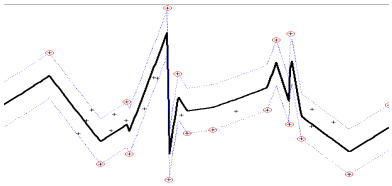


Figure 4a: Prediction of PMx concentration in January using a Polynomial kernel.

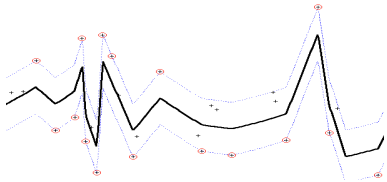


Figure 4b: Prediction of PMx concentration in October using a Gaussian kernel.

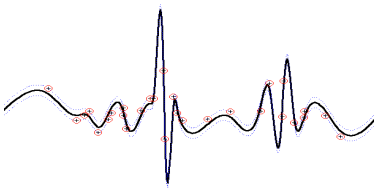


Figure 4c: Prediction of PMx concentration in January using Polynomial kernel.

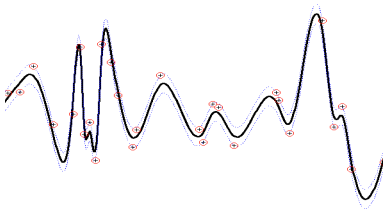


Figure 4d: Prediction of PMx concentration in October using Gaussian kernel.

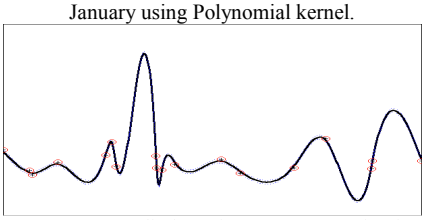


Figure 4e: Prediction of PMx concentration in January using Polynomial kernel.

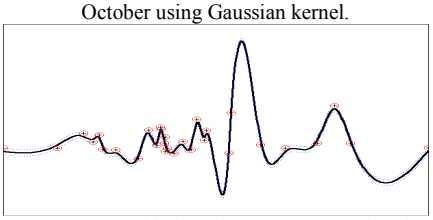


Figure 4f: Prediction of PMx concentration in October using Polynomial kernel.

Figure 4: Prediction of PMx concentration

Table 1. Summary Results

Samples 2007.	**Polynomial kernel**		**Gaussian kernel**		**Spline kernel**	
	No. of SVM	*Incorrect Forecasting*	*No. of SVM*	*Incorrect Forecasting*	*No. of SVM*	*Incorrect Forecasting*
January	17	3	26	1	17	3
February	6	5	18	3	15	5
March	12	4	28	1	24	3
April	12	4	25	0	22	2
May	8	3	23	2	19	3
June	13	1	20	1	15	2
July	18	1	26	1	17	0
August	11	2	24	1	20	3
September	16	1	25	2	20	4
October	18	0	28	1	23	3
November	17	1	24	2	21	5
December	15	3	26	2	24	6

From these results, it can be concluded that for this case study a σ of 1 gives a similar number of SVMs with respects to the number of data points. This exponentially increases the computational cost, making it unfeasible to calculate it.

5 Discussion and Conclusions

This survey has presented a modeling method of the daily atmospheric pollution by applying the support vector machine with Gaussian, Polynomial and Spline kernels functions working in regression mode. The application of SVM has enabled to obtain a good accuracy in modeling pollutant concentration of both PM10 and PM2.5. The methods, techniques and alternatives offered in the SVM field provides a flexible and scalable tool for implementing sophisticated solutions with implied dynamical and non-linear data. It is noteworthy to point that the SVM guarantees this global minimum solution and a good feature of generalization. Furthermore, implementing other kernel functions such as wavelet and hybrid functions may be implemented for future contributions.

Acknowledgement

The authors would like to thank the Air Quality Archive hosted by AEA Energy & Environment, on behalf of the United Kingdom Department for Environment, Food & Rural Affairs and the Devolved Administrations (DEFRA). Also, the authors would like to acknowledge the financial support of the Mexican government via SEP-PROMEP /103-5/09/4100 Project.

References

1. Lall R, Kendall M, Ito K, Thurston G. D.: Estimation of Historical Annual PM2.5 Exposures for Health Effects Assessment, Atmospheric Environment, Vol. 38, pp. 5217--5226. (2004)
2. Malcolm A.L, Derwent R.G, Maryon R.H.: Modelling the Long-range Transport of Secondary PM10 to the UK, Atmospheric Environment, Vol 34, pp. 881--894. (2000)
3. Vega, E. Reyes, G. Sanchez, et al.: Basic statistics of PM2.5 and PM10 in the atmosphere of Mexico City, The Science of the Total Environment, Vol. 287, pp. 167--176. (2002)
4. Querol X., Alastvey A., Ruiz C. R., et al.: Speciation and origin of PM10 and PM 2.5 in selected European cities, Atmospheric Environment, Vol. 38, pp. 6547--6555. (2004)
5. Harrison R.: Studies of the source apportionment of airborne particulate matter in the United Kingdom, Journal of Aerosol Science, Vol. 31, supplement 1, pp. S106--S107. (2000).
6. F. Wang, D.S. Chen, S.Y. Cheng, J.B. Li, M.J. Li, Z.H. Ren, "Identification of regional atmospheric PM10 transport pathways using HYSPLIT, MM5-CMAQ and synoptic pressure pattern analysis", Environmental Modelling& Software 25 , pp. 927-934 (2010).
7. Collazo-Cuevas J.I., Aceves-Fernandez M.A., Gorrostieta-Hurtado E., Pedraza-Ortega J.C., Sotomayor-Olmedo A.2, Delgado-Rosas M., "Comparison between Fuzzy C-means Clustering and Fuzzy Clustering Subtractive in urban air Pollution", CONIELECOMP 2010, 20th International Conference on Electrical Communications, pp. 174-179.
8. Aceves-Fernández M.A., Sotomayor-Olmedo A., Gorrostieta-Hurtado E., Pedraza-Ortega J.C., Tovar-Arriaga S., Ramos-Arreguin J.M., Performance Assessment of Fuzzy Clustering Models Applied to Urban Airborne Pollution, CONIELECOMP 2011, 21th International Conference on Electrical Communications, Accepted, In press (2011).
9. Lenschow P., Abraham H.J., Kutzner K., Lutz M., Preuß J.D., Reichenbächer, W.: Some ideas about the sources of PM10, Atmospheric Environment, Vol. 35, supplement 1, pp.S23--S33. (2001)
10. Vapnik, V.: The Nature of Statical Learning Theory. Springer-Verlang, New York. (1995).
11. Vapnik, V., Golowich, S., Smola A.: Support method for function approximation regression estimation, and signal processing. Advance in Neural Information Processing System 9.MIT Press, Cambridge, MA. (1997)
12. Schölkfopf B.: Smola A. J.: and Burges C.: Advances in Kernel Methods –Support Vector Learning. Cambridge, M.A.: MIT Press. (1999)
13. Osuna, E., R. Freund, F. Girosi.:Support vector machines: Training and applications. AI Memo 1602, Massachusetts Institute of Technology, Cambridge, MA 44. (1997)
14. Cristianini, N., Shawe-Taylor, J., An introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, Cambridge, UK (2000)
15. S. Osowski and K. Garanty, "Forecasting of the daily meteorological pollution using wavelets and support vector machine," Engineering Applications of Artificial Intelligence, vol. 20, no. 6, pp. 745-755, September (2007)
16. I. Sapankevych and R. Sankar, "Time series prediction using support vector machines: A survey," Computational Intelligence Magazine, IEEE, vol. 4, no. 2, pp. 24-38, (2009)
17. W. Lu and W. Wang, "Potential assessment of the support vector machine method in forecasting ambient air pollutant trends," Chemosphere, vol. 59, no. 5, pp. 693-701, April (2005).

Intelligent Decision Support System based on Data Mining

Alberto Ochoa-Zezzatti[1], Sergio Enriquez[2], Arturo Elías[2], Rubén Jaramillo[3], Ricardo Pérez[3], Alfredo Jiménez[3], Ángeles Duhart[3] & Cyntia Araiza[1]

[1]Centro de Investigaciones Sociales, UACJ.
[2] Universidad Autónoma de Aguascalientes.
[3] CIATEC Centro Conacyt

Abstract. The making decision process involves a lots of information to reduce as much as possible the probability of errors. We propose an intelligent tool for decision making with ability to organize large amounts of data and represent them in many forms besides an easy interpreting for users. The tool allows the discovery of hidden patterns and predicts the tendencies through of data mining looking for improve of the making decision process.

Keywords: Intelligent tool, decision making, data mining.

1 Introduction

The decision making is a process common in the organizations whereby several alternatives are considered to provide a solution to many kinds of problems. The range of decision making process is so wide that the majority of the occasions involve several areas of knowledge. Most important to take a decision is the information related to the subject at issue. The amount of information is very relevant; more information implies enhancing in a successful decision making. The organizations are generating of great amounts of information, but a big problem is how to organize and summarize these amounts of information for be useful and easy to interpret. An intelligent tool for the decision making allows the user to analyze, to organize and to present/display the information of way summarized and easy to interpret, and thus the takers of decisions can understand better the context or situation and so the process of decision making improves.

2 Intelligent Decision Support System

2.1 Data Mining

The data mining techniques are based on great amounts of related data and allow discovering information hides as well as predict their tendencies. The data mining is a passage in the process of knowledge discovery. The mining of data involves the use of sophisticated tools of data analysis, which can include statistical models like the

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 123-131
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

time series seen in [3], linear regression [4], mathematical algorithms and methods of machine learning [1]. As much this technique as others of Artificial intelligence can be those that contribute more to the future of some Decision Support Systems (DSS's) [2]. The tool presented in this article, uses techniques of data mining to analyze the information and to present/display it to the user, so that it is evident recognize certain existing features in the data and then predict their tendencies, which is very useful in the decision making process.

2.2 Decision Support System

The decision making is a process that is daily carried out in all type of activities by all the people. Particularly in the enterprise surroundings, this process acquires great importance since, generally the success or failure of a company is linked to right or incorrect decision making. The process of decision making is carried out to different levels with the purpose of to achieve certain objectives and where the time also is a factor to consider, the decisions may varies depending if they are short term or long term decisions. Additionally it consists of several steps and exist different models to explain them. The Fig.1. shows some common steps in the process of decision making: Generally the systems of support for the decisions have like intention fundamental to support and to facilitate this process, through the opportune and reliable acquisition of excellent information [5].

2.3 Statistical Techniques

The statistical techniques allow realize a descriptive analysis of the data, make predictions and obtain a better decisions making. Different statistical techniques were applied from the data (formed through surveys, which are used to feed the system), like the linear regression, to define an equation or function that allows us to consider the average stature of one second generation by the sex of the individuals and the statures of their grandparents as paternal and maternal.

3 System Development

3.1 Methodology

The tool that we propose was developed in language JAVA by means of the IDE (Integrated Development Environment) of Borland, JBuilder. The handling and presentation of data were devised to diverse functions within the system for. It is possible to load different archives to work with the data, to modify them and to present/display them by regions, among others options. Software re-engineering was applied in some re-usable modules, to adapt these to the requirements, for it was begun with the design of the system, being the function it bases, the load of a data file from which the other operations of the system can be used.

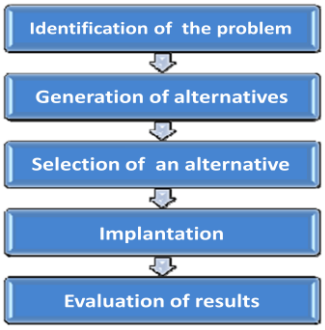


Fig. 1. *The process of Support Decisions*

4 Development Tool

This Intelligent Tool for Support Decisions (HITODE) displays to the user the information organized and summarized, in a clear way with certain characteristics of groups of people associated with properties in the collected data. Data set was obtained by surveys to a group of students of bachelor level. These data included questions on their musical, literary tastes, sports, purchases by Internet and origin municipality enters others. This data base was used to prove the functionality of the system. HITODE is composed by seven modules, which appear in the Fig. 2.

The following six modules are based on the input data that feed the system, the modules are: the file load, the graphic display of data, the mapping of regions, the reports in sequence hierarchic, the data modification and the parameters or faces of Chernoff [6]The two remaining modules, calculation of statures and migratory model were developed to divide the data of the survey and generate an equation that calculates its respective values, that are independent modules of the new data input. For the modules based on input data, the first step consists in load a data file that contains registries with n attributes. The text file requires a specific format to permit to be read by the system. The file must be in format CSV (Control On Value) and requires three lines of headed, first line has the names of the attributes, second has the amount of possible options for an attribute and finally third line contains the possible options for each attribute. When the data is loaded, the file load module qualifies its functions or buttons, areas of text, among others, allowing it to be used by the user.

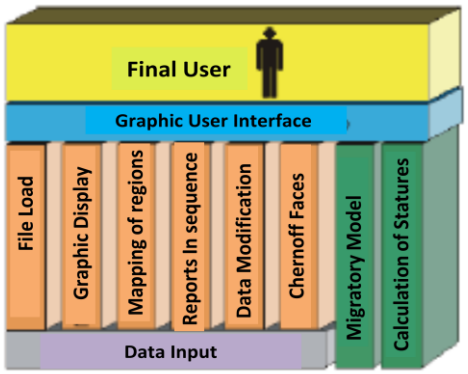


Fig. 2. Internal Structure of HITODE

The module of graphic display of data only supports attributes that are different to chain or string types. For example, the sex attribute (Fig. 3.) may be plotted because it has a defined dominion {Masculine, Feminine}. In this way, with the graphics module, it is possible show through a bar chart the amount of men and women registered in the data file of entrance.

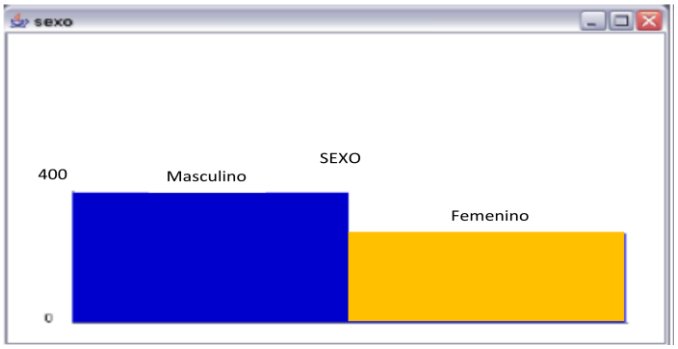


Fig. 3. HITODE Interface: Graphics Module

The module of mapping of regions, consists of presenting/displaying the information of certain attributes in a map divided by regions (Fig. 4.), where the color represents the highest percentage of occurrence of an attribute and a smaller circle in each region represents the importance of the second attribute of greater occurrence. For example when selecting the attribute "cinema", HITODE will present/display the preferences of cinematographic sort by region. Allowing the user visualize on the map predominant cinematographic sorts in each region. This function turns out particularly useful in a market study to know preferences the consumers, displayed by region. The module of faces or parameters of Chernoff displays the data by regions by means of the propose technique in [6]. It consists of using faces whose parts represent certain attributes of a certain region. For example, if a region is characterized to have a low index of unemployment, it would have a smiling mouth, and otherwise it would have a sad face. It is not possible represent

all the attributes by means of the parameters of Chernoff, only attributes whose dominion is numerical or dichotomizing, because another type of attributes would not show excellent information that can be represented through this parameters, that means if for example exists an attribute called Literature, where its dominion is {Poetry, Terror, Science fiction, Suspension, History} when trying to be presented with the size of the eyes of a face of Chernoff, would not be easy to interpret what they mean great eyes or small eyes. Although this module was realized of independent way, the original idea was taken from [6].

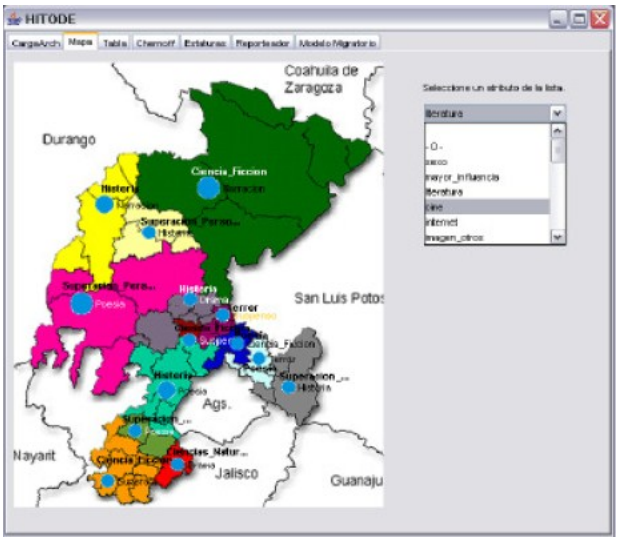


Fig. 4. HITODE Interfaces displays the attributes by regions.

The module reports in sequence hierarchic allows making a filtrate of information based on a combination of certain attributes. For example, to the first step for starting is select a region, then choice an attribute like for example the Literature type, immediately unfold options of the different types from Literature that exist in the data input together with sex, the module has the possibility of filtering by sex: Men or women solely and even both. An example of this combination of attributes could give the results of sex "masculine" that belong to "region 3" where its "literary preference" is the kind of "science fiction".

In the module of data modification, modifications from the data input file may be doing either to add new registries or to correct to the existing ones. The data appear in the form of table where each of attributes is organized by columns as it showing in Fig. 6.

The module of the migratory model is independent of the data input file of the system and makes the calculations of migration by regions, using a function generated from of data collect through surveys.

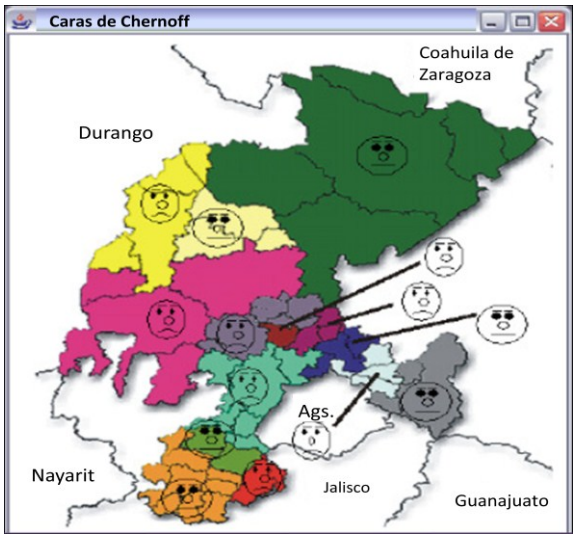


Fig. 5. HITODE Interface displays the attributes by means of Chernoff Faces

The module of statures calculation uses a function generated from statistic analysis and shows the stature average for the last two generations of the grandparents from both paternal and maternal families. Furthermore for the accomplishment of the calculation, sex of individual must be chosen. The generated equation (1) is showing follow:

$$\text{Statures}=129.5713+13.099(\text{Sex})+0.168(\text{Prom\_Abue\_Mat})+0.026(\text{Prom\_Abue\_Pat}) \quad \textbf{(1)}$$

Here, the Sex attribute takes values from 1 for masculine and 0 for feminine and Prom_abue_Pat and Prom_abue_Mat are the average of the statures of the maternal and paternal grandparents respectively. The confidence intervals and calculate of prediction from equations (2) and (3) that can to find in [4] (see this reference for more information). These values are displayed in the system followed of average stature and the symbol "±".

$$\hat{Y}_0 \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 X_0^T (X^T X)^{-1} X_0} \qquad (2)$$

$$\hat{Y}_0 \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 (1 + X_0^T (X^T X)^{-1} X_0)} \qquad (3)$$

Although the modules of the migratory model and the statures calculation working on independently new data input, they can be implemented to work with different data files. These modules were developed with the purpose of display useful and simplified information of the input data file obtained on the survey realized over the tastes and preferences of the Zacatecans young people and that contain uncommon attributes in other data files.

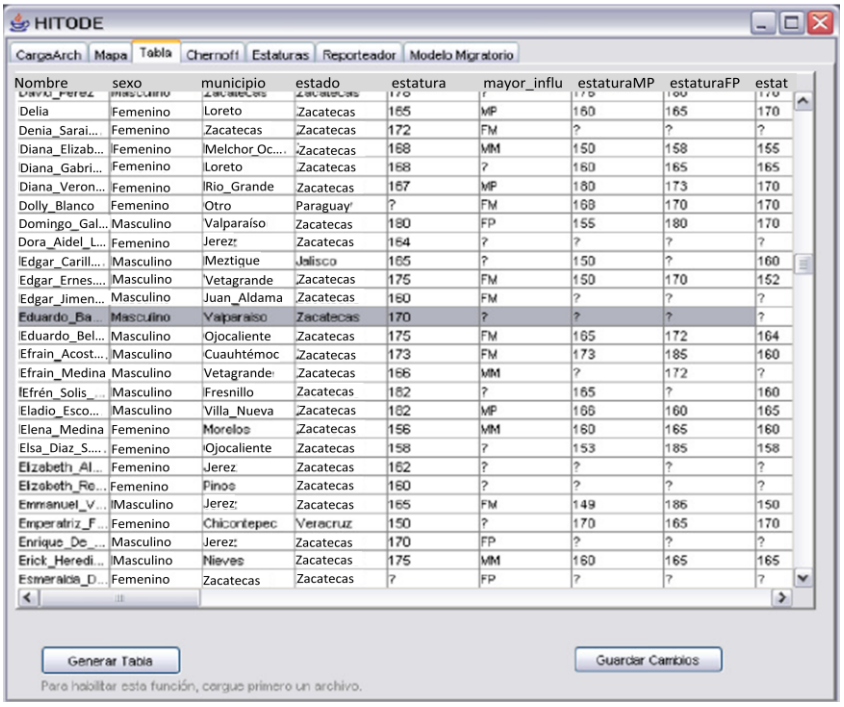


Fig. 6. HITODE Interface: Module of data modification

5 Results

The objective of this survey was to know information on Zacatecan youth to design strategies of businesses for this segment of market in specific. When using HITODE Tool with the input data file generated from this survey, was possible analyze the data of diverse forms and thus understand of simpler way the existing relations between the attributes including in the survey.

5.1 Future Research

When applying the statistical techniques, we were with the problem that high entropy in the collected data and due to this existed, was very complicated to apply some of these techniques, reason why it was necessary to make a cleaning of the data to be able to work with them. On the other hand, due to great amount of data lost percentage of prediction (security) that it was obtained with techniques as the simple linear regression were not very high when trying in particular to calculate the statures average of an individual from the statures of its grandparents, reason why it tries to realize another compilation of new data being looked for to improve the quality of these and this way to develop a function within HITODE that allows to make statistic analyses with other data. Finally HITODE is tried to migrate to Web, being implemented using XML

(Standard Markup language), to adapt the information of a comprehensible and easily analyzable way.

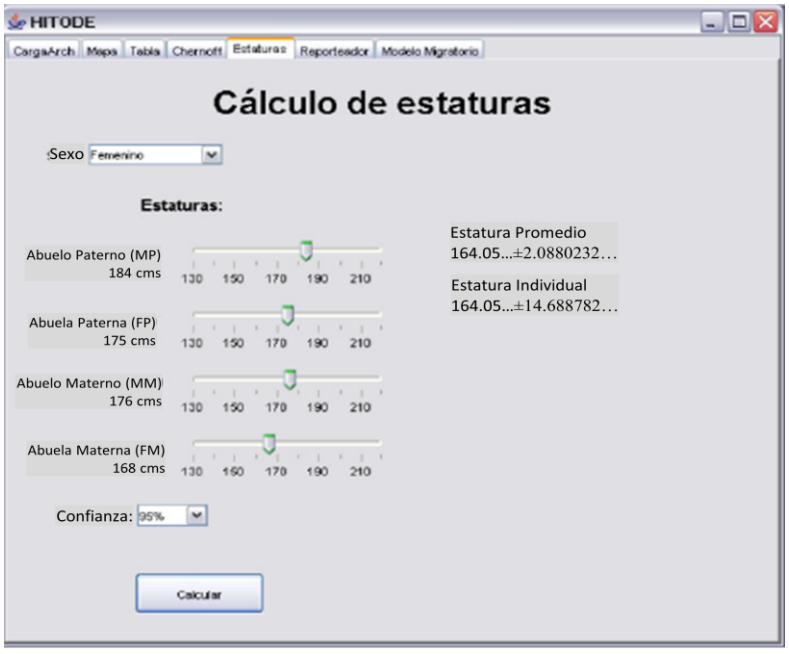


Fig. 7. HITODE Interface: Stature Calculation

6 Conclusions

An intelligent tool for the decision making allows us to compare and to corroborate information that is not so easy to visualize when great amounts of data are had. The system that we presented helps the ones in charge of decision making to visualize these data of way graphical, and divided by regions which is particularly useful at the time of making studies of market mention an example.

Acknowledges

The first author wishes to be thankful to the Drs. Hal, and Iztebegovic by its collaboration in the development of the module of the calculation of statures of the presented system as well as his collaboration for the revision of the article. Also it wishes to be thankful for to the Dr. Mendizhavili by his aid for the development and the revision of HITODE Tool and the present paper and also thank to A. Gutierrez for his support in the revision on this text.

References

[1] Seifert, Jeffrey W. "Data Mining: An overview". Congressional Research Service - The Library of Congress. 2004.

[2] Aiken, M. Liu Sheng, O. Vogel, D. "Integrating expert systems with group decision support systems". ACM Transactions on Information Systems (TOIS). 1991.

[3] Keogh, E. Pazzani, M. "Scaling up dynamic time warping for datamining applications". Conference on Knowledge Discovery in Data. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. 2000.

[4] Montgomery, D. Peck, E. Vining, G. *Introducción al Análisis de Regresión Lineal*. CECSA 3ra ed. México D.F. pp. 92 –98. 2004.

[5] Cohen Karen, D., Asín Lares, E. *Sistemas de Información para los negocios*. Mc Graw Hill, México D.F. pp. 194 – 199. 2005.

[6] Chernoff H. "The Use of Faces to Represent Points in KDimensional Space Graphically", in Journal of the American Statistical Association, Vol. 68, No. 342 (Jun., 1973), pp. 361-368.

[7] Hughes, D. *Mercadotecnia: planeación estratégica*. Ed. Addison-Wesley iberoameric. México, 1986. pp. 59 – 64

A Reputation Scheme for Peer-to-Peer Media Streaming

Francisco de Asís López-Fuentes[1]

Networks and Distributed Systems Group
[1]Departamento de Tecnologías de la Información
Universidad Autónoma Metropolitana
Av. Constituyentes 1054,
11950 México, D. F., México
{flopez}@correo.cua.uam.mx

Abstract. A challenge in peer-to-peer media streaming systems is how to select good peers in order to realize high quality streaming sessions. The selection of good peers can offer a manner to improve the quality of service via an optimal search or an efficient content delivery. This paper presents an approach for evaluating the participating peers based on their donated resources and on their behavior. This approach uses a reputation/incentives model to isolate the misbehaving peers and the non-cooperating peers, as a way to improve the system performance. Every peer builds its best path using a best-neighbor policy within its neighborhood. The search is based on the best path. A structure based on reputation/incentives policies is used by the supplying peer as a way to assign its outgoing bandwidth to the requesting peers during the media transmission phase.

Keywords: reputation, incentives, peer-to-peer systems, media streaming.

1 Introduction

During the last years, content delivery over the Internet has gained significant popularity. For example, several applications such as TV over IP, streaming and multimedia live streaming require content delivery from one-source to multiple receiver-nodes. On the other hand, peer-to-peer (P2P) networks have attracted the attention from the research community who find in these systems a fast and efficient way to deliver movies, music or software files. A P2P communication infrastructure is formed by a group of nodes located in a physical network. These nodes build a network abstraction on top of the physical network, known as an *overlay network*, which is independent of the underlying physical network with regard to the P2P procedures. An important advantage of P2P systems is that all available resources are provided by the peers. In a P2P system each peer can take the role of both, a server and of a client at the same time. During media distribution, peers contribute their resources to relay the media to others. Thus, as a new peer arrives to the P2P system the demand is increased, but the overall capacity too. This is not possible in a client-server model with a fixed number of servers.

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 132-141

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

A proper selection of peers can lead to the attainment of a good quality of service in terms of a faster search, and a faster distribution of content; but, P2P systems can be affected by misbehaving (or *free-riding*) peers, which reduce the system performance. The *reputation management systems* (RMS) are methods which alleviate this problem [3, 6, 7] through the proposal of appropriate peers, leading, with this, to a natural isolation of misbehaving or non-cooperative peers. A RMS system allows individual peers to rate one to each other according to their past experience with each other. Once a peer has been rated, its rating can be used by other peers to find the best sources of good and authentic content, keeping, at the same time, the effects of malicious peers on the network to a minimum. The proposals for these systems include solutions for the management of trust and the computation or reputation. RM Systems provide a way for building trust without trusted third parties in P2P networks [6].

This paper proposes a method for reputation computation, which involves the concept of incentives. The interest of mixing these two characteristics in this method is the following. First, peers with high reputation can cooperate to make an optimal search or a better content delivery. Second, an incentives-system can encourage the collaboration and exchange of data between peers [1, 5]. Finally, the isolation of misbehaving or non-cooperative peers can avoid the degradation system performance.

The remainder of this paper is organized as follows. Section 2 presents the model and its assumptions. Then, the protocol of the method is presented in Section 3. Section 4 describes the evaluation of our proposed model and presents the results. Section 5 concludes the paper.

2 The Proposed Method

Locating a content-supplier does no guarantee that the service of this supplier will satisfy the user [2], because some misbehaving peers may offer false information in order to maintain a cooperation impression. In order to minimize the effects of misbehaving peers these are to be detected and isolated from the system. Reputation and incentives strategies have been used in several approaches but in separate ways. The concept of reputation is used in several systems such as the online auction system eBay. In eBay's reputation system, buyers and sellers can rate each other after a transaction, and the overall reputation of a participant is the sum of these ratings over the past 6 months. The authors in [3, 9, 10 and 11] are proposing reputation systems with the purpose to ensure that peers obtain reliable information about the quality of the resources they are receiving.

The model here presented makes use of a special peer called manager-peer, which manages the reputation of all peers in the system, and considers that each one of the other peers in the system has information-reputation only of its neighbor-peers. Each peer has local table which keeps a reputation score. Every peer exchanges its local table with any others peers located at no more than 2 hops away. If a peer can be reached, from another one, in just one hop, then it is said that there is a direct link between these peers. If the number of hops is not 1 but 2, then it is said that there is an indirect-link between these peers. The presented reputation method, with incentives, consists of two parts: 1- The reputation model, and 2- The incentives model.

2.1 Reputation Model

The reputation model considers that all peers contribute with their resources to the system. The method uses two components to obtain the average reputation-score in every peer, which are: its capability and its behavior. The first component evaluates the resources of the participating peer: upload capacity, processing capacity, memory, storage capacity and number of shared files. The initial weight for each donated resources can be agreed upon by the users. The second component evaluates the behavior of the participating peer in a cooperation environment, assigning the peer a cheating level a transient level. The cheating level is assigned considering that a peer is cheating when it supplies a wrong content or when it serves with fewer or smaller resources with regard to those promised. The transient level for a peer is determined by the average length of time the peer remains in the system (service-time) and by the average length of time that it takes for the peer to return to the system after it has left. Users could be satisfied when they received content from peers with big resources and good behavior; in the other hand, users could have a bad experience when the involved peers offer low bandwidth, high error-rates, limited processing resources, or frequent disconnections. Every component in the reputation-scheme contributes with its weight, when building the final score.

For example, during a streaming session of a participating peer, the weight of the donated-bandwidth resource of the peer could be bigger than that of its available-storage resource. Initially, the method's reputation score of a peer p_i is based only on its donated resources. As time passes by, if the peer is still connected, it doesn't cheat and its stability and availability is good, then its behavior reputation is increased; otherwise its behavior reputation is to decrease.

The evaluation of the behavior-reputation considers that a transaction made by any peer can be either, performed correctly or not. Our behavior reputation scheme is based on a reputation scheme introduced in [8]. Peers interact using a reputation-approach. A complaint message is evaluated in every peer and in the manager-peer. Any peer, in order to compute the reputation of another peer, evaluates the experience of that peer's neighbors. This is indeed a distributed reputation system, in the neighborhood and in the system. This reputation is usually based on an aggregate of the feedback ratings issued by the diverse peers [6]. The manager-peer is consulted by a peer who wants to know the reputation of peers outside its neighborhood.

The behavior-reputation scheme has two scenarios. First scenario is shown in figure 1. In this scenario, peer p_1 interacts with peer p_2, p_1 may rate the transaction as satisfactory $(t(p_1,p_2)= 1)$ when a retrieval process with p_2 is successful or unsatisfactory $(t(p_2,p_1) = -1)$ if the file is no authentic, if p_2 give false information about its resources or if the download is interrupted. Thus, when peer p_2 is cheating in a retrieval process with peer p_1, a complaint (-1) is sent by p_1 to the global reputation in the manager peer and another complaint is recorded in its local table and distributed on its local neighborhood. A *reputation matrix* is built based on the total number of peers in the neighborhood or in the network. Every peer records the reputation score in the *reputation matrix*, as a local table, while the manager peer records the *reputation matrix* as a global table.

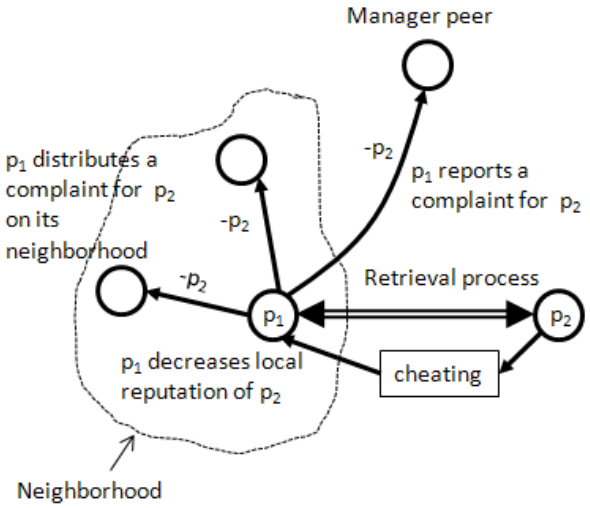


Fig. 1. First scenario for our behavior reputation scheme

Second scenario considers when a new peer joints to the system (see Figure 2). In this case, all entries of new peer are undefined, but these are update as the peer interacts with each other. Every peer updates the reputation of its local reputation matrix, while the reputation of remote peers can be derived from the manager peer.

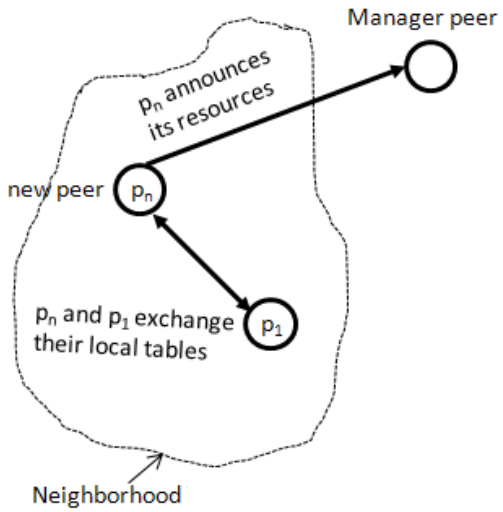


Fig. 2. Second scenario for our behavior reputation scheme

To compute the behavior reputation, the protocol periodically runs an update process in order to update the network. Then, a reputation agent updates the reputation score and the incentives of every peer based on its behavior (cheating level

and transient level). Initially, the global reputation score is based on resources only, and the behavior reputation is valued in 0. The behavior reputation score of a peer is increased if it maintains a good service time or it does not cheat, on the other way the behavior reputation score is decreased. In our model the reputation range is between 0 and 10. Where 0 indicates the minimum reputation score and 10 defines the maximum reputation score. We consider the peer status (UP/DOWN/CHEATING) to define several scenarios and to update its reputation score, whenever round expires and the update process start. If the peer status is UP and it does not cheat, the reputation agent computes the average number of rounds that a peer remains connected to the network and its behavior reputation score. If the peer status is DOWN and it does not cheat, we need to determine how long it is DOWN. If time is greater than 4 rounds, then the peer will be punished in 2 rounds when it returns to the P2P network (UP status). During the first round its reputation score will be zero and in the second round its reputation score will be based on resources only. When the peer status is CHEATING, its reputation is decreased to 0 in all its neighbor peers and in the manager peer. All peers isolate the cheating node, and they do not send, forward or receive any messages or packets from it.

2.2 Incentive Model

An incentives model needs to define some rules to motivate peers to contributes more resources and avoid the no-cooperation in the system [12, 13]. For our incentives system we assume the following rules:

- A peer cooperates with another peer based on its generosity factor [4].
- Peers with high reputation are allocated close to the source, forming rings. The internal ring will be close to the source and it will have the highest reputation.
- Peers with high reputation receive a high priority to upload contents during a contention.

We adopt an approach based on the game theory to address the no-cooperation problem in the system. In particular, we use a choking algorithms model to capture the essential tension between individual and social utility, asymmetric payoff matrices to allow asymmetric transactions between peers, and a learning-based population dynamic model to specify the behavior of individual peers, which can be changed continuously. So, our approach rewards cooperation and therefore considers upload and download rate, like a generosity factor, to translate cooperation to earn benefits or loss it to peers which cooperate or not.

Our approach considers the upload and download rate and rewards the cooperation using a generosity factor. The peer's cooperation is translated to benefits via the generosity factor. This factor measures the benefit that a peer provided relative to the benefit it obtained. On this way we try to avoid that the system collapses when we have peers consuming more services than they provide [4]. The generosity resumes the General Prisoner's Dilemma for an asymmetric payoff matrix [4, 5]. Let Up_i and Dp_i the provided service and the obtained services by the cooperating peer p_i,

respectively. A unit of provided service is a packet given successful, while an unit of obtained service is a packet received successful. Then, the p_i's generosity is given by

$$G(p_i) = Up_i / Dp_i \tag{1}$$

To reach an effective cooperation in the neighborhood, every peer uses its own generosity as a measuring stick to judge its peer's normalized generosity [4] expressed by:

$$G_{p_j}(p_i) = G(p_i) / G(p_j) \tag{2}$$

where $G(p_j)$ is the p_j's generosity and $Gp_j(p_i)$ measures p_i's generosity relative to p_j's generosity.

Using its reputation table each peer builds a hierarchical structure based on rings. We call them, reputation rings. The reputation rings is used by a supplying peer during the download phase. Peers with high reputation are allocated close to the supplying peer, while the peers with low reputation are allocated in the extern rings and distant to the source. A requesting peer with high reputation receives a high access priority (download) from the supplying peer during a contention. The number of reputation rings and its reputation thresholds are values that the system architect is free to set. A requesting peer p_i is allocated in a reputation ring R_i based in its reputation score. The incentives are distributed among peers based on this hierarchical structure. Peers close to the source receive more incentives than peers far from it. In our approach, the bandwidth is considered the main incentive. We distributed a different source bandwidth portion among the reputation ring. Thus, peers allocated into the internal reputation rings receive a bigger bandwidth portion from the supplying peer than peers allocated into the external reputation rings. We give a different percentage of incentives to every reputation ring. These percentages can be arranged by users of the system. With the reputation and the generosity factor, each source peer encloses its requesting peers in reputation/generosity rings.

Our proposed model updates the reputation of all enrolled peers running periodically a reputation process, which can be activated by every participating peer or by the manager. This process collects information about the updated resources and behavior of each peer within its neighborhood, and updates the local table. Our protocol operates in two phases, the first is called search phase and the second is the download or streaming phase.

3 Evaluation

We evaluate our reputation mechanism with incentives using a simulator based on java. A random topology is generated and different traffic scenarios are used. Thus, the simulation considers two mainly files. On the one hand the topology, and on the other hand, the traffic scenario file. Our experiments use a random topology with 40 nodes. Additionally, the resources donated by each peer such as CPU, bandwidth, memory and storage capacity are recorded. The initial reputation is based on resources only. Also, a manager peer is defined in the topology.

Regarding to the traffic scenario, we develop a generator of traffic that generate a traffic scenario based on the real internet traffic. We assume that the peer arrives and leaves (UP/DOWN) to the system following a Poisson distribution. Also, a Poisson distribution is used to model the time that a peer remains in the system. Query popularity (number of queries and time between queries) is modeled following a Zipf distribution. Every peer selects a set of videos in a random uniform way. In the beginning, there are no relationships among peers in the system. The simulator reads the topology and the traffic scenario files and traffic events such as UP's, DOWN's, CHEATING's, Queries are generated during the simulation. Every node initializes its local table with information about its neighbors and defines its shared files. The reputation mechanism collects the donated resources by each peer and monitors its behavior while the incentives system distributes the incentives based on the rules described in section 2.2. To simulate the incentives over the reputation rings, we established our hierarchical structure in 5 rings, and the reputation is distributed among them. We fix a bandwidth portion of 30%, 25%, 20%, 15% and 10% for the rings 1, 2, 3, 4 and 5, respectively. In our simulation, the peer's bandwidth is considered the main resource and its changing behavior is considered in both phases. Other resources such as storage, CPU, memory and number of files retain its initial values in the system.

The results show how the system performance is greatly affected when the rate of the misbehaving peer is increased. We measure the system performance as the number of successfully downloaded contents. We simulate the system performance using different percentage of misbehaving peer such as 0%, 15%, 30%, 45%, 60% and 75%. Figure 3 shows how the system performance is affected by the misbehaving peers. Here, we can see that the number of download content is decreased proportionally to the number of misbehaving peers. We use two alternatives, in the first alternative we sent a query using a best path only, and the second alternative each peer allocated within the query route uses its two best paths.

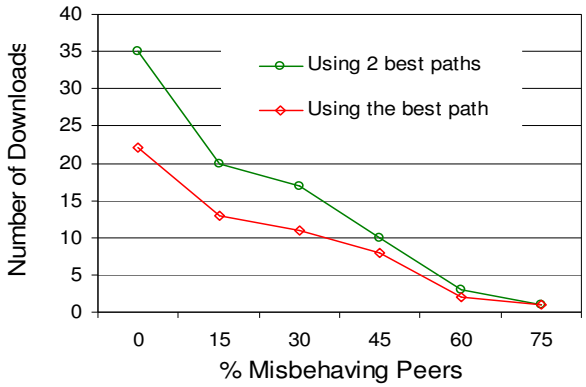


Fig. 3. System performance is affected by the misbehaving peers

Contrary, Figure 4 depicts the system performance when the cooperation between peers is increased. We can see that more files are fully downloaded when the cooperating peer percentage is increased.

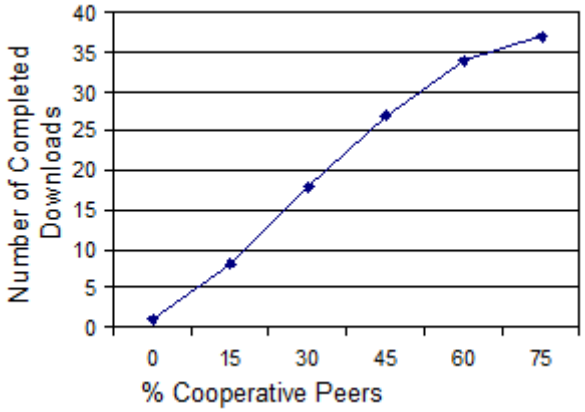


Fig. 4. Cooperation between peers increases the system performance.

Finally, our mechanism isolates misbehaving peers from good peers. On this way, any misbehaving peer cannot process, send or forward any content or query. Either the good peers cannot send queries to the misbehaving peers or receive requests from them. Figure 5 shows how a system without reputation allows download content from the misbehaving peers increasing the probability of having a greater number of corrupted contents. In this scenario we can see that misbehaving peers give bad ratings and good peers always give good ratings in the system. However, in the real electronic communities correcting the malicious peer's behavior is a hard task. Instead of correcting each such malicious peer, we need to minimize its impact in the system performance.

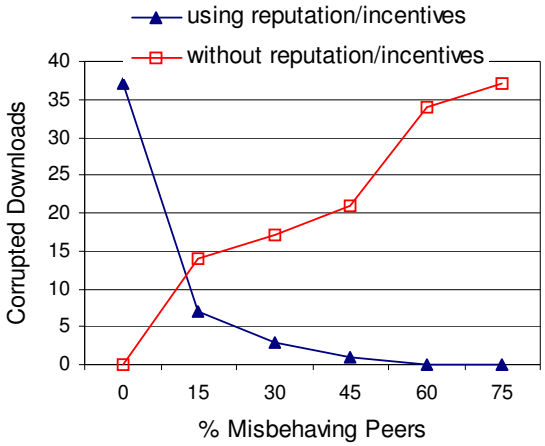


Fig. 5. Comparison of rejected corrupted files

4 Conclusions

In this contribution, we have proposed and evaluated a reputation mechanism with incentives for a P2P system. Most of the reputation systems consider correction of malicious peers by giving incentives for positive feedbacks. However, in our proposed model isolates misbehaving peers from good peers, and incentives are only used to gain most cooperation in the system. We show how the presence of misbehaving peers reduces the system performance. Our results show that using our proposed mechanism, free riding can be reduced, because the non-cooperating peers are eliminated from the system. Also, the reception of corrupted files from the misbehaving peers is eliminated. Using the reputation rings a peer can distribute its upload capacity among good peers based on its reputation score. The possible extensions for this work could be in the direction of large P2P streaming systems, where fully distributed and scalable schemes must be improved. Finally, we will perform more evaluations in order to compare our proposed scheme with other methods available in the literature.

References

1. Silverston, T., Fourmaux, O., Crowcroft, J.: Towards an Incentive Mechanism for Peer-to-Peer Multimedia Live Streaming Systems. In: Th IEEE Eight International Conference on Peer-to-Peer Computing, pp. 125--128, 2008, IEEE Press, New York (2008).
2. Marti, S., Garcia-Molina, H.: Limited Reputation Sharing in P2P Systems. In: 5th ACM Conference on Electronic Commerce, pp. 91--101. ACM press, New York (2004).
3. Kamvar, S. D., Schlosser, M. T., García-Molina, H.: The EigenTrust Algorithm for Reputation Management in P2P Networks. In: 12th International World Wide Web Conference, pp. 20-24, ACM press, New York (2003).
4. Feldman, M., Lai, K., Chuang, J., Stoica, I.: Robust Incentive Technique for Peer-to-Peer Networks. In: 5th ACM Conference on Electronic Commerce, pp. 102--111, ACM Press, New York (2004).
5. Blanc, A., Liu, Y. K., Vahdat, A.: Designing Incentives for Peer-to-Peer Routing. In: 25th IEEE International Conference on Computer Communications, pp. 374--385, vol. 1, IEEE Press, New York (2005).
6. Prasad RVVSV, Srinivas, V., Kumari, V. V., Raju, KVSVN: An Effective Calculation of Reputation in P2P Networks. In: Journal of Networks, Vol. 4, No. 5, pp. 332--342, Academic Publisher, West Bengal, India (2009).
7. Su, X., Dhaliwal S. K.: Incentives Mechanisms in P2P Media Streaming Systems. In: IEEE Internet Computing, pp. 74--81, IEEE Press, New York (2010).
8. López-Fuentes F. A. and Steinbach E.; Architecture for Media Streaming Delivery over P2P Networks. In: Springer-Verlag, LNCS 3563, pp. 72-82, Berlin Heidelberg, Germany (2005).
9. Cornelli, F., Damiani, E., Vimercati, S.D.C.D., Paraboschi, S., Samarat, S.: Choosing Reputable Servents in a P2P Network. In: Proceedings of the 11th World Wide Web Conference, Hawaii, USA (2002).
10. Gupta, M., Judge, P., Ammar, M.: A Reputation System for Peer-to-Peer Systems. In: Proceedings of ACM NOSSDAV 2003, Monterey, CA (2003).

11. Aberer, K., Despotovic, Z.: Managing Trust in a Peer-2-Peer Information System. In: Proceedings of the ACM 10th International Conference on Information and Knowledge Management, New York, USA (2001).
12. Habib A., Chuang, J.: Incentive Mechanism for Peer-to-Peer Media Streaming. In: Proceedings of the 12th IEEE International Workshop on Quality of Service, Montreal, Canada, IEEE Press, New York (2004).
13. Lai K., Feldman M., Stoica I., Chuang J.: Incentives for Cooperation in Peer-to-Peer Networks. In: Proceedings of the Workshop on Economics of Peer-to-Peer Systems, Berkeley, CA, USA (2003).

Heuristic for the Assignment of Tasks in a Real-Time Distributed Environment

Luis Alberto Gutiérrez Díaz de León[1], Carlos Alberto Franco Reboreda[1], Carlos A. Gutiérrez[2]

[1]Universidad de Guadalajara, Av. Juárez 976, Col. Centro, C.P. 44100, Guadalajara, Jal, México
[2]Universidad Panamericana Aguascalientes, Josemaría Escrivá de Balaguer 101, C.P. 20290, Aguascalientes, Ags., México
luis.gutierrez@redudg.udg.mx, carlos.franco@cuaad.udg.mx, cagutierrez@up.edu.mx

Abstract. This work consider the scheduling of periodic tasks or processes with real-time constraints in a distributed environment. Each task must be executed meeting deadlines, precedence relationships and resources constraints. The problem of scheduling tasks on a distributed environment can be viewed as a problem of assigning processes to the processors but keeping the schedulability in local environments. Thus, the problem can be divided in two phases: the first phase is assigning processes to processors and the second is to schedule assigned processes in each processor in the distributed environment. This paper focuses in the first phase. It introduces a heuristic mechanism for assigning processes in a distributed environment with real-time constraints. The heuristic mechanism has two alternative ways to perform: one takes into account the utilization factor and the other considers period. The performance of both alternatives proposed is evaluated via simulation on a high performance-computing platform.

Keywords: Real-time, scheduling, tasks, heuristic.

1 Introduction

Real-time systems are systems that must react to events in the environment within the confines of a stringent deadline. As a consequence, correctness of real-time systems does not exclusively depend on the accuracy of obtained results, but also the time these results are calculated. In hard real-time systems when there is a delay in the calculation of the results, it produces a decrease in the system performance with potentially catastrophic consequences. In hard real-time systems there are three kinds of constraints involving tasks: precedence, time and mutual exclusion on shared resources. Time constraints -known as deadlines- represent the time when a task must end its execution to prevent potential damage to the system. The tasks are classified depending on potential damage they could cause to the system in case they do not meet its deadline. For high risk of damage, the task is considered a critical task, otherwise it is non-critical. Tasks can -or not- be present periodically (in regular time

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 142-154
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

periods). Under this perspective, tasks can be classified in periodic and aperiodic. Periodic tasks are the most common and critical in real-time systems, but aperiodic tasks can also be critical or non-critical [5].

So, hard real-time systems must provide correct results (logically and within time constraints). In order to do this, they must consider and solve the problem of tasks constraints, which is known as a NP-complete problem. It means that it can not be assured that a practical solution exists in a reasonable time. However, approximate and quasi-optimal solutions are feasible and frequent enough to solve problems commonly found in practice. Distributed Hard Real-Time Systems increased its application domain to multimedia systems, virtual reality, process control, avionics, robotics, nuclear plants, quality of service, defense, military applications and so on, which would not be possible to look after outside from parallel and distributed processing environments [16].

In the distributed environment the complexity of the system increases proportionally to its application domain because it is necessary to increase the considerations (to a distributed environment) to keep the predictability level that is required in real-time. Besides, a new problem arise concerning the assignment of tasks or processes into processing entities keeping at all time the schedulability of the system, that is also a NP-complete problem [10]. However, as it is known, there are several techniques or ways to find solutions to NP-complete problems, in order to offer acceptable solutions even though they are not optimal. Regarding hard real-time distributed systems there has been development of several solutions with bounded models, less dynamic and with different scheduling criteria, but most of the proposals are for very restricted applications [4].

In this work it is proposed an algorithm for the beginning of the construction of an ordered execution plan of a set of periodic tasks with precedence relationships, time constraints and shared resource requests in a hard real-time distributed environment. This is the process assignment, which consists in the assignment of tasks or processes into processing entities that can execute those tasks or processes in such way that restrictions are met. The rest of the paper is organized as follows: description of terminology and presentation of the model, proposal presentation, and validation of the results via simulation environment.

2 Task Assignment in Distributed Environments

The problem of scheduling tasks in distributed environments or in multiple processors can be understood as the search for an execution plan that fulfill all restrictions related to tasks [3]. In this case, the search space could be represented with a tree where each node is an assignment of a task to a processor. Any path from the root to any leave of the tree is an execution plan, but not necessarily feasible, from the satisfaction of constraints perspective. Most of the algorithms perform the search trying to generate a complete path (from root to leave) that meets all restrictions. In case of several solutions in the tree, algorithms look for an optimal solution, trying to perform their search trough an heuristic function.

The search for generating a processor-task assignment can be classified in two ways: oriented to assignment and oriented to sequence. This truly depend on the semantics of edges in the search tree. When an algorithm selects a task and then tries to assign it to a processor, it is called an oriented to assignment algorithm. If the algorithm selects a processor and then tries to find a task to execute on it, the algorithm is oriented to sequence. Both algorithms traverse the tree in-depth searching for a feasible execution plan, are executed in an exponential time and produce identical results when exploring the tree in a single iteration [7].

When algorithms are static and there are no time constraints it is possible to explore the tree in a single iteration and find all possible solutions. When we consider dynamic algorithms in an environment where it is possible to accept new tasks during execution time, the search must be incremental, which increases complexity. Obtained results would be partial to specific intervals, and new tasks would be added among them.

Both options in systems with a single processor are identical, but there is a change in multiprocessor or distributed environment. In previous works it has been proved that oriented to sequence approach is not easily scalable because in each iteration it tries to select a task to execute it on the next processor. Sequence oriented schedulers are based on load balancing strategies instead of focusing on satisfy all time constraints [8]. In the other side, assignment oriented schedulers are always trying to execute the most urgent or critical task at the moment. The proposal presented in this paper is consistent with this approach, this is, oriented to assignment.

3 Terminology and System Model

Let **E** = { E_1,E_2,.., E_n }, be a processing environment, where E_i is the i-th processing entity, which is comprised of a processing resource (processor) and a set R_i of resources assigned to entity E_i and are different to processing resources. So, E_i generates a local processing environment. In such a way, the distributed environment **Dist=E+T** is comprised of the set **E** of processing entities and the communication paths **T** that are in charge of connecting processing entities. The relationship established in **Dist** can be defined by a fully connected graph.

The communication cost between any two processing entities is considered constant and is defined by variable **k**. It is constant because k is equal to the longest communication time between two processing entities, from here k = max{ T_1,T_2,.., T_m}. This allows to keep a predictable environment in execution and makes simpler the synchronization and schedulability analysis. **R** is the total resources in the distributed environment, from this **R** = {R_1,R_2,.., R_n}. Processing resources are not part of R, this is to provide more detail in the problem definition. Intersection between any two resource sets by processing entity is empty set. The distributed environment will have a set **P** of processes or tasks to be executed. The environment offers as a scheduling criteria the one defined by Earliest Deadline First [15]. The general set of processes is comprised of:

A set **t** of n periodic preemptive tasks, where:

- Each task t_i Є t□□has an execution time C_i and a period T_i in which its instances are to be activated.
- Activation of periodic tasks can be done in t=0 to for purposes of schedulability analysis. Activation of aperiodic tasks is not necessarily in t = 0 for purposes of execution.
- Periodic tasks have a deadline for its execution, known as deadline D_i.
- Deadlines of periodic tasks can be different from period.
- Periodic tasks can be run causal relationship, known as precedence constraints. Causal relationships are resolved by a proposed scheme in [12].
- The precedence relationships are represented by directed rooted trees, where the destination nodes must be executed after the source nodes.
- Periodic tasks belonging to the same precedence tree have equal periods.
- Periodic tasks are independent only if they are in different trees.
- Periodic tasks can access shared system resources. In general and in order to simplify the analysis, it is assumed that they only access a single resource (cases where access to more than one resource is required are not left out and they can be modeled as different processes serialized by precedence constraints).

A set **J** of m aperiodic tasks, where:

- Are independent.
- Do not have a specified time for execution, but are given a deadline according to MDAS [11], in order to be handled by the scheduler under EDF.
- Aperiodic requests are serviced according how they (FCFS).
- Aperiodic tasks have an execution time, known when they arrive to the system.
- Aperiodic tasks have unknown arrival time.

4 Formal Definition of the Problem

Given a set of processes **P** = **J** U **t** with time constraints (C, T, D), precedence relationships and resource restrictions. And given a distributed environment **Dist** comprised by a set **E** of processing entities and communications paths **T**. In this context the scheduling problem is to allocate resources from **R** and start times for the execution of tasks, to tasks in **(t U J)**, in order to complete all tasks with the imposed constraints. This is, maintaining schedulability property [4]. The definition of the problem includes local scheduling because in this model processors are not in **R**.

To define the distributed scheduling problem, it is divided in two stages:

a) Allocation of responsibilities for processing with guarantees of schedulability (Process Allocation).

b) Generation of local schedule schemes.

For point b, in [11] and [12] was proposed a solution that generates a local scheduling environment, this solution is consistent with current needs. The fundamental problem in this work is the full assignment of all elements of **P** to all processing entities of **E**, in a way that preserves the schedulability for every element of **E**. With the assignment of processes to the processing entities in conjunction with local scheduling, it is generated the distributed scheduling environment. The proposal for assigning processes to the processing entities is not done dynamically because the foundation is not load balancing the elements of **E**, but to make an allocation of processes to maintain the schedulability in these elements [1] [14]. Therefore in the process of assigning processes is evaluated the schedulability of the processing entity before assigning each process. The allocation strategy applies only to **t** processes because the processes of **J** are events produced in each processing entity, and therefore they must execute in the processing entity with the same aperiodic tasks dispatching scheme proposed in [11].

5 Proposed Solution

The proposed allocation of the elements of **P** to **E** processing entities is performed in three steps and in three different ways of assigning processes. Actually, when the assignment is performed, the execution plans of the processes are generated as shown in figure 1. The steps for the process assignment proposed in this paper are: I Clusters generation, II Neighbors process allocation, and III Disappearance of islands.

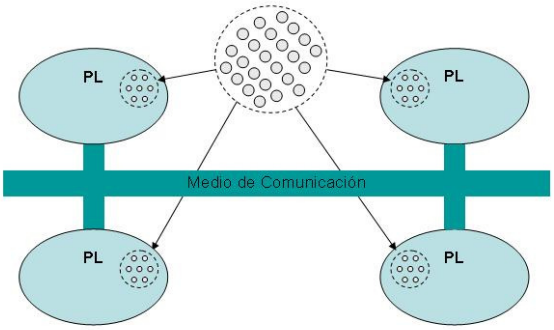


Fig. 1. Process allocation and production of local schedule plans.

5.1 Clusters Generation

With the set of processes in the initial form (fig. 2), the first step is to assign only those processes that have resource requirements from **R**.

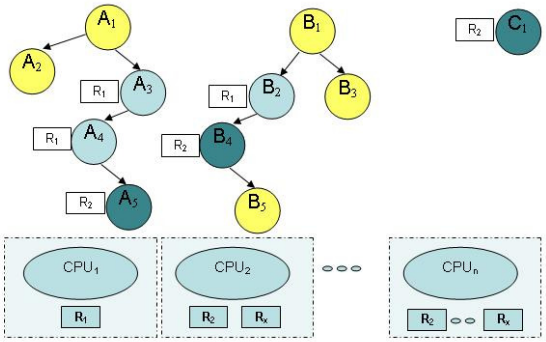


Fig. 2. Set of processes to be allocated.

Thus, any process t_i Є t, which requires a resource R_j has to be assigned to the processing entity E_k if and only if this entity contains the resource, as shown in figure 3. After the allocation of processes in this step, it is performed the schedulability analysis; if the set of processes is schedulable we can proceed to the next steps, otherwise we can not proceed with the assignment because **Dist** does not have enough resources to perform the whole set of processes meeting the restrictions imposed.

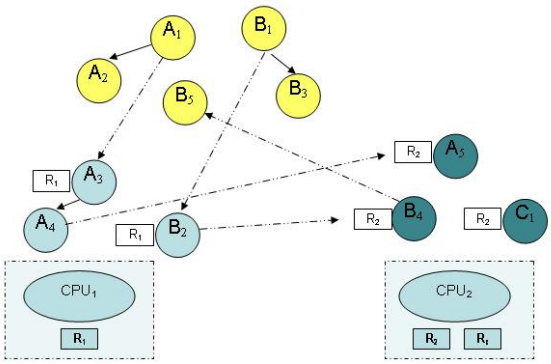


Fig. 3. Clusters generation.

5.2 Allocation of Neighbors

In this step the processes that do not have requirements from **R** are allocated. The processes that are still not allocated into processing entities may have causal relationships with processes already allocated to a processing entity of **E**. If a process that is going to be allocated has relationship wit one already allocated, the goal is that they both are allocated to the same processing entity to avoid delays in their communication due to synchronization signals. This step is known as allocation of neighbors.

To allocate processes in this step it is considered a heuristic function that leads the selection in the allocation of processes, considering reduce the communication costs (represented by constant k). However the heuristic function can be determined by two variations with the combination of factors that define the communication cost, among these factors are:

k: communication cost, o: the number of relationships that have a process when it is allocated with another process already allocated in the processing entity where it is being evaluated. (also known as occurrences), C: Execution time of the process in evaluation, T: Period of the process in evaluation.

So, the heuristic functions proposed and that are evaluated in this paper are described as follows:

$$f_1(k, o, T) = k * o / T \quad (1)$$

$$f_2(k, o, C, T) = k * o / (C/T) \quad (2)$$

These heuristics are intended to express the weight or cost that generate periodic tasks, and reduce it on their use. For this reason the first function express the cost directly related with the period or the frequency that each synchronizing task brings. The second function is a slight modification to involve the use of the processor that a task requires for its execution, involving by percentage of use the communication cost.

With these features the allocation of neighbors is conducted. In case we can not allocate a neighbor due to the lack of processing capacity, it will be evaluated and reallocated to the next processing entity where exists a relationship. In case there are no relationships with other element of **E**, then it is assigned to the processing entity with smaller utilization factor and where the allocation can keep schedulability conditions, this according to Earliest Deadline First Best Fit (EDF-BF) allocation criteria, defined in [13].

In the absence of processing entities that can withstand the process that is being assigned, then it is enabled a new processing entity and it is evaluated its allocation. In case of a new failure, it is determined the infeasibility to schedule the set of tasks in the system. This strategy allows controlling the number of processors that is actually required for system implementation.

One option might have been to enable processing entities and assign resources to processes as they were needed, but no matter how seductive this possibility may seem, there is the potential risk of obtaining a system design that does not correspond to reality by separating resources where they are really necessary.

For instance, consider the case of interaction with an actuator; the actuator must be physically in a place determined by the environment of the system. Seeking to change this assignment in theory can lead to produce unrealistic systems, which is beyond the scope of this paper.

5.3 Disappearance of Islands

In this step there may be present processes that still have not been allocated because they do not have relationship with processes with need of resources or processes that have been already assigned. So these processes remain isolated in the allocation until this step, generating what is known as process islands.

Once processes have been allocated in the previous two steps, we check the existence of islands.

In case of the existence of any island, the root process of the island is assigned to the processing entity with higher residual utilization factor and where the allocation keep the schedulability conditions, and we continue in the step of allocation of neighbors. This step is consistent with the Earliest Deadline First – Worst Fit (EDF-WF) allocation criteria, whose definition is found in [6]. The reason for using it is because root processes can have children that in the ideal case will be allocated in the same processing entity; this is to propose that precedences are local as far as possible to reduce communication costs. In case of using EDF-BF the possibility to generate remote precedences increases.

In the absence of processing entities that can withstand the process that is being assigned, then it is enabled a new processing entity and is evaluated for a possible allocation. In case of a new failure, it is determined the infeasibility to schedule the set of tasks in the system. In the absence of any island, then the allocation process has completed successfully and the next step is the preparation for execution, this is, the preset parameters in local schedulers (considering the phase of synchronization of phases in tasks addressed in the previous section).

The allocation criteria EDF-BF y EDF-WF were selected in this proposal considering their performance evaluated in the studies published in [17] [2].

6 Simulation Environment

The simulator implemented to evaluate the proposal of this work, it is really an execution environment that consists of a microkernel, which incorporates the properties of a minimal real-time system, it also includes a pitcher of execution modules that will be considered as the process to analyze their behavior, this last one was developed in this work. A MicroKernel is to add the Linux system kernel, a second kernel that acts as an interface layer between the hardware and the standard kernel. The microkernel layer controls the execution of real-time tasks and all of the conventional system. It considers the standard kernel like one more task without time constraints. The architectural design of the simulator, has a module to generate tasks as a running module. This running module and all generated by it, depend on the micro kernel, which generates real-time environment under EDF. Each of the experiments contained 100 periodic tasks whose periods were generated as a uniformly distributed variable from 10 to 1000 units of time. Each C_i/T_i was evaluated as a variable uniformly distributed in [0,1] and was standarized to provide U_P of each experiment whereas U_{Ptotal} vary from 4 to 7 units. From this, the experiment considered up to 10 processors. To fully characterize the environment, we generated a set of activities or events (aperiodic load) per processor to be varied in each graph from one point considered to 1-U_P, where for each load were carried out 1000 simulations. The arrival times for aperiodic tasks were modeled using the exponential distribution, with interarrival average calculated to match the aperiodic load. The execution times of aperiodic tasks were modeled using a uniform distribution [10,100].

7 Results

The results were obtained by 1000 runs with a length of 50,000 units over time. The simulation is run to compare the performance of the heuristic function in the two alternatives defined in (1) and (2). The first one considers the period to minimize communication costs; the second considers the utilization factor. The scope where the alternatives were evaluated with the same sets of tasks (varying processing loads) was in terms of number of processors used, percentage of schedulability, communication cost and response time. The results are shown in Figures 4, 5, 6 y 7.

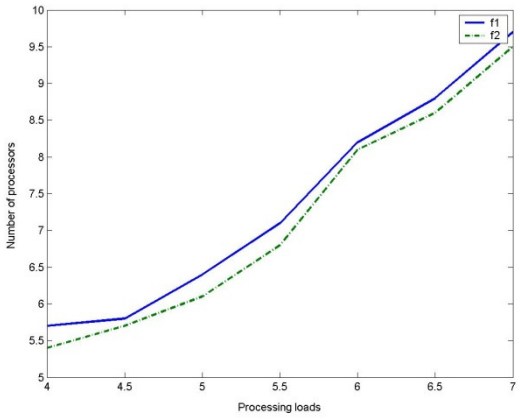


Fig. 4. Number of processors.

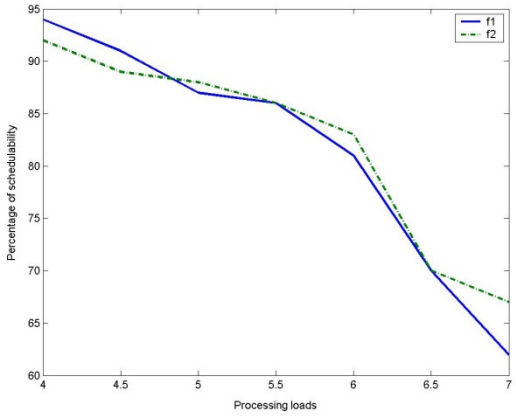


Fig. 5. Schedulability.

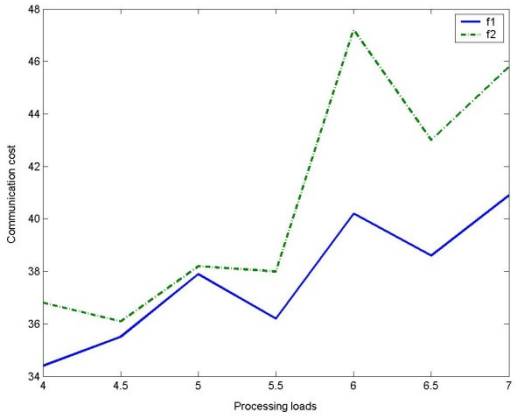


Fig. 6. Communication costs.

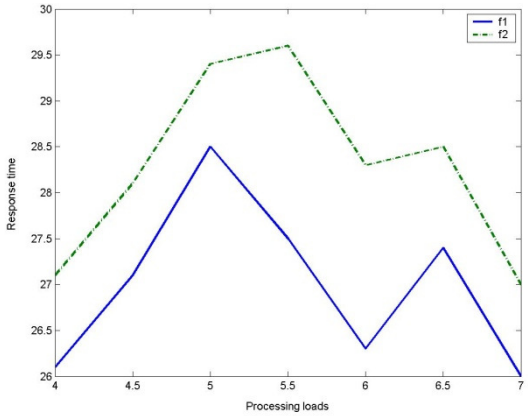


Fig. 7. Response time.

The results show that when considering the utilization factor in the heuristic function (2) it reduces the average number of processors to use, but in terms of schedulability does not generate any trend because the two heuristics work similarly (random) on a common combinatorial problem. With regard to reducing communication costs or reducing synchronization signals, the heuristic function that considers the timing of the processes (1) shows an improvement compared to the function (2) because the trend shows that there are fewer synchronization signals with function (1); this aspect correlates with response times, since these are also improved by reducing the number of synchronization signals.

8 Conclusions

In distributed real-time environments is essential the process allocation activity. This is because we now have to take care of the schedulability not only in a single processing entity but in a set of them communicated by message passing. In general the problem is classified as NP-complete. In this paper, we presented a scheme of allocation of processes to establish a real-time distributed environment trough local schedulers. The allocation scheme has a heuristic function to conduct the search assignment, trying to reduce communication costs. For the definition of the heuristic function we had two alternatives that were evaluated in a simulation environment. After analyzing the simulation results, we conclude that the function which considers the utilization factor for the allocation of processes, showed better results in terms of number of processors used, however with the function that considers only the frequency of processes there were less synchronization signals and thus response times are lower. Therefore, $f_1(k, o, T) = k * o / T$ gives better results by reducing response times due to reduction of communication costs. This work will continue with its validation through mathematical analysis of what we concluded and the evaluation of the performance of the entire distributed environment, considering the implementation of local schedulers.

9 Bibliography

[1] Atif, Y., & Hamidzadeh, B. (1998). A Scalable Scheduling Algorithm for Real-Time Distributed Systems. *Proceedings of the 18th International Conference on Distributed Computing Systems* , (págs. 352-359).

[2] Beitollahi, H., & Miremadi, S. (2005). Systems, Performance Evaluation of Fault-Tolerant Scheduling Algorithms in Real-Time Multiprocessor. En M. H. T. Fahringer (Ed.)., (pág. 738). Innsbruck, Austria.

[3] Burns, A., & Wellings, A. (2001). *Real-Time Systems and Programming Languages.* England: Addison Wesley .

[4] Buttazzo, G. C. (2002). *Hard Real-Time Computing Systems.* . United States of America: Kluwer Academic Publishers.

[5] Buttazzo, G., & Sensini, F. (1999). Optimal Deadline assignment for Scheduling Soft Aperiodic Tasks in hard real-time environments. *In IEEE Transactions On Computer* , 76.

[6] Coffman, E. G., Galambos, G., Martello, S., & Vigo, D. (1998). *Bin Packing Approximation Algorithms: Combinatorial Analysis.* Kluwer Academic Publishers.

[7] Gantman, A., Gui, P.-N., Lewis, J., & Rashid, F. (1998). *Scheduling Real-Time Tasks in Distributed Systems* . California : University of California, San Diego.

[8] Gantman, A., Guo, P., Lewis, J., & Rashid, F. (1998). *Scheduling Real-Time Tasks in Distributed Systems: A Survey.* Recuperado el 20 de enero de 2007, de Internet Draft,

www.cs.ucsd.edu/classes/fa98/sce221/OSSurveF98/index.htm
www.cs.ucsd.edu/classes/fa98/sce221/OSSurveF98/index.htm

[9] Garey, M., & Johnson, D. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W.H. Freeman and Company.

[10] Graham, R. (1969). Bounds on multiprocessing timing anomalies. *SIAM J. Appli. Math.* , 416-429.

[11] Gutiérrez, L., Franco, C., Jacinto, R., & Gutiérrez, C. (2006). Minimizing the Response Times of Aperiodic Tasks in Hard Real-time. *Electronics, Robotics and Automotive Mechanics Conference* (págs. 268-273). Cuernavaca, Moreles: IEEE Computer Society.

[12] Gutiérrez, L., Jacinto, R., & Rodríguez, F. (2002). Algoritmo para la creación de un plan ordenado de ejecución de tareas periódicas con relaciones de precedencia bajo EDF. *Memorias del Congreso Internacional de Computación 2002* (pág. 240). México, D.F.: CIC.

[13] Johnson, M. R. (1979). *computers and Intractability: A Guide to the Theory of NP-Completeness.* New York: W.H. Freeman.

[14] Khemka, A., & Shyamasundar, K. (1997). An Optimal Multiprocessor Real-Time Scheduling Algorithm. *Journal of Parallel and Distributed Computing, vol. 43* , 37-45.

[15] Liu, C., & Layland, J. (1973). Scheduling algorithms for multiprogramming in a hard-real-time environment. *Journal of the Association for Computing Machinery* , 139.

[16] Masters, M. (1998). Challenges in designing future command and control systems. *Workshop on Parallel and Distributed Real-Time systems.*

[17] Pereira Zapata, O. U., & Mejia Alvarez, P. (2002). EDF and RM Multiprocessor Scheduling Algorithms: Survey and Performance Evaluation. *Report No. CINVESTAV-CS-RTG-02. CINVESTAV-IPN, Sección de Computación.*

A Bandwidth Sharing Method under Node Autonomy and Short-Term Protection for QoS

Alfredo P. Mateos-Papis (a).

(a) Division of Sciences of Communication and Design. Universidad Autónoma Metropolitana. Cuajimalpa. Mexico DF.
amateos@correo.cua.uam.mx

Abstract. This paper proposes a method to share bandwidth between intersecting routes in networks subject to per-route distributed admission processes. With this method the current bandwidth share of a flow is guaranteed in the short-term against a possible bandwidth decrease caused by flows recently admitted in intersecting routes. However, this share-protection is decreased slowly such that, in the long term, bandwidth is distributed according to the intersecting route's observed requirements. With this method every node takes its own decisions, having an impact on the bandwidth sharing between routes and attaining, with this, global behaviors in the network. This method is expected to help preserve the Quality of Service in the routes while, because of its simplicity, allowing a simple bandwidth management in the network and possibly permitting its implementation in bigger-sized networks. The method proposed is termed: the Short Term Protection (STP) method.

Keywords. Quality of Service (QoS); Bandwidth Sharing; Distributed Traffic Control.

1. Introduction.

This article addresses in a novel way the problem bandwidth-sharing for the long-time studied QoS-aware-networks having per-route distributed admission processes, using a novel uncomplicated method where the nodes operate autonomously offering the possibility to attain global behaviors.

In bandwidth-limited data networks which offer Quality of Service (QoS), a flow-admission-process is used to foresee if the traffic increase caused by the entrance new flows to the network will originate an unacceptable deterioration of the QoS offered.

An admission process to a network can be centralized or distributed. In the centralized approach a single entity (like a bandwidth broker) makes admission decisions based on its global view of the network. This global-view requirement makes this approach less scalable compared with that of the distributed admission approach wherein there is an admission entity for each one of different parts of a network. The admission to each part is done based on a partial view of the network, with the potential risk that an admission decision could inadvertently cause an unacceptable deterioration of the QoS in other parts of the network.

The bandwidth sharing method introduced in this paper, which is called Short-Term Protection method (STP method), is aimed at networks with per-route distri-

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 155-168

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

buted admission processes, as those proposed in [1] [2], where there is no bandwidth reservation considered for any route. This method is intended to protect routes against bandwidth exhaustion derived from sudden increments in traffic in intersecting routes while still allowing the network to have a simple distributed management scheme which adjusts in accordance to the bandwidth necessities of the routes. With this method each node acts autonomously and there is no central admission authority considered. The nodes are aware neither of the existence of routes nor of the existence of flows in the network. Fig. 1(a) is a schematic representation of a network with two routes that intersect at node *x* (more specifically at its output interface) and share the bandwidth of the node's output link. In this paper these routes are called *intersecting routes* and node *x* is called an *intersection node*.

An intersection node has, at every one of its output interfaces, one queue for every one of its input interfaces. As an example, Fig. 1(b) represents a node with three input interfaces: *A*, *B* and *C*, from where three routes converge at the node and leave the node through the output interface *D* which has three queues.

Inspired by the Weighted Fair Queuing[1] (WFQ) scheduling algorithm [3], the STP method gives each queue a weight, however, each weight may change slowly according to an updating algorithm that is periodically evaluated. In this paper the STP method uses a WFQ scheduler, but the scheduler could be other kind of work-conserving scheduler [4] which could adequately operate with changing weights.

In the STP method traffic coming from the different input interfaces competes for bandwidth in a special way: the weights change slowly in favor of the queues which have bigger average lengths. For example, a route could lose up 20% of its weight in a 30-minute interval, against an intersecting route, if its queue length were constantly smaller –within this interval– compared to that of the intersecting route.

The STP method offers a simple and novel option for data networks to offer scalable QoS in comparison with the DiffServ model, studied since 13 years ago, which is oriented for scalability and flexibility in bandwidth-sharing (see section 2) but which poses a native difficulty in the evaluation of the traffic conditions in the networks.

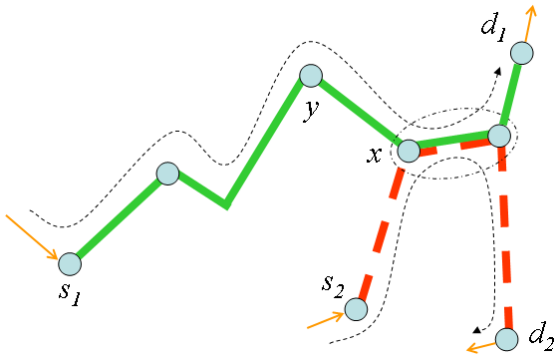
2. Background and Related Works.

The STP method has been motivated by some characteristics of the nodes used in the DiffServ[2] model [5] [6], and in the network conditions proposed in [1] which deals with networks with per-route distributed admission where the admission decisions are taken at the edge of the routes. In DiffServ the packets are classified when entering to the so-called DiffServ domain (so its admission is per-class oriented instead of per-flow oriented), and the central nodes (also called core nodes) of that domain follow a specific behavior when treating the packets of a specific class.

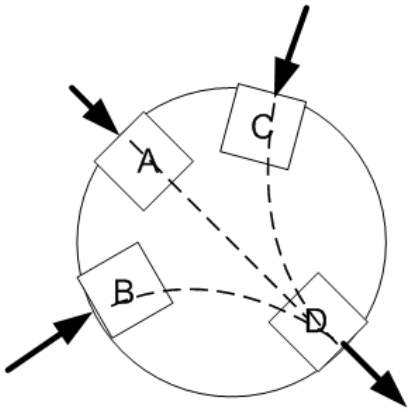
It is reasonable to consider that the behaviors provided by a node are implemented at its output interfaces [7]. Typically, an output interface of a node keeps one queue (buffer) for every one of the classes to attend, which share the total bandwidth of the output interface. Any output interface with more than one queue in operation needs a

[1] Also called PGPS (Packet-by-Packet Generalized Processor Sharing).

[2] See technical report *Rep-Tec-4-Arquitectura DS.pdf* at http://ccd.cua.uam.mx/~amateos/.



(a) Routes $s_1 \rightarrow d_1$ and $s_2 \rightarrow d_2$ in a network intersect at the output interface of node x.



(b) Representation of an intersection node with three input interfaces, *A*, *B*, and *C*, and one output interface, *D*, in which the incoming traffic converges.

Fig. 1. The concepts of two intersecting routes (subfigure a), and the concept of an intersection node (subfigure b).

scheduler to choose the next packet to be dispatched over the output interface, thus creating a bandwidth sharing process between classes.

There are methods for optimization of a global indicator in a network by obtaining the most effective bandwidth-share between classes. For example [8] presents a bandwidth-sharing method described as proactive (in contrast with *reactive* algorithms) which uses a central bandwidth manager. Another bandwidth-sharing method between classes [9] adaptively adjusts the weights of a weighted round robin scheduler in every core node and does not depend on a central bandwidth manager. It looks like the nodes can effectively operate autonomously and cooperate to obtain an expected end-to-end QoS indicator.

Centikaya et al [1] [2] propose a method of admission to a network that has predefined routes. This method is defined as *distributed* since the admissions are made per-route, not per-flow oriented, without a central management, where a route can cause a

negative impact in intersecting routes when admitting a flow. The main research goal of the paper is the admission process, not the route-intersection problem.

The STP method is oriented towards handling the negative interaction between intersecting routes with per-route admission but it is not aware of any admission process.

In [10] a *Coordinated-Schedulers* method is presented to provide delay-bounds in a network, which is not per-flow oriented. In it, the core nodes modify each arriving-packet's priority index depending on whether the packet was serviced late or early at the upstream node as a consequence of cross-traffic. The method provides natural coordination between nodes but each node operates autonomously. This method can work in the context of a single class and it seems that it could have the benefit of protecting a route against intersecting routes. The STP method is simpler although it is not in the scope of this paper to study if it can attain the benefits of the method presented by [10].

3. The STP Method.

3.1 Remarks.

The STP method was conceived to operate within one class of traffic in a network, so this paper considers traffic in a single class. The possible interaction of a class using the STP method with schemes doing bandwidth provisioning between classes, to optimize a global indicator in a network, is out of the scope of this paper.

The evaluation of the STP method is made with the results obtained from simulations with traffic generated using a mixture of constant bit rate and Pareto sources. This traffic does not represent the diversity of traffic that might exist in a network; nevertheless these results are presented as a starting point for the evaluation of this method. In the simulations a change in the amount of traffic is made through the change of the number of sources.

The evaluation of the method is limited to a single direction in the routes as it is considered that the traffic in one direction is independent of the traffic in the opposite direction.

The topology used for the simulations is rather small. It has one central node (see section 2). The STP method is used in just the central node as this paper is aimed to compare the results in a network when a single central node uses the STP method with the results when that central node is a *normal* node, that is, the node uses a single output queue for each output interface. The author has made simulations with two nodes capable of using the STP method, which will be presented in a subsequent paper.

3.2 Condensed Explanation of the STP Method.

The QoS parameter of interest in this paper is the delay. This paper evaluates the STP method based on the end-to-end delay in the routes. This delay depends on several factors, including the queuing delay in the intersection nodes which will be the

factor of interest in this paper. This paper defines the term *delay-limit* as the maximum permissible delay that a packet can experiment while traversing a route.

Consider an output interface of a node working with the STP method, where there are *N* queues being attended by the scheduler (so there are *N* routes intersecting at that interface). The weight of each queue should tend to be proportional to its relative average length compared with the sum of the lengths of the queues. The rate of weight-change should be slow, that is, the bandwidth required to satisfy a bandwidth-greedy route should be released slowly in order to protect, temporarily, the routes which would loose bandwidth.

The method starts at time t_0 where the all the weighs have the same value. At the ending of every time-interval of duration τ, the method computes an indicator for every queue of the output interface, to compare the current weight of the queue with its current relative length with regard to the lengths of the other queues. The indicator is represented with $I\Delta$ in equation (1). This time-interval should be sufficiently big as to be able to observe several arrivals and departures of packets (in this paper $\tau = 1s$). After computing this indicator the method computes the new weights of the queues and substitutes those in operation. All this computation should take place in a time much smaller than τ. In (1), the computation of this indicator is made at time $t_0 + (r+1)\tau$ for queue *i*, where *r* is an integer such that $r \geq 0$, $\phi_i^{Act}(t_0 + r\tau)$ represents the weight of queue *i*, which has been valid from $t_0 + r\tau$ to $(t_0 + (r+1)\tau)^-$), and the average lengths of the queues, obtained at time $t_0 + (r+1)\tau$, are represented with $Q_i(t_0 + (r+1)\tau)$.

$$I\Delta\left(\phi_i^{Act}(t_0 + r\tau)\right) = \frac{Q_i(t_0 + (r+1)\tau)}{\sum_{j=1}^{N} Q_j(t_0 + (r+1)\tau)} - \phi_i^{Act}(t_0 + r\tau) \quad \textbf{(1)}$$

If, from equation (1), the indicator results negative for queue *i*, then this queue should lose weight so that the result of its new weight, $\phi_i^{Act}(t_0 + (r+1)\tau)$, minus its weight in operation, $\phi_i^{Act}(t_0 + r\tau)$, should reflect a negative increment (or a decrement). Equation (2) proposes a form of calculation for this negative increment, which is represented with $\Delta\left(\phi_i^{Act}(t_0 + r\tau)\right)$.

$$\Delta\left(\phi_i^{Act}(t_0 + r\tau)\right) = \phi_i^{Act}(t_0 + (r+1)\tau) - \phi_i^{Act}(t_0 + r\tau) = \ln(1-P)\frac{\tau}{T}\phi_i^{Act}(t_0 + r\tau) < 0 \quad \textbf{(2)}$$

The two important parameters of (2) are *P* and *T*. Parameter *P* is called the *loss factor*, which is a positive constant much smaller than 1 (with values near to 0.1). When *P* is very small then $-\ln(1-P) \approx P$. The parameter *T* represents a *time span* and it might be on the order of 1200*s*. The ratio T / τ should be a big integer called *K*.

From (2) it can be seen that, as τ is much smaller than T, the decrement of the queue, from one interval to the next one, should be very small.

It can be shown that if queue i decreased its weight for K consecutive intervals of size τ within time-span T, then the ratio of its initial and ending weights, with regard to that time span, would be equal to $(1-P)$, that is: $1-P = \phi_i^{Act}\left(t_0+(l+1)T\right) / \phi_i^{Act}\left(t_0+lT\right)$. This formula can be written as: $\phi_i^{Act}\left(t_0+(l+1)T\right) - \phi_i^{Act}\left(t_0+lT\right) = -P\phi_i^{Act}\left(t_0+lT\right)$. With this it can be stated that with this method the maximum weight deterioration for a queue in a time T is P%. This result is demonstrated using the fact, which can be obtained from (1) and (2), that $\phi_i^{Act}\left(t_0+(m+K)\tau\right) = \phi_i^{Act}\left(t_0+m\tau\right)\left(1-f(P)/(T/\tau)\right)^{T/\tau}$. The variables l and m represent integers greater than, or equal to 0, and $f(P) = -\ln(1-P)$. Because of the limitation of space of this paper, all the details of this method are not developed here[3].

The STP method also proposes a calculation for the weight of every queue that should not decrease its weight, as indicated by (1), after every interval of duration τ, in such a way that the sum of the weights of all queues be always equal to 1.

Following the method proposed in [11], the average length Q_i of queue i is computed as $Q_i = (1-w_q)Q_i + w_q q_i$, where w_q is the averaging parameter and q_i is the instantaneous queue-length. This equation acts as a low-pass filter. In this equation, the smallest the value of w_q the smoother the output will be. The value for w_q is set to 0.002, to cope with the possibly burst behavior of the instantaneous queue-lengths.

4. Experimental Results

The performance evaluation of the STP method was done through a series of simulations which uses the topology shown in Fig. 2, a bounded network (or domain) with one central node, *c1*, and three edge nodes: *e1, e2* and *e3*. Outside the network boundaries there are three source nodes: *s1, s2, s3*, and three destination nodes: *d1*, *d2* and *d3*. Inside the network there are three routes: *Route1*, *Route2*, *Route3*. See that node *c1* acts as an exit node for *Route2*. Links inside the network have a bandwidth of 3*Mb/s*, for the operating Class. The links connecting the bounded network with the outside nodes have a bandwidth of 100*Mb*/*s*. The propagation delay of the links is 0.05ms. *Route1* and *Route3* intersect at the output interface of *c1* (going to *e2*). *Route2* intersects with *Route1* at the exit interface of *e1* (going to *c1*). *Route1*, then, suffers from two intersections which puts it in disadvantage against *Route3* which is not affected by *Route2*. In the experiments the situation is that *Route3* increases its traffic so that *Route1* is affected with this increase.

Node *e1* was selected to be a *normal* node in every experiment, and node *c1*, in a group of experiments uses the STP method and in other group of experiments it does not, as it is explained in the next paragraphs (remember that this paper is aimed at

[3] See technical report *Rep-Tec-2-Equations-STP-Method.pdf* at http://ccd.cua.uam.mx/~amateos/.

evaluating the effect of the STP method in just one node in the network). With the STP method *Route1*-traffic passes through one queue and *Route3*-traffic passes through the other queue of the output interface of *c1* going to *e2*. *For the sake of simplicity-of-explanation the first queue of c1 can be referred to as the queue for Route1 and the second queue as the queue for Route3*.

The results of these experiments are compared to evaluate the STP method. Because of the random nature of experiments the results are taken from the mean values of 40 to 60 experiments, for each result.

4.1. Experimental Setting

The experiments were carried out with the ns-2 simulator [12], version 2.33, also using the DS tools included in the simulator [13]. In the STP method these tools were modified to add the WFQ packet-scheduling operation, with the implementation proposed in [14], which uses the WFQ scheduler [3]. Additional modifications were made to incorporate the possibility to change the weights of the scheduler in a dynamic form, in accordance with the STP method[4].

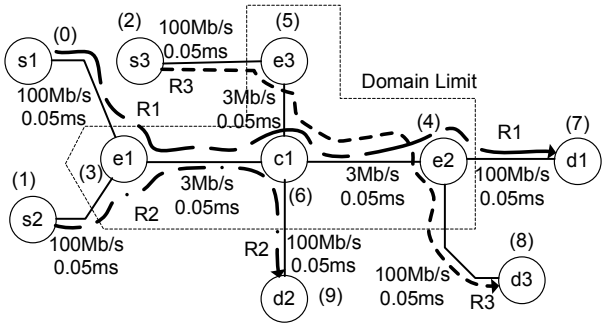


Fig. 2. The topology for the experiments has two routes: *Route1* traversing nodes *s1–e1–c1–e2–d1*, and *Route3* traversing nodes *s3–e3–c1–e2–d3*. The routes intersect at the output interface of node *c1* and separate at node *e2*. *Route1* and *Route2* intersect at the output interface of node *e1*.

The Pareto sources are *On/Off* with 250*ms* on/off duration, with 68 Kb/s during *On* periods and 95-byte packets [15][5] with Pareto shape parameter of 1.7 (infinite variance). The CBR sources have 1500-byte packets with 256Kb/s[6]. It should be clear that these rates are small compared with those actually used in the Internet, but they

[4] See technical report *Rep-Tec-3-WFQ-Changing-Weights.pdf* at http://ccd.cua.uam.mx/~amateos/.

[5] A voice packet can have 67 bytes which would make a 95-byte IP packet adding 8 bytes of UDP header and 20 bytes of IP header. A reasonable voice mean-rate of 3.0 Kbytes/s in one direction makes a traffic mean rate of 34Kb/s (3 Kbyte/s x 8 bits/byte x 95 / 67 = 34 Kb/s – considering the headers' overhead), making 68 Kb/s during the On period of On/Off sources.

[6] The rate value is taken from information from [16].

are used for the purpose of the evaluation of the STP method, within the scope of this article, as indicated in section 3.1.

It is supposed that for a long time before the beginning of each experiment the routes have had the same traffic. At the beginning of every experiment *Route3* increases its number of sources (without knowing about the possible deterioration of the QoS inflicted on *Route1*).

Three different groups of experiments were carried out, depending on the buffer used in the output interface of *c1*. In the first group, called "case *q1*", there is just one queue at the output interface of node *c1* (going to *e2*), which handles the traffic of the two routes. The second group, called "case *q2*", uses the STP method and utilizes two queues at the output interface of node *c1*, one corresponding to each one of *Route1* and *Route3*. The parameters employed for case *q2* are: $T = 1200s$, $P = 0.25$. The time τ is $1s$. The last groput, called "case *q2f*", also utilizes two queues at the output interface of node *c1*, one corresponding to each one of the routes, but the weight of each queue is fixed throughout the duration of the experiment.

For a *P* parameter value greater than 0, a *q2* case with parameter *T* being close to 0 should offer results similar to those of the *q1* case. A *q2* case with parameter *T* being close to infinite should offer results similar to those of a *q2f* case. All the other queues in all the nodes are Droptail. The queue lengths are big enough such that the total observed packet loss rate is always less than 2% in every experiment. In every experiment *Route1* and *Route2* have almost the same number of sources. *Route1* has 3 CBR sources + 17 Pareto sources, and *Route2* has 3 CBR sources + 16 Pareto sources. These numbers of sources are big enough to cause average packet-delays near to $3ms$ at the exit interface of node *e1*, without causing packet congestion losses.

Route3 initiates each experiment with 3 CBR sources + 8 Pareto sources. The initial traffic-share of *Route1* and *Route3*, at the exit of node *c1*, is: 0.5641 and 0.4359, respectively. These values are also the initial weights used for the queues of *c1* going to *e2*, for the *q2* case, which correspond respectively to *Route1* and *Rotue3*. Remember that the weights given by the STP method to intersecting routes in an output interface of a node (node *c1* in this case) tend to be the same as the bandwidth proportion that the routes take at that interface, in the long term. For the *q2f* case these values are taken as the fixed weights for the two queues of *c1* going to *e2*.

Each row of Table 1 shows the number of sources of *Route1* and *Route3*, their rates and the bandwidth percentage share at the exit interface of node *c1* going *e2* for every experiment. +PAR means the number of Pareto sources increased in *Route3* at the beginning of the experiment, for example, in the second row the total number of Pareto sources of *Route3* is 8 + 1 = 9). The heading –%R1 indicates, for a given row, the bandwidth percentage deterioration of *Route1* with regard to that which this route has in the first row. The queue-weight deterioration for a route with the STP method can not be bigger than the traffic-share deterioration of that route within a time-duration *T*). For example in the second row %R1 = 55.62 so that 1–55.62/56.41 = 0.01405 (1.41%). The maximum traffic-share deterioration for *Route1* shown in the table is 13.55%.

Table 1. Number of sources of *Route1* and *Route3*, their rates and the bandwidth-percentage share at the exit interface of node *c1*, going to *e2*, for every experiment.

CBR	CBR	PAR	PAR	+PAR	Kb/s	Kb/s	Kb/s	%	%	–%
R1	R3	R1	R3	R3	R1	R3	R3+R1	R3	R1	R1
3	3	17	8	0	1346	1040	2386	43.59	56.41	0.00
3	3	17	9	1	1346	1074	2420	44.38	55.62	1.41
3	3	17	10	2	1346	1108	2454	45.15	54.85	2.77
..	..	..	..							
3	3	17	16	8	1346	1312	2658	49.36	50.64	10.23
3	3	17	17	9	1346	1346	2692	50.00	50.00	11.37
3	3	17	18	10	1346	1380	2726	50.62	49.38	12.47
3	3	17	19	11	1346	1414	2760	51.23	48.77	13.55

4.2. Initial Results.

Fig. 3 presents the weight values for the *q2* case of the output interface in node *c1*, for four different values of *P*, through a period of time $T = 1200s$ in a special group of experiments where the traffic of *Route3* is bigger than that of *Route1*[7].

In the case of $P = 0.05$, at $t = T = 1200$ the weight of the queue attending *Route1* should be $0.564 * 0.95 = 0.5358$, and the experiments offered a weight value of 0.5291. For $P = 0.15$ this weight should be of $0.564 * 0.85 = 0.4794$, and the experiments offered a weight value of 0.4661. For $P = 0.25$ the weight obtained should be $0.564 * 0.75 = 0.4230$ and the experiments offered a weight value of 0.41. All these results are close to the projected ones.

4.3 Gain Results.

In order to assess the benefits of the STP method, this paper uses a figure of merit which consists of an overall gain for each route, for each experiment. This gain rewards with one point for every packet that traverses its route within the delay-limit of the route, and penalizes with ten points otherwise.

The first scenario considers experiments of short duration, 120*s*, for cases *q1* and *q2f*, and a delay limit is 18*ms*. The reason to make this scenario is to observe the behavior of these two cases to compare their results with those of the *q2* case, made in the next scenario. The left side of Fig. 4 shows the gains obtained for *Route1* and *Route3* (*Gain1* and *Gain3*) as a function of the number of sources added to *Route3*. Every point in the figure shows the gain obtained in each 120*s*-experiment. The upper abscissa axis shows the number of sources added to *Route3* in the experiment, and the lower abscissa axis shows the resulting rate in *Route3*. The gain in each experiment is normalized dividing it by the duration of the experiment in the simulation.

[7] *Route3* augments 30 Pareto sources at the beginning of the experiment time. Regarding Table 1, *Route1* has 40% of the traffic and *Route3* has 60% (only for this experiments the bandwidth of *c1-e2* is 4[Mb/s]).

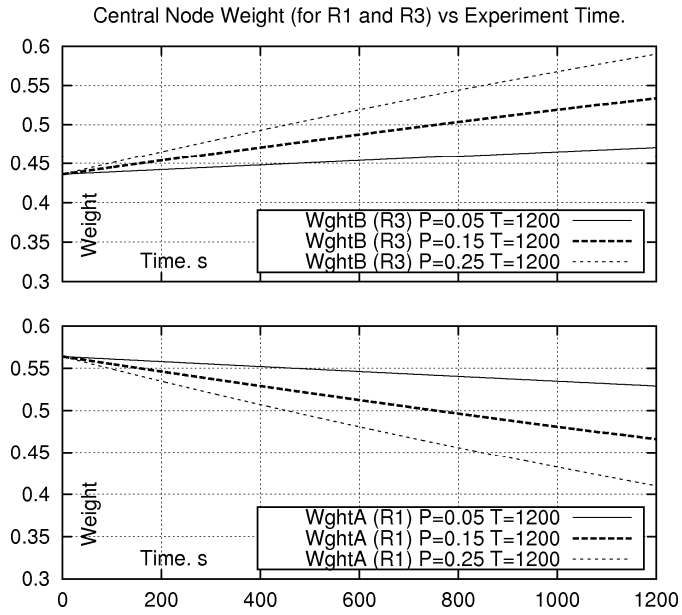


Fig. 3. Experimental results of weight-functions for the *q2* case, with different values of *P* parameter. Label "Wght A (R1) P=0.15 T=1200" stands for weight of *Route1* for the *q2* case with parameters *P*=0.15 and *T*=1200.

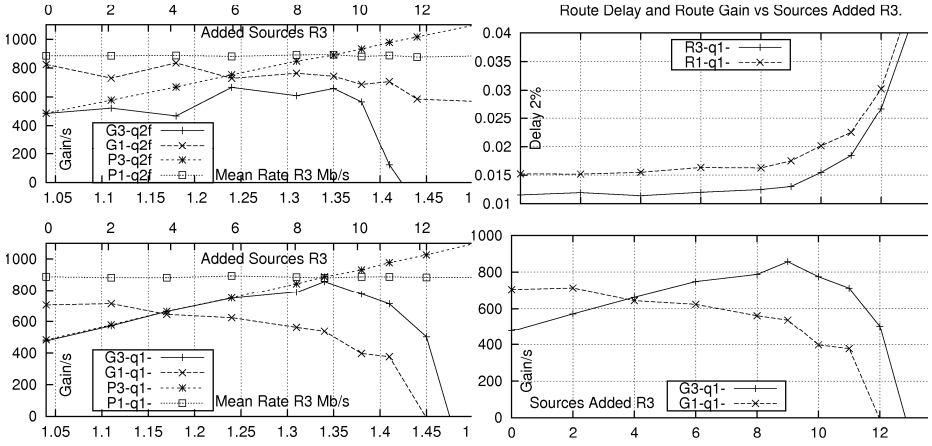


Fig. 4. The left side of the figure shows the gains of routes against number of sources added in *Route3*, for cases *q1*, and *q2f* with a delay limits of 18*ms*. Label "G3-q1" stands for *Gain3* in the *q1* case, and label "P1-q2f" stands for stands for number of packets for *Route1* for the *q2f* case. The meaning of the other labels is similar. The right side of the figure shows the gains and the delays (in *s*) of the routes for the *q1* case.

For the *q1* case *Gain3* has a maximum value at 9, decreasing afterwards. The admission process of *Route3* should avoid allowing an increase of more than 9 sources. *Gain1* deteriorates with the increase of sources in *Route3*. For the *q2f* case it is also clear that the admission process of *Route3* should neither allow more than 9 sources. From 0 to 9 sources, for the *q1* case, *Gain1* goes from 704 a 537 points, representing a loss of (704-537) / 704 = 27.7%. For the *q2f* case *Gain1* goes from 825 to 742 points, representing a loss of (825 – 742) / 825 = 10%. It is clear that the *q2f* case protects *Route1*. It is observed that the *q1* case is a good option when the available bandwidth is big compared to the traffic in the routes[8].

The right side of Fig. 4 shows the gains and the delays of the routes for the *q1* case (the delays of the *q2f* case are not shown because of space limitations in this paper). The delays are computed as the 98th percentile of the delay observed by the packets traversing their corresponding route. In other words, the delay of the route is the time exceeded by the delay of only the 2% most delayed packets. The *q2f* case is not appropriate for networks where the traffic can change in the routes and where there is no reason to give fixed protection to routes.

The second scenario is one using the *q2* case with $P = 0.25$ and $T = 1200s$, with experiments of 1200*s* duration, where it is supposed that the change in traffic at *Route3* is at $t = 0s$ of the experiment, with no further change in traffic during those 1200*s*.

It is also supposed that the traffic of the routes has been the same for a long time before time 0*s* of the experiments, so that as the STP method tends to give to the weight of each queue the same value as its relative traffic-share, then the initial traffic-share values of *Route1* and *Route3* are reflected in the initial values of the weights of the queues of *c1*, which are: 0.5641 and 0.4359, for *Route1* and *Route3* respectively. The gains are evaluated at the ending part of each subsequent interval of 120*s* duration (that is the gain at time 1200*s* is evaluated from 1080 to 1200*s*).

The left side of Fig. 5 shows how the delay in the *q2* case augments with time in every experiment for *Route1* (and decreases for *Route3*). The right side shows how the *Gain1* and *Gain3* of *q2* case tend to be equal to case *q1* at the ending of each experiment (at 1200*s*) and tend to be equal to *q2f* case at the beginning of each experiment (at 120*s*). With this, it can be seen that in the first part of each experiment the *q2* case protects *Route1* as case *q2f* does, and at the ending part of each experiment the *q2* case tend to protect *Route1* as case *q1* does.

It can also be shown, from graphics not presented in this paper, that the total gain, that is the sum of gains (*Gain1* + *Gain3*), is approximately the same for the three cases. The reason for this is that the schedulers for *q2* and *q2f* cases are work conserving, that is, the schedulers serve at full transmission rate whenever there is data to be served, ideally having a service, for the *q2* and *q2f* cases, as big as that of the *q1* case. This is clear as the STP method is not intended to maximize the total gain indicator.

The result of other experiments for *q2* scenarios with parameter $T = 120s$ (instead of 1200*s*), not shown in this paper, show that the gains and delays behave very similar to that of Fig. 5, but they change 10 times faster. This results could give an insight of how to select the parameter *T*, which could be on the order of the mean time between

[8] The standard deviation obtained for the gain was big (for most cases of increased-sources it was bigger than the mean value times 0.5. The reason of this is that the sources are of Pareto type.

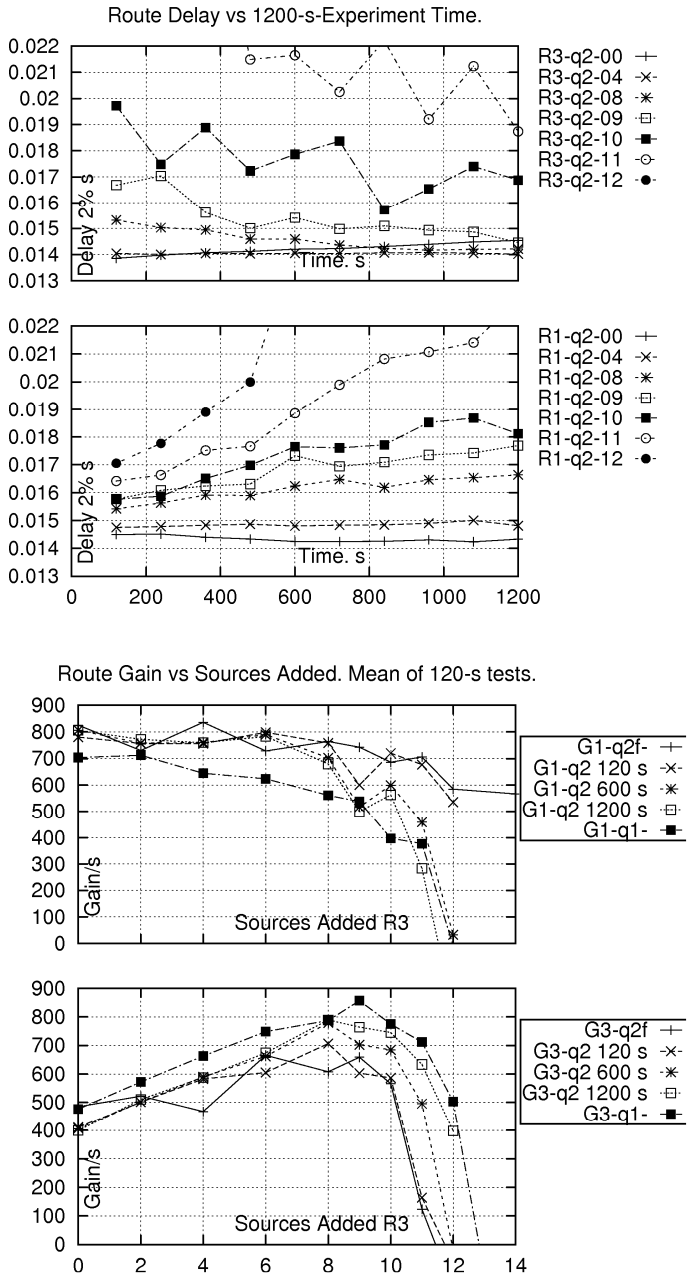


Fig. 5. The left side of the figure shows the delays of routes for *q2* case, throughout the experiment time. Label "R1-q2-04" stands for delay of *Route3* for 4 sources added in *Route3*. The right side of the figure shows the gains of routes for the *q2* case. Label "G1-q2 120 s" stands for *Gain1* of *q2* case at 120*s* of experiment. The other labels have similar meanings.

admission of flows, or the mean time of the life (duration) of the flows, in the network.

An example of a possible disadvantage of the *q2* case is, if at a certain moment *Route3* had a big weight and suddenly it decreased its traffic allowing for *Route1* to increase its traffic, and then, before *Route3* began to lose, substantially, its weight-share, it increased again its traffic, then *Route1* could suffer from this increase, even more than what it would do in a similar situation with a *q1* case.

5 Conclusions and Future Work.

This article addresses the mature problem bandwidth-sharing for QoS-aware networks with per-route distributed admission processes proposing a method which should help the network administrators to have confidence, at least within a time-window, about the amount of bandwidth they count with.

The method allows the nodes to operate autonomously, offering the possibility to attain global behaviors and being scalable. The objectives of the method are: 1- To protect, in the short term, the routes against traffic increments in intersecting routes, and 2- To dynamically assign, in the long term, a bandwidth share proportional to the average measured demands of the routes traversing congested links. As its scheduler is work-conserving this method should not impose a decrease of the transmission capacity in the network. The method's computations are separated by relatively long intervals of time (1*s*) so it should not cause noticeable performance-degradation in actual routers.

Under situation of little traffic the method is not better than the case of using single-queue configurations as there is no need to protect flows against each other.

Bandwidth reservation is still an appreciated form of service which can be a source of important economic compensations; however its overall performance depends heavily on the proper weight assignments which have to be setup by a central authority. On the contrary, the STP method is a simple method that dynamically adapts to the changing network conditions without the need of a central administration.

An analysis about how the STP parameters affect its performance is an important line of research.

References.

1 Centikaya, C., Knightly, E.: Egress Admission Control. In: IEEE INFOCOM 2000, vol. 3, pp. 1471-1480. (2000).

2 Cetinkaya, C., Kanodia, V., Knightly, E.: Scalable Services via Egress Admission Control. IEEE Transactions on Multimedia: Special Issue on Multimedia over IP, vol. 3(1), pp. 69-81. (2001).

3 Parekh, A., Gallager, R.: A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single Node Case. IEEE / ACM Transactions on Networking, vol. 1(3), pp. 346-357. (1993).

4 Neely, M.: Lecture Notes. EE 549 (from lecture 1 to 5). University of Southern California, http://www-rcf.usc.edu/~mjneely/ee549notes/EE549_Supplementary_Lecture_Notes_01.pdf

5 Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., Weiss, W.: RFC 2475. An Architecture for Differentiated Services. The Internet Society (1998).
6 Nichols, K., Blake, S., Baker, F., Black, D.: RFC 2474. Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers. The Internet Society (1998).
7 Kumar, V., Lakshman, T., Stiliadis, D.: Beyond Best Effort: Router Architectures for the Differentiated Services of Tomorrow's Internet. IEEE Communications Magazine, vol. 36(5), pp. 152-164. (1998).
8 Hui, T., Tham, C.: Adaptive Provisioning of Differentiated Services Networks Based on Reinforcement Learning. IEEE Transactions on Systems, Man, and Cybernetics, vol. 33(4), pp. 492-501. (2003).
9 Wang, H., Shen, C., Shin, K.: Adaptive-Weighted Packet Scheduling for Premium Service. In: IEEE International Conference on Communications'2001, pp. 1846-1850. (2001).
10 Li, C., Knightly, E.: Schedulability Criterion and Performance Analysis of Coordinated Schedulers. IEEE/ACM Transactions on Networking, vol. 13(2), pp. 276 – 287. ISSN: 1063-6692. (2005).
11 Floyd, S., Jacobson, V.: Random Early Detection Gateways for Congestions Avoidance. IEEE/ACM Transactions on Networking, vol. 1(4), pp. 397-413. (1993).
12 **ns-2** network simulator, http://nsnam.isi.edu/nsnam/index.php/Main_Page
13 Pieda, P., Ethridge, J., Baines, M., Shallwani, F.: A Network Simulator Differentiated services Implementation. Open IP, Nortel Networks, http://www-sop.inria.fr/members/Eitan.Altman/COURS-NS/DOC/DSnortel.pdf
14 Mrkaic, A. (tutor: U. Fiedler, supervisor: Prof. Dr. B. Plattner). Porting a WFQ Scheduler into ns-2's Diffserv Environment. Computer Engineering and Networks Laboratory (Insitut für Technische Informatik und Kommunikationsnetze). Swiss Federal Institute of Technology. Eidgenössische Technische Hochschule Zürich (2001).
15 Salman, S., Schulzrinne, H.: An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol. Department of Computer Science. Columbia University. 2004, http://www1.cs.columbia.edu/~salman/publications/skype1_4.pdf
16 Skype, https://support.skype.com/en/faq/FA1417/How-much-bandwidth-does-Skype-need;jsessionid=5A4C03A96FCF4A2AE9CE5D72D15093CD?frompage=search&q=how+much+bandwidth+for+a+call&fromSearchFirstPage=false

Modeling information repositories consistency and auditing through Alloy and CCalc

Hector G. Ceballos, Ramon F. Brena and Francisco J. Cantu

Tecnologico de Monterrey
Monterrey, Mexico
{ceballos, ramon.brena, fcantu}@itesm.mx

Abstract. We present a case study on which a model checking language (Alloy) and a non-monotonic causal language (CCalc) are used for modeling the auditing process of an information repository through a MultiAgents System. Alloy and CCalc are used for identifying inconsistencies and simulating an automatic correction process carried out by intelligent agents supervised by human users. Similarities and strengths of both languages are pointed out.

Keywords: MultiAgent Systems, Model checking tools, Independent Choice Logic, Simulation

1 Introduction

For a University, the intellectual production of its professors is an important asset valued by accreditations, rankings and founding organisms [1]. Nevertheless, this information does not reside in a unique repository; rather than it is managed by multiple entities such as editorial companies, organizations and institutions. In most of the cases, each professor compiles and organizes his own publication record.

Maintaining an institutional repository updated and consistent is a permanent time-consuming activity that involves the participation of professors and auditors. This kind of tasks requires intensive application of knowledge in regular basis. The goal is not only performing routines or processes periodically, but translating daily operation in a source of knowledge and learning. Guidance of auditors is indispensable in this case for the correct design and application of policies and norms.

This paper presents the use of two formal languages for the description of the elements and interactions on the described domain: Alloy and CCalc. The former is a model checker used to describe in detail the objects on the domain and the rules that allow verifying the consistency of the information stored in the repository. The latter was used for modeling a MultiAgent System on charge of controlling the auditing process.

This paper is organized as follows. First it is presented the domain and the scenario. Next there are briefly introduced Alloy and CCalc. In section 4 it is described how these languages were used for modeling the domain. Finally we conclude comparing both languages in section 5.

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 169-178

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

2 Information auditing in a University repository

The Tecnologico de Monterrey counts with an information system where information about publications is fed in an institutional repository by the professor itself. Human auditors are responsible for classifying and complementing the information [2]. Auditing is made asynchronously and the results are notified to the professor when changes affect his/her personal record. Additionally, after a publication is registered, its coauthors are notified and empowered to provide additional information or do corrections.

2.1 The Publications Repository

The publications repository registers the scientific production of professors organizing it in an institutional taxonomy. For instance, articles in journals, articles in proceedings and thesis are different types of publications. The information stored in the repository is actually the metadata of the publication; hence we have common data elements like authors, publication date, title, etc.

It is considered additional information and constraints for each type of publication. For example, a journal article is published by a journal, meanwhile that a proceedings article is published in the proceedings of a conference. It is important maintaining a differentiated catalogue of journals and conferences that allows not only quantifying but qualifying professors' scientific production.

There are some common inconsistencies that expert auditors have already detected and modeled. One of them is the duplicity of the publication in the repository and consists on the existence of two publications in the repository having such a degree of similarity that make the auditor suspect that both are in fact the same publication registered twice. This and other types of inconsistencies must be corrected off-line, on regular basis and after information modification.

2.2 The automated auditing process

We propose to implement a MultiAgents System as back-end platform on charge of monitoring, auditing and correcting information feed by professors. The architecture of the system is shown in Figure 1.

Expert auditors define and supervise inconsistency and correction rules. The *LogMonitor* Agent keeps track of actions performed by auditors and professors in the web information system that updates the repository. The *RepGuardian* Agent is responsible for instantiating service agents required for auditing the publications. *Auditor* agents evaluate auditing rules on demand; they are responsible for gathering the necessary information to audit the record and request a correction whenever there is one available and trustworthy. *Corrector* agents apply the correction rules. The *Notifier* Agent communicates the result of auditing to the human expert or professor responsible through a *User* agent.

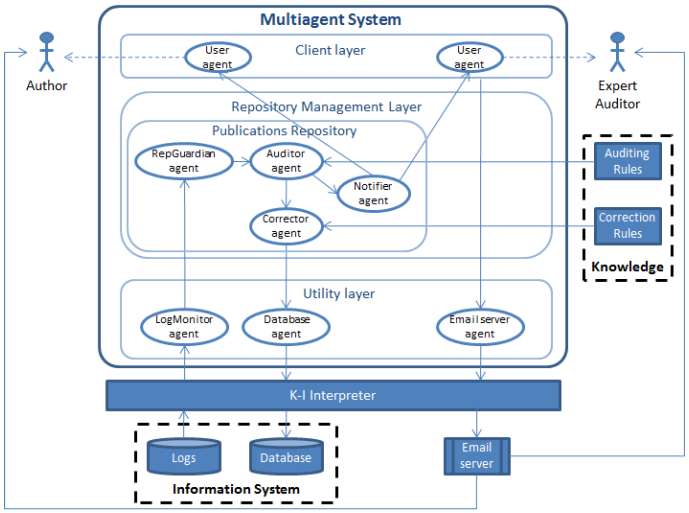


Fig. 1. MultiAgent system architecture for publications auditing.

3 Formal Languages

Current formal languages have started to incorporate modeling primitives such as classes and inheritance that eases the expression of a domain. These object oriented characteristics are exploited during the inference process, allowing generalizing the application of rules and limiting the branching factor. In this paper we explore two formal languages: Alloy and CCalc.

3.1 Alloy Analyzer

Alloy Analyzer [3] is a lightweight modeling language for software design. Its notation uses the sets and relations nomenclature. Alloy is considered a model checking tool because generates all the possible combinations in a certain scope and tries to identify counterexamples for the declared axioms in the model.

Every value in Alloy logic is a relation: a set is a unary relation, scalars are singleton sets, and predicates are n-ary relations. Rows are unordered, columns are ordered but unnamed and all relations are first-order. Alloy syntax define some set constants and operators for representing set operations. Additionally counts with Boolean operators, quantifiers that can be used to declare sets or for construct quantified expressions, and cardinality expressions.

Alloy has upper level constructors such as signatures, represented `sig`, which are used to represent sorts of similar things. Signatures have attributes and constraints associated. A signature can extend other signature, inheriting its attributes and constraints. Facts, represented `fact`, introduce constraints that are assumed to always hold. Functions, represented `fun`, are named expression with declaration parameters and a declaration expression as a result invoked by

providing an expression for each parameter. Predicates, represented `pred`, are named formula with declaration parameters.

Assertions, denoted `assert`, are constraints intended to follow from facts of the model. Assertions doesn't constraint the model, but are checked in the generated model with the command `check`. Check instructs the analyzer to search for counterexamples to a given assertion within a scope. The scope is expressed in quantities of individuals for all available signatures or for every signature individually. The `run` command can be used for instructing the analyzer to search for instances of a given predicate or a given function within a scope.

Alloy can be used for elaborating static and dynamic models. Static models describe a single state where properties are invariant. Facts constraint all possible relations between individuals of the defined signatures. Predicates are used to express definitions, i.e. a named pattern showed in a (set of) individual(s).

Dynamic models describe transitions between states meanwhile predicates are used for describing operations. One way of representing a dynamic model is defining the `state` signature as ordered which allows that each instance of this signature represent a state of the transition model. Predicates receiving two parameters representing contiguous states are used to indicate valid transitions in the model. These predicates are the equivalent of action predicates and hence on its definition must include preconditions and post-conditions; in the post-conditions it must be included frame properties, i.e. indications of the things that must remain unaltered. A special fact denominated `traces` is used to relate all the consecutive states or time frames indicating the valid transitions.

3.2 CCalc

The Causal Calculator (CCalc) is a system for representing commonsense knowledge about action and change. It implements a fragment of the causal logic C+ [4]. The semantics of the language of CCalc is related to default logic and logic programming. Computationally, CCalc uses ideas of satisfiability planning. CCalc runs over Prolog and connects to a SAT solver for generating possible plans.

CCalc uses the nonmonotonic causal logic C+ which through formulas and axioms expresses causal rules. This logic advantages to other action description languages by distinguishing between variant and invariant symbols and assuming that nothing changes unless there is a cause for it. A causal theory in CCalc is constituted by sorts, variables, constants and axioms (causal laws). CCalc allows to express sorts of things identified by a single name. A mechanism of simple inheritance is represented. Variables are typed with the defined sorts and are used in the construction of axioms (rules) to indicate the possible values that can be used in a given predicate.

CCalc define constants instead of predicates. Constants are typed too, allowing not only associating Boolean values to them but objects (instances of sorts) too, which allows expressing the current value of an object property in a given time. Constants can be of type fluent or action. Fluent constants receive any kind of value and can be inertial or rigid, depending of its value is allowed to

change or not in time. Nevertheless, the value of fluent constants doesn't change unless there is an axiom indicating so. A fluent formula is a formula such that all constants occurring in it are fluent constants; meanwhile an action formula is a formula that contains at least one action constant and no fluent constants.

CCalc has a single construct for expressing rules or axioms, called causal laws. Causal laws can express static and dynamic laws which relate events occurring in the same time or in consecutive time frames, respectively. Causal laws have the general form `caused` F `if` G, where F and G are formulas. Through this formalism CCalc can express *static laws* (if F and G are fluent formulas), *action dynamic laws* (if F is a fluent formula and G is an action formula), and *fluent dynamic laws* (if it is added `after` H to the causal law, being H any formula).

Time in CCalc is expressed explicitly at two levels: in dynamic causal laws (through the `after` clause) and on the declaration of facts or queries (indicating the time slice, which ranges from 0 to the `maxstep` variable). CCalc represents a state with a set of instantiated fluent formulas and transitions with events that result of the instantiation of (action or fluent) dynamic causal laws.

CCalc provides a set of abbreviations over its general form that synthetically expresses causal axioms. For instance, the statement `nonexecutable` F `if` G is an abbreviation for `caused` $\perp$ `after` $F \wedge G$, and indicates that the action formula F cannot be executed if G holds.

4 Modeling publications auditing

Alloy was used for modeling the properties of the domain and representing constraints of the domain. On the other hand, CCalc was used for modeling the process involved in information auditing.

4.1 Modeling repository consistency with Alloy

The different types of publications were declared by extending the signature `Publication`. Common metadata was declared as attribute of `Publication`. Specific constraints were declared on each publication type. For instance, see the declaration of articles published in proceedings of conference in Figure 2. The attribute `publishedIn` is constrained to be a single instance of `Proceedings`.

```
sig InProceedingsArticle extends Publication {
publishedIn: one Proceedings
}
```

Fig. 2. Declaration of the InProceedingsArticle signature.

Information like publication's status and the current,previous and actual year were declared through typed constants. The signature `Year` was declared ordered with respect to the predicate `next`.

Consistency rules were expressed through predicates indicating the possible inconsistencies. Figure 3 shows an example of an inconsistency rule. The similarity between publications uses simple comparison between titles, but given that titles are defined as signatures it is possible to use an attribute representing similarity measures. This rule also compares the list of authors.

```
pred samePublication[p1, p2: Publication] {
p1 != p2
p1.title = p2.title and p1.year = p2.year
all a: Person | a in p1.author <=> a in p2.author
}
```

Fig. 3. The SamePublication inconsistency predicate.

For modeling the valid operations in the repository, the signature `Repository` was declared ordered, indicating on each time step which publications were contained in it. Publications were included at model generation and it was simulated their insertion, deletion and modification by introducing and extracting them from the `Repository` state.

Figure 4 shows the operation for adding a publication to the repository; the first two lines represent preconditions and the last one is the postcondition. `r` and `r'` are used for relating two repository states. In the `traces` fact there are indicated four valid operations between consecutive repository states (`r'=r.next`). We also modeled as operations: the deletion of a publication, the change of the publication year and the change of the publication status.

```
pred addPublication[r, r': Repository, p: Publication] {
some p2: Publication | p.id = p2.id => p2 not in r.contains
p not in r.contains
r'.contains = r.contains + p
}
```

Fig. 4. The addPublication operation.

Consistency of the repository was expressed through predicates `IsConsistent(r,p)` and `AllConsistent(r)`, which verifies that all consistency predicates hold for every publication p contained in the repository at state r.

For determining if the given definitions are capable of generating valid models, we used the command `run` with the predicate `AllConsistent`. Limiting to a single repository state we validated (statically) the existence of models satisfying the definitions of publication types and consistency constraints. The Alloy visualizer was useful for inspecting graphically these models.

For validating actual information from the repository, we introduced publication information using constants. Properties like title similarity was calcu-

lated previously. Additionally, the initial conditions for the repository were given through a set of facts.

The detection of inconsistencies was made through asserts. See for instance the assert in Figure 5, which through a command `check` produces models on which the duplicated record inconsistency holds.

```
assert assert_duplicated_pub {
all r: Repository | all disj p1, p2: pubsInRep[r] | not
samePublication[p1, p2]
}
```

Fig. 5. Duplicated publication verification through an assert.

The last task with Alloy was generating a recovery plan for a given example or scenario. The following command asks Alloy for possible models where the last state of the repository is consistent (within 3 states):

`run AllConsistent[ro/last] for 3`

Plans for recovering the consistency when publications $P1$ and $P2$ are the same include: 1) removing $P1$, 2) removing $P2$, and 3) removing $P1$ and $P2$. Identified plans were inspected in the visualizer projecting the solutions along the `Repository` signature. Nevertheless, the predicate that produces the transition was not shown graphically.

4.2 Modeling the Multiagents System with CCalc

CCalc use was used for modeling the agents constituting the proposed Multiagents System. Modeling was divided in three sections: the MAS framework, the domain and the actions. Agents modeling followed the definitions of the Independent Choice Logic (ICL) [5]. The main sorts of the MAS framework were: `agents`, `beliefs` and `entities`. FIPA ACL[1] messages were also represented as sorts.

In the domain were modeled classes of agents, types of beliefs and discourse objects like publications, inconsistencies and persons. Discourse objects are used during agent communication and reasoning and were defined without properties. Unlike Alloy, in CCalc the declaration of a type of object is made separately through sorts and constants. See for instance the declaration of the LogMonitor agent class in Figure 6.

According to ICL, an agent has observables (beliefs) and can perform actions; both are defined through constants where, by convention, the first argument identify the agent class. Observables are defined as inertial fluents (that can be true or false along time); meanwhile actions identifiers are defined as exogenous actions (controlled by the agent itself). Additional agent characteristics can be

[1] FIPA Agent Communication Language Specifications. `http://www.fipa.org/repository/aclspecs.html`

```
:- sorts
agent >> agLogMonitor.
:- constants
believes(agLogMonitor, belNewPublication, publication, person) ::
inertialFluent;
actInform(agLogMonitor, agRepGuardian, belNewPublication, publication,
person):: exogenousAction.
:- constants
monitors(agLogMonitor, repository) :: inertialFluent.
```

Fig. 6. LogMonitor agent class definition in CCalc.

expressed through constants, like in the inertial fluent `checks` that expresses the capability of a LogMonitor agent for monitoring some repository.

External events to the MAS are defined and controlled through two constants: a fluent identifying the occurrence of the event (controlled on the execution) and an action identifying the consequences of the event. For instance, Figure 7 shows the constants used for simulating the insertion of a new publication in the repository.

```
:- constants
exNewPublication(publication, person) :: inertialFluent;
evNewPublication(publication, person) :: exogenousAction.
```

Fig. 7. New publication event constants.

Specific beliefs were declared as subclasses of the sort `belief` in order to reduce the branching factor on model generation. For instance, `belNewPublication` represents the belief an agent has regarding the existence of a new publication in the repository. For instance, see how this belief is used in the declaration of the *LogMonitor* (Figure 6).

Agent's beliefs were initialized on time 0. For example,

```
0: [/\ LM /\ P | -believes(LM, BNP, P, PER)].
```

indicates that all *LogMonitor* agents (`LM`) believe that there is no new publication in the repository (`BNP`) at time 0.

An example of perception of an external event by an agent is shown in Figure 8; whenever a new publication is registered in the repository, the *LogMonitor* agent becomes aware of it. `evNewPublication` is an auxiliary predicate that might consider additional information in the transformation of perceptions into beliefs. Other changes in agent beliefs were also represented by causal rules.

Action descriptions were codified in terms of agent beliefs. For instance, Figure 9 shows the causal laws used for constraining the execution (preconditions) and indicating the effects (postconditions) of the action on which an *Auditor*

```
evNewPublication(P, PER) causes believes(LM, BNP, P, PER),
-exNewPublication(P, PER).
```

Fig. 8. Example of agent perception.

agent (AUD) informs a *RepGuardian* (RG) that the publication P is inconsistent (INC) but has no correction (CO). Note the use of the universal quantifier (\/) and negation (-) for triggering this action when *some* inconsistency has no correction.

```
nonexecutable actInform(AUD, RG, BNR, P) if -[\/ INC \/ CO |
believes(AUD, BIC, P, INC, CO) & believes(AUD, BNR, P, INC)].

caused believes(RG, BNR, P, AUD), -believes(AUD, BIC, P, INC, CO),
-believes(AUD, BNR, P, INC) after believes(AUD, BIC, P, INC, CO) &
believes(AUD, BNR, P, INC) & actInform(AUD, RG, BNR, P).
```

Fig. 9. Example of action description.

Similarly to Alloy, we used constants for declaring actual agents and publications. Their invariant properties were declared through unconditional static causal laws like: `caused checks(aud1, duplicated)`.

CCalc queries were used for: 1) a progressive specification of agents, and 2) identification of plans. In the first case, the specification of the scenario was given step by step and it was verified its feasibility. In this way we debugged agent actions and determined the rules and beliefs required for goal achievement.

Once that agents were fully specified, we calculated those plans capable of satisfy such specifications. For instance, the query in Figure 10 was used for generating a valid plan given two new publications in the repository, each publication with a different type of inconsistency. The goal is expressed in the step previous to the last by indicating that the *RepGuardian* must believe that both publications were audited and that the *User* agent of the person responsible for p2 was notified of the automatic correction of its inconsistency.

```
:- query
label :: 4;
maxstep :: 15;
0: exNewPublication(p1, per), -exNewPublication(p1, per2);
0: exNewPublication(p2, per), -exNewPublication(p2, per2);
0: -aprioriInconsistent(p1, status), aprioriInconsistent(p1, duplicated);
0: aprioriInconsistent(p2, status), -aprioriInconsistent(p2, duplicated);
maxstep-1: believes(rg, audited, p1), believes(rg, audited, p2),
believes(us, corrected, p2, status, changeStatus).
```

Fig. 10. Auditing plan generation.

CCalc generated valid plans showing the sequence of actions executed and those predicates holding on each time step. This query produced a valid plan on 15 steps where each one contained the simultaneous execution of 2 or 3 actions. Running this query with less time steps didn't produce a valid plan, meanwhile a higher number of steps produced many other valid plans.

5 Conclusions

Alloy and CCalc demonstrated to be useful tools in the modeling and construction of the MultiAgent System for the auditing scenario. Both tools use SAT solvers which makes them very efficient for an agile testing and execution.

Alloy's compatibility with the Object Oriented Paradigm eased modeling the domain. Beyond that similarity, Alloy permitted to express internal conditions that should be satisfied in every instance of a given class. Its rich set of operators and constructors allowed to make complex definitions in a few lines. Its tools for visualizing the generated cases eased the specification of domain constraints.

Nevertheless, even when Alloy provides the facilities for representing dynamic models, the specification of frame conditions was cumbersome. In order to satisfy the model, Alloy produces changes during the execution that must be controlled with more frame conditions. Besides, tracking the operations performed during a simulation requires analyzing every two consecutive states, which is not necessary with CCalc due to its printing of the executed actions.

CCalc demonstrated superiority in the specification of frame conditions through the notions of causal laws and inertial fluents. Its abbreviations for expressing causal laws provided an easy implementation (and reading) of the model. Its support for specifying and monitoring concurrency was remarkable.

In both languages it was possible to generate plans towards a given goal and evaluate given scenarios. The Alloy model will permit to evaluate a set of similar publications extracted from the database and evaluate their consistency (in a single state scenario). The CCalc model will allow generating plans that a set of agents should follow in order to audit and correct automatically a set of publications.

References

1. M. Zuckerman, "America's best graduate schools. news & world report." U.S. News & World Report, USA, 2004.
2. F. Cantu, H. Ceballos, S. Mora, and M. Escoffie, "A knowledge-based information system for managing research programs and value creation in a university environment," in *Proceedings of the Eleventh Americas Conference on Information Systems*, (Omaha NE, USA), Association for Information Systems (AIS), August 11-14 2005.
3. D. Jackson, *Software Abstractions: Logic, Language and Analysis*. MIT Press, 2006.
4. E. Giunchiglia, J. Lee, V. Lifschitz, N. McCain, H. Turner, and J. L. V. Lifschitz, "Nonmonotonic causal theories," *Artificial Intelligence*, vol. 153, p. 2004, 2004.
5. D. Poole, "The independent choice logic for modelling multiple agents under uncertainty," *Artificial Intelligence*, vol. 1-2, no. 94, pp. 7–56, 1997.

An Aspect Oriented Approach for the Synchronization of Instance Repositories in Model-Driven Environments

Juan Castrejón

ITESM, Campus Ciudad de México
Calle del Puente 222, 14380, México, D.F., México
A00970883@itesm.mx

Abstract. Software development based on the transformation of meta-model abstractions into particular models and code artifacts is an active research line within software engineering. However, there has been less emphasis on the interoperability of the applications in charge of producing runtime instances of these models. As a consequence, model instances are tied to particular combinations of models and generation tools. If these combinations are to be changed, the data associated to the model instances is at risk of being lost. In this paper, an aspect oriented approach is proposed in order to allow the synchronization of the instance repositories associated to a common meta-model, by using a model bus in charge of receiving and distributing notifications of updates made to the particular object instances. Finally, in order to demonstrate the benefits of the proposed approach, an implementation based on the Eclipse Modeling Framework and Spring Roo is presented.

Keywords: Metamodeling, Software engineering, Software maintenance.

1 Introduction

Model-Driven Software Development (MDSD) has recently gained a great deal of attention in the software engineering area. Reasons for this situation include a perceived effectiveness in avoiding the lack of compliance between design documents and implementation artifacts [1], as well as an increased productivity due to its focus on model abstractions rather than on implementation details [1].

According to this discipline, development teams should aim their attention at modeling the highest abstraction levels of software systems, and then rely on transformation procedures to generate low-level abstractions. In doing so, the model abstractions generated by the development teams do not depend on any specific platform, language or implementation details. This mechanism allows the generation of a set of low-level models that depend on a common high-level model. This is especially useful when a system targets a heterogeneous environment, involving different platforms, tools and programming languages [1].

In this scenario, each environment provides its own model definition (for example UML or Ecore), instance repository technology and access mechanisms (for instance Relational Databases or XML records). These differences can potentially lead to interoperability issues among the different model abstractions [2]. This in turn may prevent an effective synchronization of the data generated by these low level models

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 179-189

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

and the orchestration of the services that their associated tools provide [2]. One of the consequences of this situation is that the changes made in each environment are only kept locally. Therefore, if a particular environment becomes unavailable or if its instance repository is corrupted, the set of instance data associated to this environment is at risk of being lost.

In this paper, an approach based on Aspect Oriented Programming (AOP) techniques is proposed to allow the synchronization of the object instances associated to the repositories of a MDSD environment. It should be noted that the use of AOP is intended in order to conduct the synchronization mechanisms in a non-intrusive manner. Moreover, the addition of the AOP constructs to the application base code is performed during the model transformation procedures, in accordance to the MDSD approach. In this regard, there is a growing set of tools and frameworks intended to support software development based on MDSD principles [1]. Two of these tools are of particular interest for this study, the Eclipse Modeling Framework [3] (EMF) and Spring Roo [4].

EMF represents one of the most matured efforts to support the MDSD approach. It is built on top of the Eclipse platform [3] and is considered by many researches as one of the main environments for model-driven development, due to the size of its community and the number of experimental tools developed around it [3].

Spring Roo is an open source project intended to generate enterprise web applications by means of a set of commands executed through a command-line shell [4]. These commands deal not only with functional requirements, but also with a subset of the non-functional requirements usually associated to web applications [4].

The remaining of this paper is organized as follows. Section 2 introduces work related to the integration and synchronization of models in a MDSD environment. Section 3 contains the full description of the approach proposed by this study. In Section 4, a case study is presented in order to analyze the effectiveness of the proposed approach. Finally, Section 5 summarizes the ideas presented in this paper and introduces future work that might derive from this study.

2 Related Work

Regarding the integration of models and their associated tools in MDSD environments, we can refer to the work by Hein et al. [2], where a model bus is proposed in order to facilitate the orchestration of modeling services. These services represent automated operations over a set of models, such as *creation*, *editing*, *transformation*, *verification* and *execution* [2]. The model bus provides an abstraction layer intended to separate implementation details from the modeling services. This allows for *model representation*, *access* and *location* transparency between the different tools that provide these services. Unlike the approach proposed in this paper, the model bus deals with high-level model abstractions, instead of object instances generated from these models, that is, the lowest-level abstractions in MDSD.

In Breu et al. [5], the requirements and architectural concepts of an infrastructure to support model evolution are discussed. This infrastructure allows the cooperation of different tools that act upon a common system model. Its main functionalities include *model versioning*, *change identification*, *propagation* and *notification* [5]. Its change-

driven nature is similar to the one proposed in this paper. However our main intention is to keep up-to-date instance repositories as opposed to high-level models.

Regarding model evolution, we can mention the work by Herrmannsdoerfer [6]. In this work, the author proposes a workbench to handle *meta-model adaptation*, that is, the evolution of models due to changing requirements and technological progress [6]. The workbench includes the recording of changes made to meta-models, and the corresponding migration processes that the associated models require to conform to these changes. Based on this history of changes, the intention is to automatically migrate the models associated to a common meta-model, during its evolution. The general idea of maintaining a coupled evolution between models is similar to the one proposed in this study. However, the approach described in this paper is more limited in scope, because it does not directly consider the evolution of meta-models.

3 Instance Model Bus

This section introduces the *Instance Model Bus* (IMB), an approach to synchronize instance repositories in model-driven environments. First, the general concepts of this approach are presented, followed by an implementation using EMF and Spring Roo.

3.1 General Approach

As discussed previously, when different models are generated from a common meta-model in a MDSD environment, there is risk of incompatibility between the object instances that each of these models generate. In order to overcome this limitation, a general approach is described in this section in order to automatically synchronize instance repositories. The approach is divided in two main phases, as follows. The first phase is in charge of generating an IMB instance, along with the constructs that allow the modeling tools to send and receive notifications of changes made to the object instances that they manage, using this particular IMB instance. In the second phase, the notification constructs are invoked during the execution of the modeling tools. Fig. 1 summarizes the set of elements, and their interactions, required in a MDSD environment using the IMB.

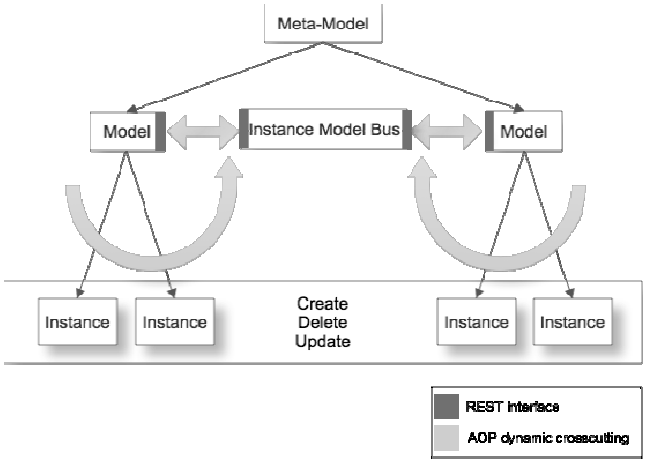


Fig. 1. Elements and activities associated to the Instance Model Bus.

The notification constructs are implemented with the aid of AOP crosscutting techniques. These constructs are generated as part of the transformation procedures from the common meta-model to each of the low-level models. In particular, these constructs advice the tools associated to each of the generated models, in regard to the following operations over the object instances: *create*, *update* and *delete*. When these AOP constructs are executed, they notify the change data to the IMB associated to this particular MDSD environment. In this regard, the IMB instance provides a set of REST interfaces [7] that the modeling tools can invoke for this purpose.

In order to avoid interoperability issues between the data formats used by each instance repository, the definition of transformation procedures that can convert each of these formats into a common data representation is required. In this regard, the use of XML [7] is proposed as the interchange data format for all of the communications to and from the IMB. The combination of REST interfaces and XML data interchange provides a great deal of flexibility, which makes it relatively easy to provide support to a wide range of modeling tools and instance repositories.

Regarding the runtime execution phase, when the IMB receives a notification of change by any of the modeling tools, it propagates the notification through all of the remaining modeling tools registered in the particular MDSD environment. In order to do so, each of the modeling tools should also provide REST interfaces to allow this interchange of data. These interfaces should also be generated during the transformation procedures that generate each low-level model.

Finally, when a modeling tool receives a notification of change by the IMB, it transforms the notification data to the particular format used in its associated instance repository, and performs the requested operation, that is, *create*, *update* or *delete* a particular object instance.

Considering that the previous activities rely on REST communications, the change notifications need to be made persistent by all of the participants. This allows a fallback mechanism in the event of failures preventing communication between the modeling tools, such as network delays or physical failures in their environments.

It should also be noticed that since the operations of the modeling tools are advised through AOP crosscutting, there is no need to manually modify their base code. This

is particularly helpful if development teams have access to the source code of the modeling tools and are in charge of their maintenance. An example of such scenario would be a generated web application acting as a modeling tool. In this way, the proposed approach is non-intrusive at the source code level.

3.2 Implementation

An implementation of the general approach discussed in the previous section is now presented for Java based systems. It is based on EMF and Spring Roo, regarding the generation of the modeling tools following a MDSD approach. For the generation of XML data, to and from the instance repositories, the use of the JAXB specification [8] is proposed. Finally, for the implementation of AOP crosscutting techniques, the use of the AspectJ project [9] is intended. Fig. 2 summarizes the set of tools that were selected for this implementation.

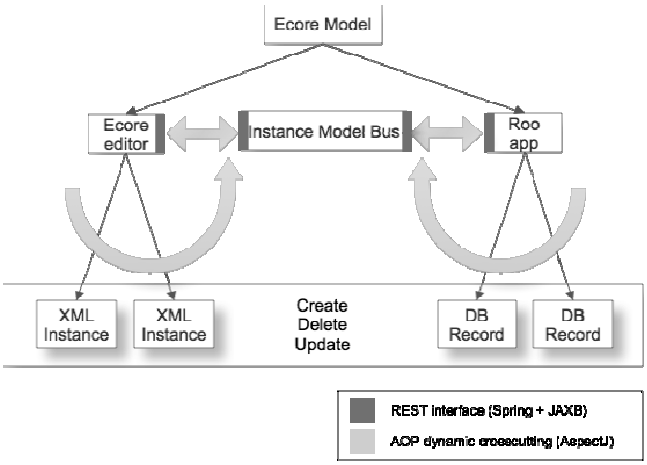


Fig. 2. Elements associated to the Instance Model Bus implementation.

In this implementation, the main meta-model is an instance of the Ecore model, which is the meta-model of the EMF framework [3]. Using the Ecore constructs we can define the static structure of an application. This would be the high-level model in a MDSD environment. We can then generate a set of low-level models taking as base the Ecore model, by using a set of transformation procedures. The current implementation provides two transformations, one to generate an Ecore editor, and a second one to generate a Spring Roo web application. Their details are described next.

The Ecore editor is created using the facilities of the EMF generator projects [3], along with a custom process to customize the data types specified in the associated meta-model. Moreover, in order to generate the REST interfaces and the XML data transformation support, an Eclipse plugin was developed [3]. This plugin is responsible for the creation of a web application associated to the Ecore editor, which

contains the required interfaces and XML support, along with aspect files that advice the *create*, *update* and *delete* operations of the Ecore editor.

The Spring Roo web application is created by generating Spring Roo commands, following a similar approach to the one described in [10]. In order to add REST and XML support to the generated web application, a Spring Roo add-on [4] was developed. This add-on provides a set of Roo commands that generate the aspects files in charge of providing the required interfaces to communicate with the IMB.

It also necessary to generate an IMB instance for this particular MDSD environment. This is achieved by executing an operation of the IMB Eclipse plugin. It should be noted that the generated IMB instance is also a Spring Roo web application. This allows us to reuse some of the Spring Roo generation logic described earlier.

Finally, as appreciated in Fig. 2, the object instances created by the Ecore editor and the Spring Roo application have incompatible data formats. While the editor generates XML files, the web application stores the instance data in database records. The IMB tools allow them to synchronize the data of their instance repositories, by transforming the data from these formats into a common XML definition, and vice versa. This transformation is conducted with the aid of JAXB tools, according to the static structure defined in the common Ecore meta-model.

In order to conclude this section, Fig. 3 depicts the operations of the Eclipse plugin that was developed for the current implementation of the IMB.

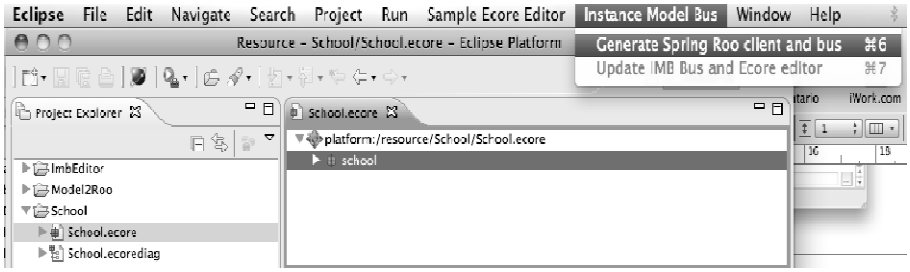


Fig. 3. Operations of the IMB Eclipse plugin.

4 Case Study

This section describes the development of a *Todo list* application, following a MDSD approach. The intention is to demonstrate the benefits of the approach proposed in Section 3, along with the operation of the IMB Eclipse plugin and the Spring Roo add-on described in the previous section.

The *To-do List* is a simple application intended to keep a list of pending tasks for a single user. The modeling of this application can be conducted with the Ecore Diagram editor [3], which is part of EMF. The resulting model is depicted in Fig. 4.

Once the Ecore model of the *To-do list* application is created, we can generate the corresponding Ecore editor, Spring Roo web application and an instance of the IMB for this environment. To this end, we use the IMB Eclipse plugin, and the facilities of the EMF generator projects. The resulting artifacts of this process are described next.

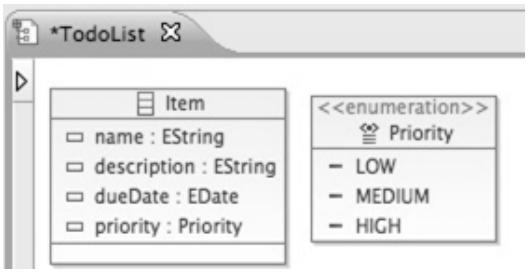


Fig. 4. Modeling of the To-do list application in the Ecore Diagram Editor.

For the creation of the IMB instance, a Spring Roo script is generated. The following program listing depicts the set of commands contained in this script file.

Set of Spring Roo commands that generate the IMB instance for the *To-do list* application.

```
  project --topLevelPackage mx.itesm.todolistbus
  persistence setup --provider HIBERNATE
  --database HYPERSONIC_IN_MEMORY
  jaxb    xsd    compiler    setup    --generateDirectory
src/main/java
  http resource representation setup

  enum type --class ~.domain.ToolName
  enum constant --name Spring
  enum constant --name Eclipse

  entity --class ~.domain.Tool
  field enum --fieldName name --type ~.domain.ToolName
  field string --fieldName description
  field string --fieldName ipAddress
  field string --fieldName port
  field string --fieldName contextPath
  controller all --package ~.web

  enum type --class ~.domain.ImbNotificationType
  enum constant --name CreateEntity
  enum constant --name UpdateEntity
  enum constant --name DeleteEntity
  entity --class ~.domain.ImbNotification
  field string --fieldName url
  field string --fieldName entity
  field string --fieldName entityName
  field string --fieldName returnTypeClass
  field number --fieldName entityId --type java.lang.Long
  field enum --type ~.domain.ImbNotificationType
  --fieldName type
  imb generate notificationScheduling
```

We can appreciate that the Spring Roo commands generate not only the static structure of the application, but also XML and notification scheduling support. In this

case, the static structure represents the modeling tools that can be registered in this MDSD environment, that is, either *Spring* Roo applications or *Eclipse* Ecore editors. The notification scheduling is required in order to persist the change notification data.

The next application to be generated by the IMB Eclipse plugin is the Spring Roo web application. Just like the IMB instance, this application is created from a Spring Roo script, as depicted in the following program listing.

Subset of Spring Roo commands that generate the Spring Roo *To-do list* application.

```
project --topLevelPackage mx.itesm.todolist
persistence setup --provider HIBERNATE
--database HYPERSONIC_IN_MEMORY
jaxb xsd compiler setup
--generateDirectory src/main/java
http resource representation setup

enum type --class ~.domain.Priority
enum constant --name Low
enum constant --name Medium
enum constant --name High

entity --class ~.domain.Item
field string --fieldName name
field string --fieldName description
field enum --fieldName priority
--type ~.domain.Priority
field date --fieldName dueDate --type java.util.Date
controller all --package ~.web
imb update controllers
imb generate schemas
perform command --mavenCommand compile
http resource add oxm
--class imb.domain.todolist.itesm.mx.Item
```

The last application that needs to be generated is the Ecore editor. This is achieved using the standard EMF Generator projects facilities [3]. In particular, we need to create an *EMF Generator Model* from the common meta-model, and then select the *Generate All* option from its associated menu. This last process is depicted in Fig. 5.

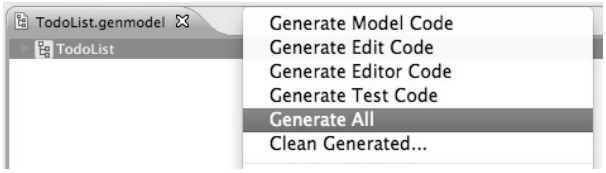


Fig. 5. Generation of the Ecore editor using an EMF Generator model.

Once the generated Spring Roo scripts are executed and the Ecore editor is generated, the aspect files that manage the synchronization of the instance repositories are created. As explained previously, this aspect files depend on the AspectJ project.

The following program listing depicts a subset of the aspect file generated for the Spring Roo application. The first operation advises the execution of the *create* method for instances of the *Item* class. This advice generates a change notification and then registers it with the scheduling program associated to the web application. The scheduling program will in turn deliver this notification to the IMB instance.

The second operation shown in this program listing provides a REST interface that the IMB instance can invoke upon reception of a change notification. This interface transforms from XML to the data format required to create an object instance in the database associated to the web application. Finally, it should be noted that the aspect files generated for the other applications are very similar to the ones described in this section.

Aspect file associated to the Spring Roo *To-do list* application.

```
  after(Item object, BindingResult result, Model model,
        HttpServletRequest request) returning :
  execution (*(mx.itesm.todolist.web.ItemController).
        create(..))   &&   args(object,   result,   model,
request){
      object.setImbId(System.currentTimeMillis()
      + new Random().nextInt());
      object.merge();
      new Thread(new NotificationThread(
          ImbNotificationType.CreateEntity,
          properties.getString("bus.address") +
          "/imb/create/{object}/Spring",
          ItemController_Roo_Imb.transformToImbItem
          (object), "item", object, marshaller)).start();
}

  @RequestMapping(value = "/create/item",
                  method = RequestMethod.PUT)
  public void ItemController.imbCreate(
              @RequestBody Item object) {
      Item modelObject = ItemController_Roo_Imb.
                         transformFromImbItem(object);
      modelObject.persist(); }
```

In order to conclude this section, the interfaces of the applications that were generated in the previous steps are depicted in Fig. 6. These applications synchronize their instance repositories using the aspects advising their operations to *create*, *update* and *delete* instances. For example, if an instance is *created* using the Ecore editor, the Spring Roo application receives a change notification by the IMB, and then creates the instance in its own repository. The same process applies when an object instance is *updated* or *deleted*, either in the Ecore editor or in the Spring Roo web application.

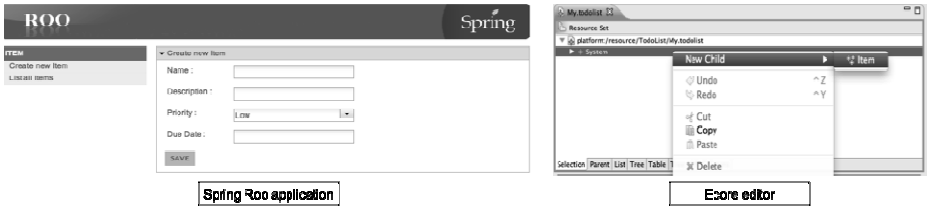


Fig. 6. Applications generated for the To-do list.

5 Conclusions

In this paper, an approach to synchronize the instance repositories within a MDSD environment was proposed. This approach relies on AOP to intercept the operations in charge of maintaining the particular object instances. The change notifications are sent through REST interfaces associated to the modeling tools. Both the aspects and the communication interfaces are generated during the transformation procedures from a common meta-model to particular low-level models.

An implementation of this approach was presented, based on EMF and Spring Roo. The implementation of the REST interfaces was conducted using the facilities of the Spring framework, and the AOP crosscutting was applied using the AspectJ project.

The main contribution of the approach proposed in this paper lies on the ability to automatically synchronize the instance repositories generated from a common meta-model. This is, changes performed over an object instance by any of the modeling tools, are automatically reflected in all of the instance repositories.

It should also be noticed that the current implementation of the *Instance Model Bus* deals with a subset of common non-functional requirements (NFRs), such as maintainability and interoperability of the generated applications. However, support for other types of NFRs, such as debugging and security, will be considered in future implementations. Finally, support for a wider set of programming tools and environments is also intended for future work.

Acknowledgments. The author would like to thank Rosa López-Landa and Dr. Rafael Lozano, both from ITESM Campus Ciudad de México, for their help in discussing and analyzing early ideas that led to the implementation of the approach proposed in this paper.

References

1. Stahl, T., Völter, M.: Model-Driven Software Development. Wiley, New York (2006)
2. Hein, C., Ritter, T., Wagner M.: Model-Driven Tool Integration with ModelBus. Workshop Future Trends of Model-Driven Development (2009)
3. Steinberg, D., Budinsky, F., Paternostro, M.: EMF: Eclipse Modeling Framework. Addison-Wesley Professional, Boston (2008)
4. Spring Roo, http:// www.springsource.org/roo

5. Breu, M., Breu, R., Löw, S.: Living on the MoVE: Towards an Architecture for a Living Models Infrastructure. In: Fifth International Conference on Software Engineering Advances, pp. 290 -- 295. IEEE Computer Society, New York (2010)
6. Herrmannsdoerfer, M.: COPE – A Workbench for the Coupled Evolution of Metamodels and Models. In: 3rd International Conference on Software Language Engineering (2010)
7. Richardson, L., Ruby, S., Heinemeier, D.: Restful Web Services. O'Reilly Media, Sebastopol (2007)
8. Java Architecture for XML Binding, http://jaxb.java.net/
9. Laddad, R.: AspectJ in Action, 2 edition. Manning, Greenwich (2009)
10. Castrejón, J., López-Landa, R., Lozano, R.: Model2Roo: A Model Driven Approach for Web Application Development based on the Eclipse Modeling Framework and Spring Roo. In: 21st International Conference on Electronics, Communications and Computers (2011)

Comparison Between Phase Unwrapping Algorithms for Depth Estimation in 1-D Wavelet Profilometry

Jesus Carlos Pedraza-Ortega[1], Efren Gorrostieta-Hurtado[1], Emilio Vargas-Soto[1], Juan Manuel Ramos-Arreguín[1], Marco Aceves-Fernández[1], Carlos Alberto Olmos-Trejo[1], Sandra Luz Canchola-Magdaleno1, Saúl Tovar-Arriaga[1], Artemio Sotomayor-Olmedo[1],

[1] UAQ, Facultad de Informática, Querétaro, México.
carlos.pedraza@uaq.mx, efrengorrostieta@gmail.com, emilio@mecatronica.net, juan.ramos@uaq.mx, marco.aceves@uaq.mx, caolmos@uaq.mx, sandra.canchola@uaq.mx , saulotovar@hotmail.com, artemio.sotomayor@uaq.mx

Abstract. Transform based profilometry has been widely used for three-dimensional (3-D) surface shape measurement using a projected fringe pattern. These methods require a single image with a sinusoidal fringe pattern projected on it. The projected pattern has a known spatial frequency and its information is used to avoid any discontinuities in the fringes with high frequency. Among single projected fringe pattern, most of the methods use Fourier or Wavelet transforms to extract the phase information. However, they focused only on the transform method and not on the phase unwrapping algorithms. In this paper, a 1-D wavelet profilometry method is presented considering two different wavelet transforms. Later, different phase unwrapping algorithms are used to extract the depth information considering local and global analysis. Several computer simulations and experiments are carried out to validate the proposed method. The merits and limitations of each of these variations on the method are indicated and the error is estimated.

Keywords: Phase unwrapping; depth estimation; algorithms; wavelet transform

1 Background

In order to extract the 3D information of an object, several contact and non-contact measurement techniques have been employed. The main idea is to extract the useful depth information from an image or set of images in an efficient and automatic way. The result of the process (depth information) can be used to guide various tasks such as synthetic aperture radar (SAR), magnetic resonance imaging (MRI), automatic inspection, reverse engineering, 3D robot navigation, interferometry and so on [1]. The contact measurement techniques provide a better way to realize this process by using a vision system together with a tool in contact with the object, like laser, fringe projection and so on. Among all the diverse techniques, one of the most widely used is the fringe projection. Fringe processing methods are widely used in non-destructive testing, optical metrology and 3D reconstruction systems. Some of the desired characteristics in these methods are high accuracy, noise-immunity and fast

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 190-203
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

processing speed. The most known fringe processing methods are the Fourier Transform Profilometry (FTP) method [2], Phase-shifting or phase stepping [3, 4], digital phase locked loop [5], direct phase detection [6], and Wavelet Transform Profilometry [12, 16]. Among all the methods, one of the main challenges is the problem of wrapped phase information problem due to the fact that the phase of a periodically varying intensity pattern is encoded or wrapped and it contains the depth information of the object. The phase unwrapping problem has been attacked by several researchers who have attempted to solve it in many ways. Historically, one of the first algorithms to deal with this problem was proposed by Takeda and Mutoh in 1982 [2]. Later Berryman [7] and Pedraza [8, 17, 18] proposed a modified Fourier Transform Profilometry by carrying out global and local analyses in the phase unwrapping step. Then, unwrapping algorithms (temporal and spatial) were introduced and modified [7-10]. Phase unwrapping techniques use exhaustive data computations and approximations, however, these approaches have a small contribution to understand the cause of failure in the phase unwrapping process. This research presents an implementation of phase unwrapping algorithms, considering the problem of residues.

Generally, most of the proposed methods used a Fourier Transform Profilometry, and another suitable solution is to use the wavelet transform extract the information.

Wavelet transform offers multi-resolution in time and space frequency, and it is a tool that offers advantages over the Fourier transform [9-10]. The computation in the method can be carried out by analyzing the projected fringe patterns using a wavelet transform. Mainly, this analysis consists of demodulating the deformed fringe patterns and extracting the phase information encoded into it and hence the height profile of the object can be calculated, quite similar to Fourier transform.

Different wavelet algorithms are used in the demodulation process to extract the phase of the deformed fringe patterns. Those algorithms can be classified into two categories: phase estimation and frequency estimation techniques. The phase estimation algorithm employs complex mother wavelets, here, the extracted phase suffers from 2π discontinuities and a phase unwrapping algorithm is required to remove these 2π jumps. Zhong et al. [9] have applied Gabor wavelets to extract the phase distribution where a phase unwrapping algorithm is required. The frequency estimation technique estimates the instantaneous frequencies in a fringe pattern, which are then integrated to estimate the phase. The phase extracted using this technique is continuous; consequently, phase unwrapping algorithms are not required for 2D Wavelet Profilometry. Complex or real mother wavelets can be used to estimate the instantaneous frequencies in the fringe pattern. Dursun et al. [14] and Afifi et al. [15] have used Morlet or Paul wavelets, separately, to obtain the phase distribution of projected fringes. Also, Gdeisat et al. [16] have proposed a 1D continuous wavelet transform approach to retrieve phase information in temporally and spatially fringe patterns.

Most of the previous research is focused on using the Fourier and wavelet transforms to obtain the 3D information from an object; pre-filtering the images, extracting the phase information of fringe patterns, using phase unwrapping algorithms, and so on.

In the present research, a comparison between two phase unwrapping algorithms is presented in 1D Wavelet Profilometry is presented in order to obtain the 3D information from an object. First, the spatial frequency of the projected fringe pattern

is obtained; later the mathematical model is obtained and used together with the spatial frequency in order to establish the problem. Then, a 1D Wavelet Profilometry is applied considering the most suitable wavelets for the analysis. Later, three phase unwrapping algorithms are used to obtain the desired 3D information. One contribution of this research is the proposed methodology, because in previous works there are no comparison among different phase unwrapping (PU) methods in 1D Wavelet based profilometry. The results show that it's suitable to compare the present work with other similar researches. To test the method, some virtual objects were created for use in computer simulations and also some experiments were carried out.

2 Profilometry Basics

As described early, there are several fringe projection techniques which are used to extract the three-dimensional information from the objects. The two mostly used techniques (Fourier Transform and Wavelet Profilometry) are presented.

2.1 Fourier Transform Profilometry (FTP)

The image of a projected fringe pattern and an object with projected fringes on it, as shown on figure 3, can be represented by:

$$g(x,y) = a(x,y) + b(x,y) * \cos[2 * \pi f_0 x + \varphi(x,y)] \quad (1)$$

$$g_0(x,y) = a(x,y) + b(x,y) * \cos[2 * \pi f_0 x + \varphi_0(x,y)] \quad (2)$$

where $g(x,y)$ and $g_0(x,y)$ are the intensities of the images at the point (x,y), $a(x,y)$ represents the background illumination, $b(x,y)$ is the contrast between the light and dark fringes, f_0 is the spatial-carrier frequency and $\varphi(x,y)$ and $\varphi_0(x,y)$ are the corresponding phase to the fringe and distorted fringe pattern.

The phase $\varphi(x,y)$ contains the desired information. This angle $\varphi(x,y)$ is the phase shift caused by the object surface end the angle of projection, and its expressed as:

$$\varphi(x,y) = \varphi_0(x,y) + \varphi_z(x,y) \quad (3)$$

where $\varphi_0(x,y)$ is the reference plane projected phase angle, and $\varphi_z(x,y)$ is the object's height distribution phase.

In Pedraza et al work [17, 18], the Equation 5 can be rewritten as:

$$\varphi_z(x,y) = \frac{h(x,y)2\pi f_0 d_0}{h(x,y) - l_0} \quad ; h(x,y) = \frac{l_0 \phi_z(x,y)}{\phi_z(x,y) - 2\pi f_0 d_0} \quad (4)$$

where the value of $h(x,y)$ is measured and considered as positive to the left side of the reference plane. Also, the Equation 4 expresses the height distribution as a function of the phase distribution.

The Equation 1 can be rewritten as:

$$g(x,y) = \sum_{n=-\infty}^{\infty} A_n r(x,y) \exp(in\varphi(x,y)) * \exp(i2\pi n f_0 x) \quad (5)$$

where r(x,y) is the reflectivity distribution on the diffuse object [3,4]. Then, a FFT (Fast Fourier Transform) is applied to the signal in the x direction only. Thus, the following equation is obtained:

$$G(f,y)=\sum_{-\infty}^{\infty}Q_n(f-nf_0,y) \tag{6}$$

where Q_n is the 1D Fourier Transform of $An\ exp[in\varphi(x,y)]$.

Here $\varphi(x,y)$ and $r(x,y)$ vary very slowly in comparison with the fringe spacing, then the Q peaks in the spectrum are separated from each other. It is also necessary to consider that if a high spatial fringe pattern is chosen, the FFT will have a wider spacing among the frequencies; this behavior helps to identify the fundamental peak f_0. In FTP, next step is to remove all signals but positive fundamental peak f_0. Then, the result is shifted and centered. Later, the IFFT (Inverse Fast Fourier Transform) is applied in the x direction only. Here, is necessary to separate the phase part of the result from the rest because it contains the depth information:

$$\begin{aligned}\varphi_z(x,y)&=\varphi(x,y)+\varphi_0(x,y)\\&=\mathrm{Im}\{\log(\hat{g}(x,y)\hat{g}_0^{*}(x,y))\}\end{aligned} \tag{7}$$

The whole phase map is obtained by applying the same procedure for each x line. The result is that the values of the phase map are wrapped at some specific values whose range lie between π and $-\pi$. Then, to recover the true phase it is necessary to restore the measured wrapped phase by an unknown multiple of $2\pi f_0$ [17]. However, to analyze and describe signals, it requires information from both domains time and frequency, therefore Fourier is not a suitable solution to express those signals and another way is proposed, wavelet transform.

2.2 Wavelet Transform Profilometry

The wavelet transform (WT) is considered an appropriate tool to analyze non-stationary signals. This technique has been developed as an alternative approach to the most common transforms, such as Fourier transform, to analyze fringe patterns. Furthermore, WT has a multi-resolution property in both time and frequency domains which solves a commonly know problem in other transforms like the resolution.

A wavelet is a small wave of limited duration (this can be real or complex). For this, two conditions must be satisfied: firstly, it must have a finite energy. Secondly, the wavelet must have an average value of zero (admissibility condition). It is worth noting that many different types of mother wavelets are available for phase evaluation applications. The most suitable mother wavelet is probably the complex Morlet one [2]. The Morlet wavelet is a plane wave modulated by a Gaussian function, and is defined as:

$$\psi(x)=\pi^{1/4}\exp(icx)\exp(-x^2/2) \tag{8}$$

where c is a fixed spatial frequency, and chosen to be about 5 or 6 to satisfy an admissibility condition [11]. Figure 1 shows the real part (dashed line) and the imaginary part (solid line) of the Morlet wavelet.

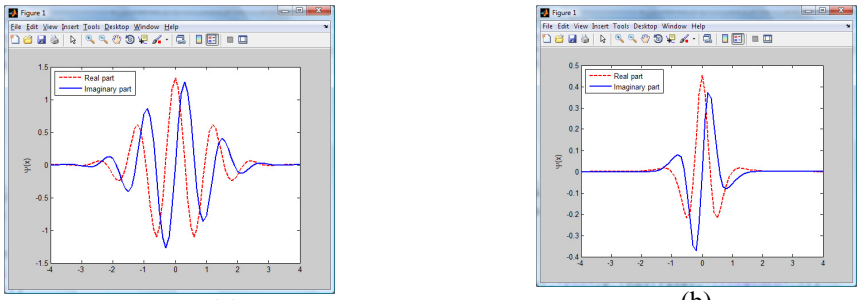


Figure 1. Mother Wavelets: (a) Complex Morlet and (b) Paul.

Also, Paul Wavelet is considered as one choice to perform the phase evaluation and is defined as:

$$\psi(x) = \frac{2^n n!(1-ix)^{-(n+1)}}{2\pi\sqrt{\frac{(2n)!}{2}}} \tag{9}$$

where n is the order of the Paul mother wavelet and chosen to have the value of 5.

The one-dimensional continuous wavelet transform (1D-CWT) of a row $f(x)$ of a fringe pattern is obtained by translation on the x axis by b (with y fixed) and dilation by s of the mother wavelet $\psi(x)$ as given by:

$$W(s,b) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} f(x)\psi^*\left(\frac{x-b}{s}\right)dx \tag{10}$$

here, * denotes complex conjugation and $W(s,b)$ is the calculated CWT coefficients which refers to the closeness of the signal to the wavelet at a particular scale.

In this research, the phase estimation and frequency estimation methods are used to extract the phase distribution from two dimensional fringe patterns. In the phase estimation method, a complex Morlet and Paul wavelets will be applied to a row of the fringe pattern. The resultant wavelet transform is a two dimensional complex array, where the phase arrays can be calculated as follows:

$$abs(s,b) = |W(s,b)| \tag{11}$$

$$\varphi(s,b) = \tan^{-1}\left(\frac{\Im\{W(s,b)\}}{\Re\{W(s,b)\}}\right) \tag{12}$$

To compute the phase of the row, the maximum value of each column of the modulus array is determined and then its corresponding phase value is found from the phase array. By repeating this process on all rows of the fringe pattern, a wrapped phase map results and an unwrapping algorithm is then needed to unwrap it.

In the frequency estimation method, a complex Morlet wavelet and Paul wavelet are applied to a row of the fringe pattern. The resultant wavelet transform is a two dimensional complex array. The modulus array can be found using Equation (14) and hence the maximum value for each column and its corresponding scale value can be determined. Considering that we are interested in the 1D signal:

$$f(x) = a(x) + b(x)\cos(2\pi f_0 x + \varphi(x)) \tag{13}$$

Considering the Euler identity for cos(x), we can re-write the Equation 13 as:

$$f(x) = a(x) + b(x)\cos(\varphi(x)) = a(x) + b(x)\frac{e^{i\varphi(x)}}{2} + b(x)\frac{e^{-i\varphi(x)}}{2} \tag{14}$$

The analytic function *f* in an open interval A, where $z_0 \in A$ can be decomposed into Taylor series:

$$f(z) = \sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!}(z - z_0)^k \tag{15}$$

Therefore:

$$\varphi(x) = \varphi(x) + \varphi'(x)(b)(x-b) + \frac{\varphi''(x)}{2!}(x-b)^2 + \frac{\varphi'''(x)}{3!}(x-b)^3 + \ldots \tag{16}$$

If :

$$\varphi(x) = \varphi(b) + \varphi''(b) \approx 0,\ \varphi'''(b) \approx 0,\ \varphi'^{v}(b) \approx 0, \cdots, \varphi^k(b) \approx 0 \tag{17}$$

Then, the function can be reduced as:

$$\varphi(x) = \varphi(b) + \varphi'(b)(x-b) \tag{18}$$

Moreover, the Morlet Wavelet is defined as $\psi(x) = e^{i\varpi_0 x}e^{-\frac{x^2}{2}}$, this wavelet will be applied to the mother wavelet $W(s,b) = \frac{1}{s}\int_{-\infty}^{\infty} f(x)\psi^*\left(\frac{x-b}{s}\right)dx$:

If s = 1, then:

$$\begin{aligned} W(s,b) &= \frac{1}{s}\int_{-\infty}^{\infty} f(x)\psi^*\left(\frac{x-b}{s}\right)dx \\ &= \int_{-\infty}^{\infty}\left[a(x) + b(x)\frac{e^{i\varphi(x)}}{2} + b(x)\frac{e^{-i\varphi(x)}}{2}\right]\psi^*\left(\frac{x-b}{s}\right)dx \\ &= a\int_{-\infty}^{\infty}\psi^*\left(\frac{x-b}{s}\right)dx + \frac{b}{2}\int_{-\infty}^{\infty}\frac{e^{i[\varphi(b)+\varphi'(b)(x-b)]}}{2}\psi^*\left(\frac{x-b}{s}\right)dx + \frac{b}{2}\int_{-\infty}^{\infty}\frac{e^{-i[\varphi(b)+\varphi'(b)(x-b)]}}{2}\psi^*\left(\frac{x-b}{s}\right)dx \end{aligned} \tag{19}$$

By solving Equation 19, the following equation is obtained:

$$a\sqrt{2\pi}e^{-\frac{1}{2}\varpi_0^2} + \frac{b}{2}\sqrt{2\pi}e^{-\frac{1}{2}(\varpi_0 + s\varpi_s)^2}e^{-i\varpi_s b} + \frac{b}{2}\sqrt{2\pi}e^{-\frac{1}{2}(\varpi_0 - s\varpi_s)^2}e^{i\varpi_s b} \tag{20}$$

Then the instantaneous frequencies are computed using the next Equation [11]:

$$\hat{f}(b) = \frac{c + \sqrt{c^2+2}}{2s_{\max}(b)} - 2\pi f_o \tag{21}$$

where f_0 is the spatial frequency. At the end, the phase distribution can be extracted by integrating the estimated frequencies.

The same procedure can be developed to get the instantaneous frequencies, which lead us to have the wrapped phase and therefore it is necessary to apply a phase unwrapping algorithm.

3 Phase Unwrapping

Since two decades ago, phase unwrapping has been a research area and many papers have been published, presenting some ideas that solves the problem. Several phase unwrapping algorithms have been proposed, implemented and tested.

The phase unwrapping process is not a trivial problem due to the presence of phase singularities (points in 2D, and lines in 3D) generated by local or global undersampling. The correct 2D branch cut lines and 3D branch cut surfaces should be placed where the gradient of the original phase distribution exceeded π rad value. However, this important information is lost due to undersampling and cannot be recovered from the sampled wrapped phase distribution alone. Also, is important to notice that finding a proper surface, or obtaining a minimal area or using a gradient on a wrapped phase will not work and one could not find the correct branch in cut surfaces.

The phase unwrapping has many applications in applied optics that require an unwrapping process, and hence many phase unwrapping algorithms has been developed specifically for data with a particular application. Moreover, there is no universal phase unwrapping algorithm that can solve wrapped phase data from any application. Therefore, phase unwrapping algorithms are considered as a trade-off problem between accuracy of solution and computational requirements. However, even the most robust and complete phase unwrapping algorithm cannot guarantee in giving successful or acceptable unwrapped results without a good set of initial parameters. Unfortunately, there is no standard or technique to define the parameters that guarantee a good performance on phase unwrapping.

To face the phase unwrapping problems, algorithms can be divided in two categories: local and global phase unwrapping. Local phase unwrapping algorithms find the unwrapped phase values by integrating the phase along a certain path. This is called path-following algorithms [6].

Global phase unwrapping algorithms locate the unwrapped phase by minimizing a global error function and are also called local phase unwrapping algorithm and a global phase unwrapping algorithm, by following the methodology proposed by Pedraza in [1]. The unwrapped phase values and the wrapped phase can be related with each other as:

$$\Psi(n) = \varphi(\pi) + 2\pi k(n) \quad -\pi < \Psi(n) \leq \pi \tag{22}$$

$$\varphi(n) = \Psi(n) + 2\pi v(n) \quad -\infty < \varphi(n) \leq \infty \tag{23}$$

here Ψ(n) holds the wrapped phase values and φ(n) holds the unwrapped phase values, k(n) is the function containing the integers that must be added to the wrapped phase φ to be unwrapped, n is an integer and v(n) is the function containing a set of integers that must be added to the wrapped phase Ψ.

Noting that;

$$v(n) = -k(n) \tag{24}$$

The wrapping operation ω which converts the unwrapped phase is defined by:

$$\varpi\{\varphi(n)\} = \arctan\left[\frac{\sin(\varphi(n))}{\cos(\varphi(n))}\right] \tag{25}$$

3.1 Local Phase unwrapping

Local phase unwrapping algorithms finds the unwrapped phase values by integrating the phase along certain path that covers the whole wrapped phase map. The local phase unwrapping defines the quality of each pixel in the phase map to unwrap the highest quality pixels first and the lowest quality pixels last (quality-guided phase unwrapping). For this purpose, the methods known as residue-balancing are proposed, which attempts to prevent error propagation by identifying residues (the source of noise in the wrapped phase). The residues must be balanced and isolated by using barriers (branch-cuts), therefore, it aims to produce a path-independent wrapped phase map. Path-dependency occurs to the existence of residues.

Residue-balancing algorithms search for residues in a wrapped-phase map and attempt to balance positive and negative residues by placing cut lines between them to prevent the unwrapping path breaking the mesh created. The residue is identified for each pixel in the phase map by estimating the wrapped gradients in a 2 × 2 closed loop, as shown in Figure 2.

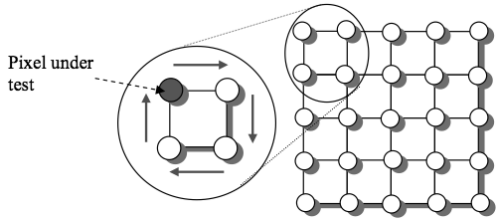


Figure 2. Identifying residues in a 2 × 2 closed path.

This is carried out using the following equation:

$$r = \Re\left[\frac{\Psi_{i,j} - \Psi_{i+1,j}}{2\pi}\right] + \Re\left[\frac{\Psi_{i+1,j} - \Psi_{i+1,j+1}}{2\pi}\right] + \Re\left[\frac{\Psi_{i,j+1} - \Psi_{i,j}}{2\pi}\right] \tag{26}$$

Where $\Re[]$ rounds its argument to the nearest integer, $\Psi_{x,y}$ is the wrapped pixel. The equation 13 can only take three possible results: 0, +1, and -1. A pixel under test is considered to be a positive residue if the value of r is +1, and it is considered to be a negative residue if the value is -1. Conversely, the pixel is not a residue if the value of r is zero. After identifying all residues in the wrapped phase map, these residues have to be balanced by means of branch cuts. Branch-cuts act as barriers to prevent the unwrapping path going thorough them. If these branch cuts are avoided during the unwrapping process, no errors propagate and the unwrapping path is considered to be path independent. On the other hand, if these branch cuts are penetrated during the unwrapping, errors propagate throughout the whole phase map, and in this case the unwrapping path is considered to be path dependent.

3.2 Global phase unwrapping

In the previous section, it was stated that local phase unwrapping algorithms follow a certain unwrapping path in order to unwrap the phase. They begin at a grid point and integrate the wrapped phase differences over that path, which ultimately covers the entire phase map. Local phase unwrapping algorithms (residue-balancing algorithms) generate branch cuts and define the unwrapping path around these cuts in order to minimize error propagation.
In contrast, global phase unwrapping algorithms formulate the phase unwrapping problem in a generalized minimum-norm sense [6]. Global phase unwrapping algorithms attempt to find the unwrapped phase by minimizing the global error function as shown in equation 14

$$\varepsilon^2 = ||\, solution - problem \,||^2 \tag{27}$$

Global phase unwrapping algorithms seek the unwrapped phase whose local gradients in the *x* and *y* direction match, as closely as possible.

$$\varepsilon^2 = \sum_{i=0}^{M-2}\sum_{j=0}^{N-1}\left|\Delta^x\varphi(i,j) - \hat{\Delta}^x\psi(i,j)\right|^p + \sum_{i=0}^{M-1}\sum_{j=0}^{N-2}\left|\Delta^y\varphi(i,j) - \hat{\Delta}^y\psi(i,j)\right|^p \tag{28}$$

Where $\Delta^x\varphi(i,j)$ and $\Delta^y\varphi(i,j)$ are unwrapped phase gradients in the *x* and *y* directions respectively, which are given by:

$$\Delta^x\varphi(i,j) = \varphi(i+1,j) - \varphi(i,j) \tag{29}$$

$$\Delta^y\varphi(i,j) = \varphi(i,j+1) - \varphi(i,j) \tag{30}$$

$\hat{\Delta}^x\psi(i,j)$ and $\hat{\Delta}^y\psi(i,j)$ are the wrapped values of the phase gradients in the *x* and *y* directions respectively, and they are given by:

$$\hat{\Delta}^x\psi(i,j) = \omega\{\psi(i+1,j) - \psi(i,j)\} \tag{31}$$

$$\hat{\Delta}^y\psi(i,j) = \omega\{\psi(i,j+1) - \psi(i,j)\} \tag{32}$$

Finally the wrapping operator is defined by the equation 25.

4 Setup and proposed Methodology

Considering Figure 3, we have a fringe which is projected from the projector, the fringe reaches the object at point H and will cross the reference plane at the point C. By observation, the triangles DpHDc and CHF are similar and since:

$$\frac{CD}{-h} = \frac{d_0}{l_0} \tag{33}$$

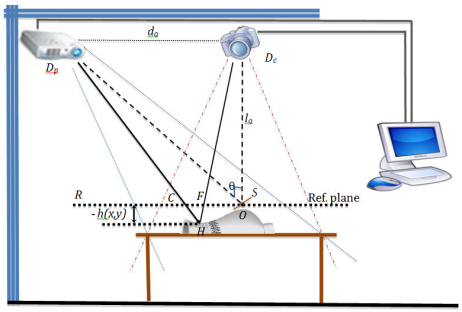


Figure 3. Experimental setup.

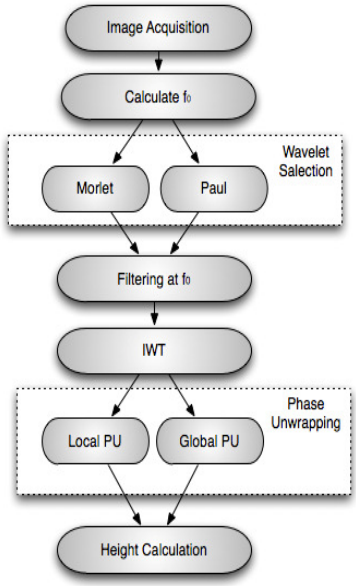
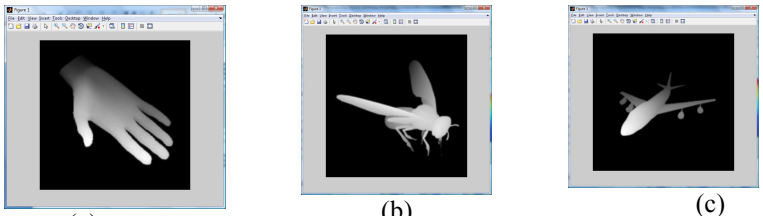


Figure 4. Proposed methodology.

The experimental setup shown in Figure 3 is proposed and during the experiments the methodology in Figure 3 is applied. The first step is to acquire the image. Due to the nature of the image, sometimes a filtering to eliminate the noise is necessary, and a filter is used. Next, the fundamental frequency *fo* is estimated. Later, the mother is selected (Morlet or Paul) and applied. The filter at *fo* is carried out and the Inverse Wavelet transformation is done. At this stage, the information of the height is phase wrapped and two phase unwrapping algorithms are proposed: Local and Global Analysis Algorithm and Graph Cuts Algorithm. The final step is to obtain the object reconstruction and in some cases to determine the error (in case of virtual created objects). The experimental setup uses a high-resolution digital CCD camera and a high resolution digital projector.

The object of interest can be any three-dimensional object and for this work, three objects are considered, which are shown on figure 5.

It is also important to develop software able to produce several different fringe patterns. To create several patterns, it is necessary to modify the spatial frequency (number of fringes per unit area), and resolution (number of levels to create the sinusoidal pattern) of the fringe pattern. It may also be necessary to include into the software development a routine capable of performing phase shifting as well as to include the horizontal or vertical orientation projection of the fringe pattern.



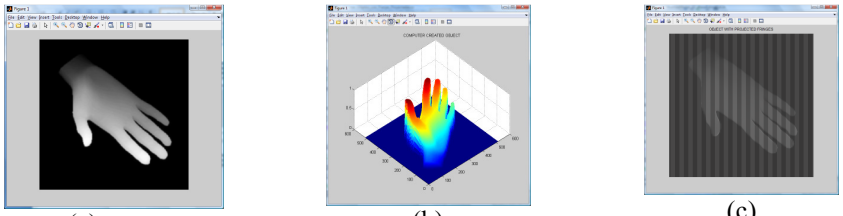
(a) (b) (c)

Figure 5. Virtual objects used in the test: (a) Hand, (b) Fly, and (c) Airplane.

5 Results

To test the methodology, first an object with Hand shape is used. Then, a sinusoidal fringe pattern of known spatial frequency is created with 128 fringes and added to the shape of the created object. The resulting image is shown in Figure 6(c). It is worth noting the distortions of the fringe pattern due to the object´s shape.

The wrapped phase and its mesh are shown in Figure 6. The reconstructed Hand using the Morlet Wavelet Transform and applying the Local PU Analysis and the Global PU Analysis can be seen in Figure 7. Notice that, by applying this method, the shape of the Hand looks almost equal, but it has an error magnitude of about 3.2 and 2.1% respectively.



(a) (b) (c)

Figure 6. Computer created Hand: (a) Object image, (b) Object mesh, and (c) fringes projected on it.

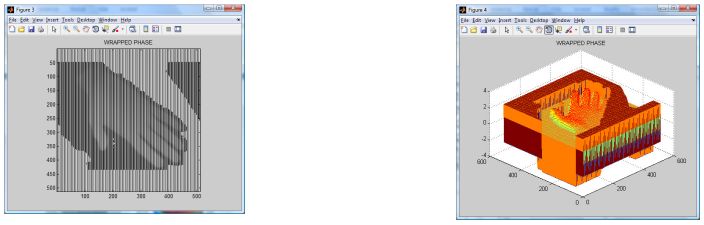


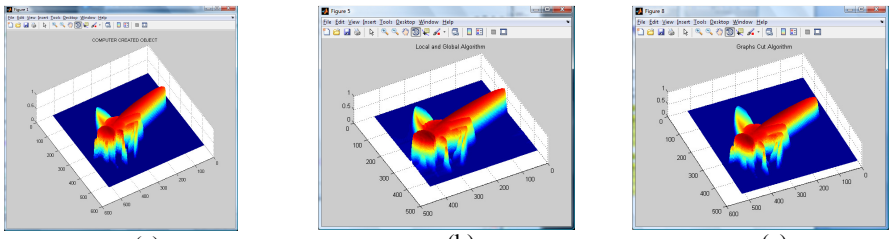
Figure 7. Wrapped phase (image and mesh).

(a) (b)

Figure 8. Reconstructed object using Morlet Wavelet and (a) Local Phase Unwrapping Algorithm, and (b) Global Phase Unwrapping Algorithm.

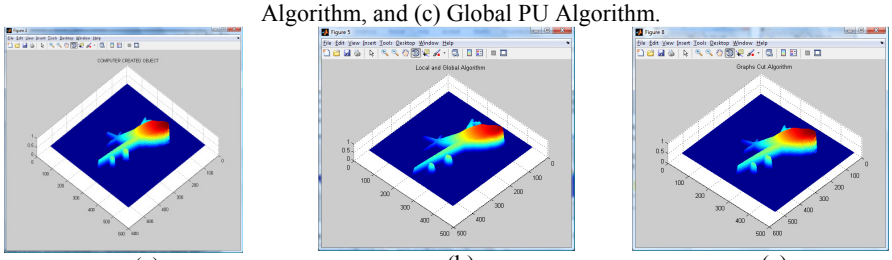
The results on Figure 8 show that the whole volume presents an acceptable error and the shape is well defined. The mother wavelet used was the Morlet but the same experiment was conducted for the Paul wavelet ant the results are presented on. The computer simulation allowed us to test and proposed methodology.

To validate the whole methodology, more experiments were conducted considering the objects observed on Figure 5. Those objects have different shapes (computer created), where the height is known in every point in the object. Then, the Morlet and Paul mother Wavelets are considered as well as the two different phase unwrapping algorithms. As a second experiment, Paul Wavelet is used and height of the virtual object was compared with each one of the analysis and the results are presented in figure 9.



(a) (b) (c)

Figure 9. Fly object and its reconstruction using Paul Wavelet: (a) Object, (b) Local PU Algorithm, and (c) Global PU Algorithm.



(a) (b) (c)

Figure 10. Airplane object and reconstruction using Morlet Wavelet: (a) Object, (b) Local PU Algorithm, and (c) Global PU Algorithm.

Third experiment was conducted with Morlet Wavelet and the object used was the airplane, and the respective results of the reconstruction process can be seen in Figure 10. Later, all the experiments are joined and the error magnitude is enclosed on tables

1 and 2. The results show that the better performance was obtained by using the Morlet wavelet together with the Global Phase Unwrapping Algorithm in final step to do the 3D reconstruction process.

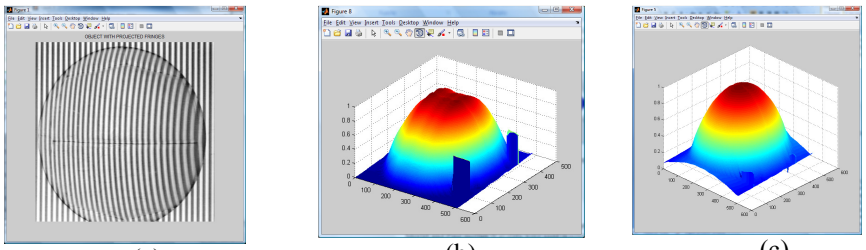
Table 1. Error table using Morlet Wavelet.

Object	Local and Global	Graph Cuts
Hand	3.26	2.11
Fly	3.47	2.21
Airplane	3.51	2.18

Table 2. Error table using Paul Wavelet.

Object	Local and Global	Graph Cuts
Hand	4.37	3.43
Fly	4.65	3.76
Airplane	4.71	3.55

Finally, the performance of the proposed methodology was tested in a real object (Volleyball) and both Morlet and Paul wavelets were used considering the Global Phase Unwrapping Algorithm for the phase unwrapping and the results can be observed in Figure 11.



(a) (b) (c)

Figure 11. Real object and reconstruction: (a) Object, (b) Using Morlet Wavelet, and (c) Paul Wavelet with Graph Cuts Algorithm.

6 Conclusions and Future Work

In this work, an enhanced Wavelet based Profilometry was presented and tested. Both Morlet and Paul mother Wavelets were used in conjunction with Local and Global techniques, as well as Graph Cuts Algorithms in the phase unwrapping process. Three different objects generated by the computer were utilized (Hand, Fly and Airplane). The object's projected fringe pattern has a known spatial frequency. Also, a real object was chosen and the methodology was carried out with an accurate reconstruction of the object. Among the proposed wavelets, the one who shows a better performance was the Morlet wavelet in comparison with Paul wavelet, because

Morlet was the one that produced a minimal error. As a conclusion, we can say that the proposed methodology could be used to digitize diverse objects with good results. As a future work, the software performance can be improved in order to implement it inside an embedded system.

References

1. Gokstorp, M. Depth Computation in Robot Vision; Ph.D. Thesis, Department of Electrical Engineering, Linkoping University: Linkoping, Sweden, 1995.
2. Takeda, M.; Ina, H.; Kobayashi, S. Fourier-Transform method of fringe pattern analysis for computed-based topography and interferometry. J. Opt. Soc. Am. 1982, 72, 156-160.
3. Grevenkamp, J.E.; Bruning, J.H. Phase-shifting interferometry. In Optical Shop Testing; Malacara, D. Ed.; Wiley: New York, NY, USA, 1992.
4. K. Creath, in: W.R. Robinson, G.T. Reid (Eds.), Interferogram Analysis: Digital Fringe Pattern Measurement Techniques, Institute of Physics, Philadelphia, PA, 1993.
5. M. Gdeisat, D. Burton, M. Lalor, Appl. Opt. 39 (2000) 5326.
6. Y. Ichioka, M. Inuiya, Appl. Opt. 11 (1972) 1507.
7. Berryman, F.; Pynsent, P.; Cubillo, J. A theoretical Comparison of three fringe analysis methods for determining the three-dimensional shape of an object in the presence of noise. Opt. Lasers Eng. 2003, 39, 35-50,
8. Pedraza, J.C. et al, Image Processing for 3D Reconstruction using a Modified Fourier Transform Profilometry Method; Springer-Verlag: Berlin, Germany, 2007; pp. 705–712.
9. Rastogi, P.K. Digital Speckle Pattern Interferometry and Related Techniques; Wiley: New York, NY, USA, 2001.
10. Itoh, K. Analysis of the phase unwrapping algorithm. Appl. Opt. 1982, 21, 2470-2486.
11. Wu, L.S. Research and development of fringe projection-based methods in 3D shape reconstruction. J. Zhejiang Univ. Sci. A 2006, 7, 1026-1036,
12. Zhong, J.G.; Wang, J.W. Spatial carrier-fringe pattern analysis by means of wavelet transform: wavelet transform profilometry. Appl. Opt. 2004, 43, 4993-4998.
13. Zang, Q.; Chen, W.J.; Tang, Y. Method of choosing the adaptive level of discrete wavelet decomposition to eliminate zero component. Opt. Commun. 2009, 282, 778-785.
14. Dursun, A.; Ozder, S.; Ecevit, N. Continuous wavelet transform analysis of projected fringe pattern analysis. Meas. Sci. Tech. 2004, 15, 1768-1772.
15. Afifi, M. et al, Wavelet-based algorithm for optical phase distribution evaluation. Opt. Commun. 2002, 211, 47-51.
16. Gdeisat, M.A.; Burton, D.R.; Lalor, M.J. Spatial carrier fringe pattern demodulation using a two-dimensional continuous wavelet transform. Appl. Opt. 2006, 45, doi:10.1364/AO.45.008722.
17. Pedraza-Ortega, J.C. et al, Three-Dimensional Reconstruction System based on a Segmentation Algorithms and a Modified Fourier Transform Profilometry. In Proceedings of the IEEE Electronics, Robotics and Automotive Mechanics Conference, Cuernavaca, Morelos, September, 2009.
18. Pedraza-Ortega, J.C. et al, A Profilometric Approach for 3D Reconstruction Using Fourier and Wavelet Transforms. In MICAI 2009: Advances in Artificial Intelligence; Springer Berlin-Heidelberg: Berlin, Germany, 2009; pp. 313-323.
19. Gdeisat, M.A.; et al, Spatial and Temporal Carrier Fringe Pattern Demodulation using the One-Dimensional Continuous Wavelet Transform: Recent Progress, Challenges and Suggested Developments. Opt. Lasers Eng. 2009, 47, 1348-1361.

A review on Enterprise Resource Planning System Selection Process

Augusto A Pacheco-Comer, Juan C González-Castolo

Universidad de Guadalajara, CUCEA Information Technologies Department, Periferico Norte 799, esquina Av. Parres Arias, Nucleo Universitario Los Belenes, Zapopan, México, C.P.45100
augustopacheco@acm.org, jcgcastolo@cucea.udg.mx

Abstract. Enterprise Resource Planning system is one of the information system most implemented by organizations. Their use can be seen at small, medium and big enterprises. It is one of the most important business optimization projects that an enterprise could attempt. At this paper we present a research literature review regarding enterprise resource planning implementation models, critical success factors and selection and evaluation criteria. A proposal of implementation methodology and there are suggestions for future research on selection processes and methodologies using Petri Nets and Multi Agents System.

Keywords: ERP; enterprise resource planning; selection; implementation model; information system; evaluation criteria.

1 Introduction

Information and communication technologies (ICT) allow an organization to gain a competitive advantage, but this cannot be done by ICT by itself [1]. The organization must have: Standardized business processes and people that know and have the ability to apply them [2].

An Enterprise Resource Planning (ERP) is an information system, which integrates most of the data that an organization can process and use in their operations[3]. Its implementation requires: Money, time, a great amount of people effort; and, as an enterprise system, enforce a change in the organizational culture.

ERP systems are increasingly important in today´s business, as they have the ability to support organizational strategies, integrate the flow of information and enhance competitive advantage and individual performance [4]. It has a central database that contains all of the transactions that an organization could register; depending on its set of functional modules. Those functional modules [5] could be, but not restricted to: material management, production, sales, marketing, distribution, financial services, human resources, reports, etc. As its impact affects the whole organization, the ERP system implemented should be the right one [6].

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 204-213

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

The present paper is organized as follows: Section 2 introduces ERP selection process as a Critical Success Factor (CSF), presenting ERP system concept and CSFs. Section 3 presents an ERP literature review, introducing literature ERP implementation methodologies and a propose integrated methodology, its phases and descriptions. Section 4 introduces ERP selection criteria and evaluation categories research literature review and four important criteria categories. Section 5 concludes with discussion regarding ERP selection process, Petri Nets, Multi Agents System and suggestions for future research.

2 ERP Selection Process as a Critical Success Factor

The main goals of an ERP system are to automate business processes [7], to improve interactions and communications inside and outside organization [8] and to eliminate patch work to legacy systems [9]. Its implementation success is affected by CSFs [5, 10-13]. As mentioned by Garcia-Sanchez [12]: "CSF is defined as the limited number of areas in which results, if satisfactory, will ensure successful competitive performance for the organization".

CSF for Umble [14] are: clear understanding of strategic goals, commitment by top management, excellent project management, organizational change management, a great implementation team, data accuracy, extensive education and training, focused performance measures, multi-site issues and ERP selection.

Meanwhile Aloini [10] said that the top ten CSFs are: inadequate ERP selection, ineffective strategic thinking and planning strategic, ineffective project management techniques, bad managerial conduction, inadequate change management, inadequate training and instruction, poor project team skills, inadequate Business Process Reengineering (BPR), low top management involvement, low key user involvement.

Mabert [15] defined thirty CSFs grouped in three categories: planning, implementation decision and implementation, considering schedule and budget as important factors on project goal.

Ranzhe [16] presented: top management involvement, department's participation, funds support, cooperation between enterprise and software company, reasonable expectation with definite target, open and honest communication, training, group structure, project management, enterprise information management, outsider competition pressure, level of the supplier of ERP and service of the supplier of ERP; as CSFs.

Garcia-Sanchez [12] defined that: top management support, BPR, project management, project champion, end users involvement, training and support for users, having external consultants, change management plan, ERP system selection, vision statement and have an adequate business plan to facilitate of changes in the organizational structure in the "legacy systems" and in the IT infrastructure, communication, teamwork composition for the ERP project and problem solutions are the most important CSFs.

Pabedinskaite [17] claim there are internal, external and mixed factors, depending on the aim of responsibility. Most of the CSFs are internal, so the organization is responsible of the success of the project.

Suebsin [18] claims that ERP implementation success is determined by ERP adoption process and this can be addressed by: customary requirement, quality of human resources, inadequate support, change management skills from, clarification of project scope and individual value recognition.

In his paper, Pabedinskaite [17], identifies that selecting the appropriate system is the second most important factor of all CSFs. Aloini [10] identifies inadequate ERP selection as the first factor. Iskanius as the eighth [19]. Garcia-Sanchez as the sixth [12]. Upadhyay as the fourteenth [20]. Ahituv [7], Umble [14], Somers [6], Haghighi [21], Esteves [22], Chiesa [23], Pastor[24], Adam[25], Burqués [26], Muñiz [27], Ayag [28], Kahraman [29], Tomb [30], Stefanou [31], Vilpola [32], Supramaniam [33], Kyung-Kwon [34] have mentioned in their list of CSFs.

Tsai [35] claims that ERP software selection criteria is a factor that influences ERP software quality, information quality and ERP success.

3 ERP Literature Review

ERP selection is an important factor for the success of an ERP implementation. Research literature indicated that: If the business wishes to increase the grade of implementation success, an implementation model should be used [11]. Table 1 and 2 present most representative ERP implementation methodologies in research literature reviewed, each column represent an author's model, rows each of the phases.

Table 1. ERP Implementation methodologies phases

Ahituv [7]	Tomb [30]	Umble [14]	Stefanou [31]
System Selection	Get of requirements	Selection	Business vision
Definition and implementation plan	Develop business case	Implementation	Analysis of business needs against boundaries and change wish
Implementation, process component	Design technological solution		Evaluation and selection
Operation	Implementation		Implementation

Table 2. ERP Implementation methodologies phases (continue)

Esteves[22] Dery[36]	Ayag [37]	Haghighi [21]	Aloini [10]
Adoption	ERP Selection. Requirements analysis	Strategic plan development	Concept: Strategic planning
Acquisition	Application of selection algorithm, defining values, weight attributes and utility	Pre-selection	Concept: Selection
Implementation	Search of candidates and first selection	Selection	Implementation: Deployment
Usage	ERP selection using the best qualified	Post-selection	Implementation: Integration
Evolution	Negotiation with ERP provider	Implementation	Implementation: Stabilization
Retirement	Selection of IT Infrastructure Implementation and post-implementation		Post-implementation: Progress Post-implementation: Evolution

Neves [38] claims that the model used for the implementation is affected by the complexity of the organization and the quantity of functional modules that is trying to operate. Meanwhile, Ahituv [11] and Grenci [7], claim for the use of a System Development Life Cycle (SDLC) model for implementation. Lau [9] indicates that the organization must be prepared to study organization´s needs, get users commitment and assemble an expert project team as a way to improve the success.

Most of the literature specifies a multi-phases methodology as the normal way of implementation process. Some of the authors include business strategic as a provider of requirements and needs to be fulfilled by ERP system. Other saw ERP system as a software development project integrated on business process using software development lifecycle, but most of them understand that an ERP System is an information system that must be aligned with business needs. Our point of view agrees with most of the literature, an ERP system is a business tool and project, not an ICT project.

We also think than the methodology for an ERP system implementation could be a mix between software development lifecycle and business project lifecycle. And inside the methodology the selection process could be a CSF that should be attended with care because it could affect the implementation methodology since solution providers has their own implementation model.

Our own propose is a literature integrated implementation methodology for ERP system. The description of the phases and their sequence is presented at Table 3:

Table 3. ERP Implementation methodology propose, phases and description of phases

Propose phase order	Description
Strategic business analysis	The organization identifies their vision, mission, strategic objectives and business environment to align the new system.
Obtains business requirements list and business readiness	Creating a function list with organization's needs. Evaluating how ready is the organization in terms of the adoption of a new technology, technical, functional, processes, communication, cultural, administrative, resource and commercial change.
First candidates list and filter	Organization does a market search identifying those ERP systems that could fulfill requirements and needs. Preliminary contact with providers. Compile of informations regarding each solution. Discard of systems than did not fulfill obligatory or priority needs. Ending with a list of three to six ERP systems to evaluate.
Identify and agree evaluation method	Determine which method or combination of methods will be used for evaluate ERP packages.
Evaluate of short list candidates	Elaborate demonstration guide line for providers that need to be fulfilled. Integrate evaluation team who's going to attend provider's demonstrations and visiting providers.
Elaborate evaluation matrix	Apply evaluation method and prepare evaluation matrix.
Decision-making	Present evaluation matrix to the evaluation committee, top management and stakeholders. Use defined criteria for analysis and discussion. Main delivery is the decision of whish system the organization going to implement.
Negotiate contract	Defines scope, deliveries, cost, resources and key process indicators (KPI).
Plan ERP implementation	Elaborate the project implementation plan, including data migration and responsibility definition.
Implementation	Executes ERP implementation plan until total deployment of ERP system with training, configuration, etc.
Maintenance	Negotiate maintenance plan with providers, top management and end users to be aware of aspects related to functionality, usability and adequacy to the evolving business processes.

Before the implementation of an ERP, the project lead team needs a clear understanding of business objectives, goals and metrics to be fulfilled. As most of the critical success factors are human related, a culture change plan must be elaborated.

4 ERP Selection Criteria

Implementing an ERP system is not an inexpensive or risk-free venture. An estimated 40-70% of ERP implementations experienced some degree of failure [39]. That is why an organization should select the most appropriated ERP systems for their business needs[10].

Economic research data show that the average mean investment for an ERP implementation could be between fifty thousand dollars to several millions, by ERP implementation attempt [40], in Small and Medium Enterprises (SME). Mexican economical census [41] shows that there are more than 77 thousand economic units in Guadalajara. More than 1% of them could attempt to implement an ERP System. If we calculate the amount spent in the attempts that could fail, an estimated of at least 15 million dollars could be wasted, not including labor cost and business opportunities losses (770 economic units by the minimum dollars spend of 50 thousand dollars per attempt by 40% of fail). In this matter, the importance of a good selection process could be evident.

Literature presented different approaches to selection and evaluation process.

Razmi [42] presented a hybrid multi-criteria model evaluating five categories in two consecutive wave, the first one evaluating with Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) and the second one with PROMETHEE.

Haghighi [21] included a three phase process where a criteria weight table must be filled.

Chiesa [23] also used a weight criteria weight table, but identifying six categories to compare. In both cases, the algorithm for calculation is a weight by evaluation matrix.

Ahituv [7] did not mention an evaluation criteria, just the necessity to submit a feasibility report to the steering committee.

Llal [43] proposed a data envelopment analysis approach using linear programming to measure and compare the relative efficiencies of decision making units. This evaluation approach use two sets of criteria: Meet of business needs and vendor attributes.

Karaarslan [44] proposed an Analytic Hierarchy Process (AHP) defining a hierarchical structure of the problem and forming a pair wise comparison matrix to compute weight of each element and their aggregated weight.

Kahraman [29] proposed a fuzzy heuristic multi-attribute conjunctive approach, using fuzzy heuristics to eliminate the worst alternatives among all at the first stage and then using a fuzzy conjunctive method to select the best among alternatives.

Ayag [28] preferred a Fuzzy Analytic Network Process (ANP), to take quantitative and qualitative elements to evaluate ERP software alternatives. Because the nine-point scale pairwise comparison in the conventional ANP could be insufficient and imprecise for reflecting the right judgments of decision makers fuzzy logic was integrated.

Perera [45] presented a multi-criteria decision model using AHP based in a study from Sri Lankan manufacturing organizations that found seven criteria categories using Expert Choice Software.

Burqués [26] and Pastor [24] introduced SHERPA methodology and a NoFUN notation. This methodology uses weight criteria tables in an incremental process until the best option emerges.

Nikoukaran [46] claims that a hierarchical framework for evaluation software is essential for proper selection.

Reuther [47] claims that the duration of the selection process has an impact on the success of the ERP implementation and the system functionality requirements is the

highest critical selection criteria. Carvallo [48] claims that the enterprise can organize selection criteria into a criteria catalog built for a scope, which can be either a domain or a category of domains arranged in a hierarchical tree-like for the comparison and evaluation. Dimitrova [49] refers than the application of ANP in a framework, together with a Priority Matrix, offers guidelines in the evaluation and selection of alternatives. Jadhav [50] proposed a hybrid knowledge based system integrated with rule based and case based reasoning components. He claims that: "this system assists decision makers, not only in evaluation and selection of the software packages, but also to specify requirements of the desired software package using well defined set of evaluation criteria and criterion values". Bernroider [51] presented preliminary results in terms of ERP systems selection criteria. The paper support the view that ERP decision making is dominated by quality and project related criteria rather than potential benefits on the organizational or individual level. Lv [52] claims than return of investment, functional match, flexibility, complexity, consultants and vendor profile are the most important factor to take care about them to select and appropriate ERP system. Ya-Yueh [53] use six factors to evaluate a useful ERP system. Fuzzy AHP method is used to measure the related weights between different factors based on pair-wise comparisons. Ya-Yueh claim that chief enterprise officer decision, system functionality versus business processes needs and purchasing cost are the more important decision factors.

Different computational methods has been attempted to find the better way to select the appropriate ERP system for different businesses. We think that the best evaluation method should include functional, technical, economical and operational criteria´s, because each of these criteria is an organizational resource concern.

5 Discussion and Conclusions

This survey helps to identify that this topic keeps attention of several research communities. Those communities claim for a general lifecycle model divided on phases. We present our own propose methodology on Table 3, integrated by methodologies of literature. CSFs must be covered before, during and after ERP implementation. An adequate selection process is one of the CSF found the most on research literature. And, as selection is one of the first steps on the ERP life cycle, the whole implementation is based in this important process that could affect the whole implementation process and future business operations. Since main goal of the enterprise top management is the improvement of its operations, if they are unable to use the ERP appropriately after implementation project, because they choose a too small or too inflexible system to their needs, or other cause, they could have poor operations and financial results that could affect the own business existence.

In selection process, different ways to identify and evaluate the selection criteria have been attempted. We think that multivariable criteria used until now are right, if it includes functional, technical, economical and operational criteria´s. But future research could be done to evaluate how the methodologies proposed in literature have improved the ERP implementation success and how the quantification of improvement can be measured.

Also, there are other computational modeling and simulation methodologies that can be researched to get more knowledge regarding this multi-criteria decision problem. As an ongoing research, our opinion is that: Petri nets could be one of those computational modeling and simulation methodologies because they allow to model complex behavior affected by state changes [54], the selection process could be seen as consecutive change of states. Also multi agents could be used to model the selection process because they allow to define behaviors that can be applied by different agents [55]; each agent could be programmed considering stakeholders and their different point of view and evaluation criteria. On both cases, we have to identify clearly, how the evaluation criteria variables interact against each other. For that purpose, we think that an empirical study, focused on that matter, could help us to identify interactions and variables used in Guadalajara, Mexico, by decision makers on the ERP selection process.

Acknowledgments

CONACYT, http://www.conacyt.gob.mx.

References

1. Laudon, K.C., Laudon, J.P.: Management Information Systems: New Approaches to Organization & Technology. Prentice Hall, New Jersey (1998)
2. Davenport, T.H.: The Future of Enterprise System-Enabled Organizations. Information Systems Frontiers 2, 163-180 (2000)
3. Davenport, T.H.: Putting the enterprise into the enterprise system. Harvard Business Review 76, 121 (1998)
4. Bravo, E., Santana, M.: Impacto de la implementación de los sistemas de planeamiento de recursos empresariales ERP en el desempeño individual. In: Americas Conference on Information Systems, pp. 265. Association for Information Systems, (2010)
5. Ali, H.B., Saad, H.B.: Enterprise resource planning: a review and a STOPE view. International Journal of Network Management 15, 363-370 (2005)
6. Somers, T.M., Nelson, K.G.: A taxonomy of players and activities across the ERP project life cycle. Information & Management 41, 257-278 (2004)
7. Ahituv, N., Neumann, S., Zviran, M.: A system of development methodology for ERP systems. Journal of Computer Information Systems 42, 56 (2002)
8. A., M., R., R., E., S.: Challenges in enterprise resource planning implementation: state-of-the-art. Business Process Management Journal 16, 537 - 565 (2010)
9. Lau, L.: Developing a successful implementation plan for ERP: Issues and challenges. In: International Association for Computer Information Systems, vol. 4, pp. 223-229., Las Vegas, Nevada, USA (2003)
10. Aloini, D., Dulmin, R., Mininno, V.: Risk management in ERP project introduction: Review of the literature. Information & Management 547-567 (2007)
11. Grenci, R.T., Hull, B.Z.: New Dog, Old Tricks: ERP and the Systems Development Life Cycle. Journal of Information Systems Education 15, 277-286 (2004)

12. Garcia-Sanchez, N., Perez-Bernal, L.E.: Determination of critical success factors in implementing an ERP system: A field study in Mexican enterprises. Information Technology for Development 13, 293 (2007)
13. Maldonado, M.: El Impacto de los Factores Críticos de Éxito en la Implementación de Sistemas Integrados de ERP. (Spanish). Cuadernos de Difusión 13, 77-118 (2008)
14. Umble, E.J., Haft, R.R., Umble, M.M.: Enterprise resource planning: Implementation procedures and critical success factors. European Journal of Operational Research 146, 241-257 (2003)
15. Mabert, V.A., Soni, A., Venkataramanan, M.: Enterprise resource planning: Managing the implementation process. European Journal of Operational Research 146, 302-314 (2003)
16. Ranzhe, J., Xun, Q.: A Study on Critical Success Factors in ERP Systems Implementation. IEEE, pp. 1-6. (2007)
17. Pabedinskaitė, A.: Factors of Successful Implementation of ERP Systems. Economics & Management 691-697 (2010)
18. Suebsin, C., Gerdsri, N.: Key factors driving the success of technology adoption: Case examples of ERP adoption. In: Portalnd International Center for Management of Engineering and Technology Proceedings, pp. 2638-2643. Portland, Oregon, USA (2009)
19. Iskanius, P.: Risk Management in ERP Project in the Context of SMEs. Engineering Letters 17, 266-273 (2009)
20. Upadhyay, P., Dan, P.K.: ERP in Indian SME's: A Post Implementation Study of the Underlying Critical Success Factors. International Journal of Management Innovation Systems 1, 1-10 (2009)
21. Haghighi, H., Mafi, O.: Towards a Systematic, Cost-Effective Approach for ERP Selection. Proceedings of World Academy of Science: Engineering & Technology 61, 231-237 (2010)
22. Esteves, J., Bohorquez, V.: An updated ERP systems annotaded bibliography: 2001-2005. Communications of AIS 2007, 386-446 (2007)
23. Chiesa, F.: Metodología para selección de sistemas ERP. Reportes técnicos en ingenieria de software 6, 17 (2004)
24. Pastor, J., Estay, C.: Selección de ERP en Pequeñas y Medianas Empresas con un Proyecto de Investigación – Acción. (2000)
25. Adam, F., Sammon, D.: The Enterprise Resource Planning Decade: Lessons Learned and Issues for the Future. Idea Group Publishing, London (2004)
26. Burqués, X., Franch, X., Pastor, J.A.: Formalizing ERP Selection Criteria. (2000)
27. Muñiz, L.: ERP Guía práctica para la selección e implantación. Ediciones Gestión 2000, España (2004)
28. Ayag, Z., Ozdemir, R.G.: An intelligent approach to ERP software selection through fuzzy ANP. International Journal of Production Research 45, 2169-2194 (2007)
29. Kahraman, C., Büyüközkan, G., Ruan, D.: A fuzzy heuristic multi-attribute conjuntive approach for ERP software selection. World Scientific Publishing, Estambul (2004)
30. Tomb, G.: Implementing Enterprise Resource Planning: Lessons Learned from the Front. January, (2006)
31. Stefanou, C.: The Selection Process of Enterprise Resource Planning (ERP) Systems. In: Americas Conference on Information Systemas Proceedings. AIS Electronic Library, 418, (2000)
32. Vilpola, I.H.: A method for improving ERP implementation success by the principles and process of user-centred design. Enterprise Information Systems 2, 47-76 (2008)
33. Supramaniam, M., Kuppusamy, M.: Investigating the Critical Factors in Implementing Enterprise Resource Planning system in Malaysian Business Firms. Proceedings of World Academy of Science: Engineering & Technology 57, 332-341 (2009)

34. Kyung-Kwon, H., Young-Gul, K.: The critial success factors for ERP implemetation: an organizational fit perspective. Information and management 40, 25-40 (2002)
35. Tsai, W.H., Lee, P.L., Shen, Y.S., Yang, C.C.: The relationship between ERP software selection criteria and ERP success. In: Proceedings of International Conference on Industrial Engineering and Engineering Management, pp. 2222-2226. (2009)
36. Dery, K., Grant, D., Harley, B., Wright, C.: Work, organisation and Enterprise Resource Planning systems: an alternative research agenda. New Technology Work and Employment 21, 199-214 (2006)
37. Ayag, Z., Özdemir, R.G.: An intelligent approach to ERP software selection through fuzzy ANP. 45, (2007)
38. Das Neves, D., Fenn, D., Sulcas, P.: Selection of enterprise resource planning (ERP) systems. South African Journal of Business Management 35, 45 (2004)
39. Frano, J.: ERP System Acquisition Project Planning. In: Toolkit, I. (ed.) ERP, vol. 2010, pp. (2008)
40. Jutras, C., Castellina, N.: ERP Plus in process industries Managing Compliance in the Pursuit Profits. (2010)
41. INEGI: Censos Económicos 2004. INEGI, México (2009)
42. Razmi, J., Sangari, M.S.: A hybrid multi-criteria decision making model for ERP system selection. In: International Conference on Information and Automation and Simulation for Sustainability, pp. 489-495. (2008)
43. Lall, V., Teyarachakul, S.: Enterprise Resource Planning (ERP) System Selection: A Data Envelopment Analysis (DEA) Approach. Journal of Computer Information Systems 47, 123-127 (2006)
44. Karaarslan, N., Gundogar, E.: An application for modular capability-based ERP software selection using AHP method. International Journal of Advanced Manufacturing Technology 42, 1025-1033 (2009)
45. Perera, H.S.C., Costa, W.K.R.: Analytic Hierarchy Process for Selection of ERP Software for Manufacturing Companies. Vision (09722629) 12, 1-11 (2008)
46. Nikoukaran, J., Hlupic, V., Paul, R.J.: Criteria for simulation software evaluation. In: Proceedings of winter simulation conference, pp. 399-406. (1998)
47. Reuther, D., Chattopadhyay, G.: Critical factors for enterprise resource planning system selection and implementation projects within small to medium enterprises. International Engineering Management, vol. 1, pp. 851-855 (2004)
48. Carvallo, J.P., Franch, X., Quer, C.: Determining Criteria for Selecting Software Components: Lessons Learned. Software, IEEE 24, 84-94 (2007)
49. Dimitrova, V.A.: Application of the Analytic Network Process (ANP) in a framework of ERP systems implementation success. In: 4th International IEEE conference "Intelligent Systems", pp. 20-19-20-24. (2008)
50. Jadhav, A., Sonar, R.: A Hybrid System for Selection of the Software Packages. IEEE, pp. 337-342. (2009)
51. Bernroider, E.W.N., Sudzina, F., Pucihar, A.: European criteria for assessing enterprise resource planning (ERP) systems: Preliminary results from multiple empirical studies In: BLED Proceedings. pp. 19. Association for Information Systems, (2009)
52. Fan, L., Jinliang, C.: Influencing factors on ERP system selection. IEEE, pp. 671-673. (2010)
53. Ya-Yueh, S.: A Study of ERP Systems Selection via Fuzzy AHP Method. IEEE, pp. 1-4. (2010)
54. Petri, C.A.R., Wolfgang: Petri Net. Scholarpedia T1 - Petri net 3, 6477 (2008)
55. Wooldridge, M.: An introduction to multi agents systems. John Wiley & Sons Ltd (2002)

A Comparison of Responses of Voltage Propagation in Transmission Lines Using Wave Digital Filters Simulation

Jesús García-Guzmán, Farah Helúe Villa-López, Alfredo Ramírez-Ramírez

Facultad de Ingeniería Mecánica Eléctrica, Universidad Veracruzana
Xalapa, México
jesusgarcia@ieee.org

Abstract. Due to the development on programmable digital devices, Digital Signal Processing concepts have generated a lot of interest as an alternative technique to solve systems of partial differential equations. This contribution looks at the problem of voltage propagation in transmission lines represented through a linear hyperbolic one-dimensional system, using a method based in Wave Digital Filters (WDF). In this method, the set of partial differential equations is matched up with a conventional and hypothetical electric circuit that is directly converted into a group of standard networks, which in turn correspond to basic WDF operations. Transition from analogue to digital world is accomplished by applying a bi-linear transformation and the resulting circuit is a discrete model of the system. The method is used here for the simulation and analysis of voltage propagation in different types of transmission lines. A comparison of results obtained by programming the model in Matlab is presented.

Keywords: Wave digital filter, PDE numerical solution, transmission lines simulation.

1 Introduction

Physical systems in the areas of electromagnetism, acoustic, fluids, solid mechanics, heat transfer, among others, can be represented by sets of partial differential equations (PDE). Modelling these systems is of great importance for the understanding of the related phenomena. Numerical methods are often sought because in most cases of practical interest the equations cannot be solved analytically. Some of these methods are based in classical techniques such as finite elements and finite differences, offering acceptable solutions in many situations. In this report we make use of a method originally proposed by Fettweis [1] using multidimensional wave digital filter theory [2, 3] for the integration of PDE systems. Our contribution is supported by the use of an original method described elsewhere [4], which is based on magnetically coupled circuits, and the simulation results reported here are focused to the study of different types of transmission lines under different boundary conditions.

System models in which the functions depend on the time and a space variable are one-dimensional models. If the system involves at least two spatial variables it is considered a multi-dimensional model. The propagation of voltage waves through a

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 215-225

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

transmission line is one of the classical examples of a linear hyperbolic one-dimensional phenomenon.

In communication and power distribution systems in which sets of conductors are used to transport information and electric or electromagnetic energy between two points, transmission lines are very important. They are also widely used in microwaves circuits as resonators, filters, delays equalizers and many more systems; hence the propagation of voltage waves through transmission lines is a problem of great interest in electrical and electronic systems. Power and communication systems are sensible to the effects of voltage distribution along transmission lines and the simulation of their behaviour is very useful for the design and operation of such systems.

Matlab code was written in order to implement the multidimensional (MD) WDF model for the analysis of voltage propagation throughout transmission lines, using representative sets of parameters, exciting the networks with different driving functions, and working under different coupling conditions at the terminations of the line.

The method of multidimensional wave digital filters has been used for the simulation of voltage propagation in transmission lines for several years and, in the current state of the art, it is investigated by putting together recent developments in different areas of research, in particular, merging advances in digital electronics and signal processing techniques. In some cases, digital filters are used in conjunction with finite differences methods, e.g. for the analysis of transmission line coupling [5] and for the simulation of complex systems based on the travelling-wave model [6].

The relevant features of the method are all included in the algorithm used in this report. The benefits of its full parallel processing technique have been positively reported before [7], and the recent trends in the use of ideal transformers networks for the introduction of parallel and series adaptors of WDF [8], support the advantages of the MD-WDF structures implemented. Our algorithm, however, can still be improved by taking advantage of new developments in the subject, and introducing alternative filter structures that would add benefits in terms of efficiency and reliability [9].

2 The PDE Model for a Two-Conductor Transmission Line

Figure 1 shows a simplified representation of a two-conductor transmission line system. The parameters resistance r, inductance l, conductance g and capacitance c, are considered constants throughout this analysis, and they are subject to the following restrictions:

$$r \geq 0, \quad l > 0, \quad g \geq 0, \quad c > 0 . \tag{1}$$

The source end is represented by the series combination of U_S and R_S, whereas the resistive load end is given as R_L. The voltage between the conductors is $u(x,t)$ and the current through the line is $i(x,t)$. Variable x is the spatial coordinate and t corresponds to time.

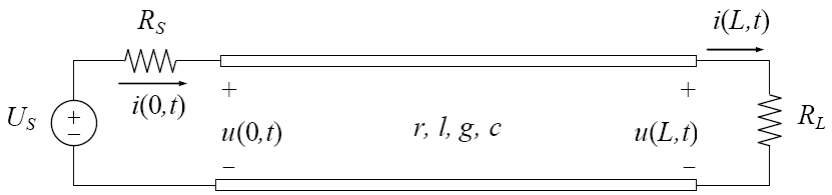


Figure 1. Transmission line with terminations in both ends.

The behaviour of this simplified system is represented by the following set of partial differential equations:

$$\frac{\partial u}{\partial x} + l\frac{\partial i}{\partial t} + ri = 0\,, \tag{2a}$$

$$\frac{\partial i}{\partial x} + c\frac{\partial u}{\partial t} + gu = 0\,. \tag{2b}$$

Given that the length of the line is L, the values of the variables are restricted to $x \in [0, L]$ and $t \geq 0$. From the elements connected to the ends of the line, the boundary conditions are given as follows:

$$x = 0: \qquad u(0,t) = U_S(t) - R_S\, i(0,t)\,, \tag{3a}$$

$$x = L: \qquad u(L,t) = R_L\, i(L,t)\,. \tag{3b}$$

The PDE system defined in (2) together with the set of conditions (3) define the initial version of the model to be solved, for a two-conductor transmission line with a source termination in one end and a pure resistive load in the other end. This model needs to be transformed and arranged through several steps before an algorithm based on WDF can be developed.

3 From the PDE Set to the Equivalent Kirchhoff Network

In order to obtain a network equivalent of the PDE set, complex quantities are introduced and a multidimensional Laplace transform needs to be applied to the system. Using the complex wave number s_x and the complex frequency s_t, defined within a complex vector $\mathbf{s} = (s_x, s_t)^{\mathrm{T}}$ and a vector of variables $\mathbf{t} = (x,t)^{\mathrm{T}}$, a 2-D Laplace transform can be performed. Expressing the functions voltage and current in exponential form:

$$u(x,t) = r_2 I_2 e^{\mathbf{s}^{\mathrm{T}}\mathbf{t}}, \qquad i(x,t) = I_1 e^{\mathbf{s}^{\mathrm{T}}\mathbf{t}}\,. \tag{4}$$

The complex constants I_1 and I_2 are interpreted as complex amplitudes and r_2 is an arbitrary constant which allows the substitution of the variable voltage with the second current unknown I_2.

Applying the corresponding 2-D Laplace transform, and rearranging terms, (2a) and (2b) are transformed into the following algebraic equations:

$$(s_t l + r) I_1 + s_x r_2 I_2 = 0\,, \tag{5}$$

$$s_x r_2 I_1 + (s_t c + g) r_2^2 I_2 = 0\,. \tag{6}$$

A hypothetical network which corresponds to (5) and (6) can be obtained when these equations are interpreted as the result of the application of Kirchhoff's voltage law around a mesh. After applying the method described in [4] and choosing $r_2 = \sqrt{l/c}$, a symmetrical T-circuit is obtained and this is transformed into the symmetric lattice shown in Figure 2, in which all coefficients are positive. This lattice is used as a reference Kirchhoff network for the transformation to the discrete domain, in order to obtain a digital filter circuit.

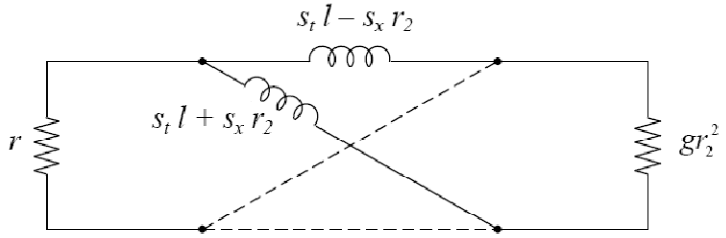


Figure 2. Lattice equivalent to the T-circuit.

4 Designing the Discrete WDF Structure

Inductive terms of the symmetric lattice are used for the design of the corresponding structure in the discrete domain. A WDF model is obtained by applying the following bi-linear transformation to the circuit:

$$s_t l - s_x r_2 \to \frac{2r_2}{h}\psi'\,, \quad s_t l + s_x r_2 \to \frac{2r_2}{h}\psi''\,. \tag{7}$$

with

$$\psi' = \tanh\frac{1}{2}(s_t T - s_x h), \quad \psi'' = \tanh\frac{1}{2}(s_t T + s_x h)\,. \tag{8}$$

The selected value for r_2 produced the symmetry obtained for both the lattice and the corresponding bi-linear transformation, making that the integration path in the **t** domain follows the direction of wave propagation, thus achieving all major advantages of the method. The spatial step size h is related to the time step size T by

$$l\, h = r_2 T\,. \tag{9}$$

Using $r_2 = \sqrt{l/c}$, the step size and the port resistances of the lattice are:

$$T = \sqrt{lc}\, h\,, \tag{10}$$

$$r_0 = \frac{2r_2}{h}\,. \tag{11}$$

In the solution of the filter circuit, voltages and currents are replaced by *voltage wave quantities* for every port with resistance *R*. Introducing wave quantities and using the lattice equivalent connected to the resistive elements through two-port adaptors, the final structure shown in Figure 3 is obtained. Due to the absence of sources, the reflected voltage waves a_1 and a_2 are equal to zero. The port resistances R_1 and R_2 are equal to the actual resistances connected to the ends of the lattices:

$$R_1 = r\,, \qquad R_2 = gr_2^2\,. \tag{12}$$

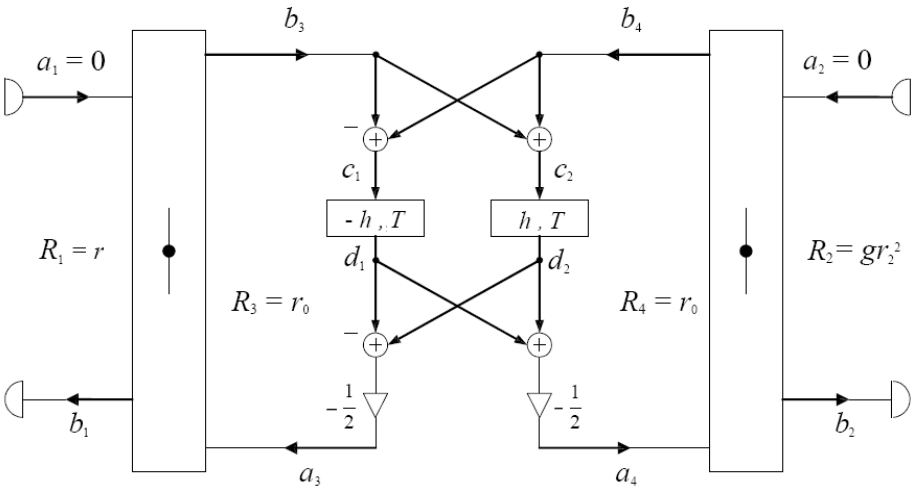


Figure 3. Structure of the MD-WDF for the simulation of the transmission line.

From the two-port series adaptors, we obtain the following equations:

$$\gamma_1 = \frac{2r}{r + r_0}\,, \qquad \gamma_2 = \frac{2gr_2^2}{gr_2^2 + r_0}\,, \tag{13a}$$

$$\gamma_3 = \frac{2r_0}{r + r_0} = 2 - \gamma_1\,, \qquad \gamma_4 = \frac{2r_0}{gr_2^2 + r_0} = 2 - \gamma_2\,. \tag{13b}$$

where the number of multipliers can be simplified as required. The reflected waves are obtained knowing that:

$$b_1 = -\gamma_1 a_3\,, \qquad b_2 = -\gamma_2 a_4\,, \tag{14a}$$

$$b_3 = a_3(1-\gamma_3)\,, \qquad b_4 = a_4(1-\gamma_4)\,. \tag{14b}$$

The solution of the circuit is obtained following the method described in [4] and [10], yielding the following equivalent expressions, all of them valid for the voltage propagation through the transmission line:

$$u = \frac{r_2}{2r_0}(a_4 - b_4) = \frac{h}{4}(a_4 - b_4) = \frac{h}{4}\gamma_4 a_4\,. \tag{15}$$

Since lattice impedances in the reference network are purely inductive and they contain both space and time derivatives, an explicit discrete algorithm can be implemented, given that every spatial shift h occurs associated to a delay by T. Applying the bi-linear transformation and using m and k to denote discrete steps in space and time respectively, the difference equations for the shift operators are derived:

$$d_1[m,k] = c_1[m+1,k-1]\,, \tag{16a}$$

$$d_2[m,k] = c_2[m-1,k-1]\,. \tag{16b}$$

The inputs to the shift operators are obtained from the inputs to the lattice:

$$c_1 = b_4 - b_3\,, \qquad c_2 = b_4 + b_3\,. \tag{17}$$

Reciprocally, the outputs of the lattice are obtained from the outputs of the shifters:

$$a_3 = \tfrac{1}{2}(d_1 - d_2)\,, \quad a_4 = -\tfrac{1}{2}(d_1 + d_2)\,. \tag{18}$$

5 Boundary and Initial Conditions

Expressions in (16) are not applicable at every spatial boundary, since the structure for $x = 0$ and $x = L$ would require values that are outside the transmission line. Values for d_1 and d_2 on the boundaries need to be calculated in a special way.

For $x = 0$, $c_2[m-1, k-1]$ is not available and d_2 cannot be obtained by using (16b). However, d_1 is still obtainable and an alternative expression for d_2 can be derived by using the given boundary conditions. Replacing a_4 from (18) in (15), the voltage u is expressed in terms of d_1 and d_2:

$$u = -\frac{r_2}{4r_0}\gamma_4(d_1 + d_2)\,. \tag{19}$$

Substituting (19) in the boundary condition (3a), an expression valid for $x = 0$ is obtained and hence the value of d_2 can be expressed as:

$$d_2 = \frac{(R_S\gamma_3 - r_2\gamma_4)d_1 - 4r_0U_S(t)}{R_S\gamma_3 + r_2\gamma_4}\,. \tag{20}$$

Also, using (3b), an expression to compute d_1 for $x = L$ is obtained:

$$d_1 = \frac{(R_L\gamma_3 - r_2\gamma_4)d_2}{R_L\gamma_3 + r_2\gamma_4}. \quad (\mathbf{21})$$

Initial conditions also incorporated to the algorithm. Assuming there is no initial energy in the system, the functions voltage and current are zero for $t = 0$. So, $a_3 = 0$ and $a_4 = 0$. Then (14) also implies that b_3 and b_4 are equal to zero in the initial step. By (17) c_1 and c_2 are zero at $t = 0$ and hence the shifters start the procedure from zero.

6 Experimental results

A Matlab algorithm was developed following the MD-WDF model described by the foregoing equations. The algorithm was initially validated through the simulation of voltage propagation in a coaxial cable quoted by Rabenstein and Krauss [11] after an example solved originally using a finite difference method. The length of the cable is L = 10 km and the parameters of the line are r = 46.8 mΩ/m, l = 5.6 μH/m, g = 0 and c = 120 pF/m. The voltage source is given as

$$U_S(t) = \sin^2\left(\frac{\pi t}{\tau}\right)\cdot[\delta_{-1}(t) - \delta_{-1}(t-\tau)], \quad (\mathbf{22})$$

where δ_{-1} is used to represent the unit step function. The value of τ is equal to $24T$, which is the duration of the voltage pulse in the source. The number of spatial sampling points was equal to 32. The source resistance is $R_S = \sqrt{l/c} = 216\Omega$, for a good coupling between the line and the source, and the load resistance is $R_L = 10^4 \cdot R_S$ = 2.16 MΩ, giving an approximately open circuit in the load end. The simulation produced the same results as reported in [11]. The propagation of voltage throughout the line is shown in Figure 4(a), where the potential distribution reflects the coupling conditions at the terminations of the line: there is low reflection in the source side and an open termination at the load end. The boundaries voltages for this line are plotted in Figure 4(b), along with the shape of the output for $x = L/2$.

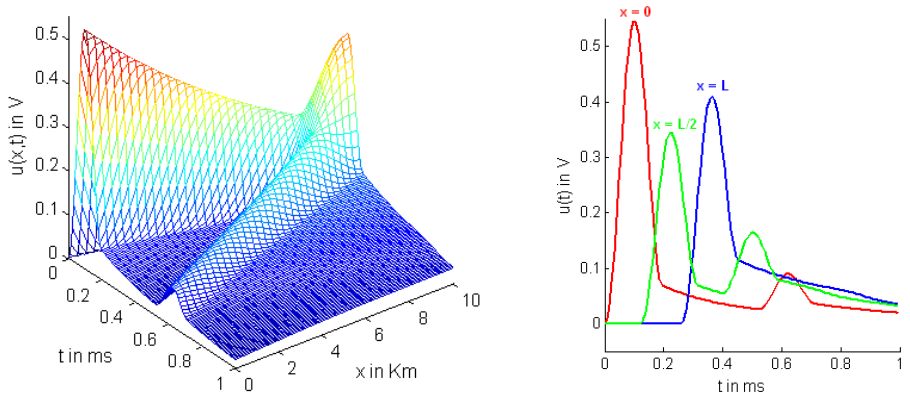


Figure 4. (a) Voltage simulation for a coaxial cable transmission line with open circuit at $x = L$; (b) boundary voltages in a coaxial cable and simulation results for $x = L/2$.

A second simulation was performed for the same line with a short circuit at $x = L$ and the results obtained were similar to those reported in literature [12, 13]. Figure 5 presents the output of the simulation, revealing multiple reflections, distortion and loss that are characteristic of wave propagation in a transmission line.

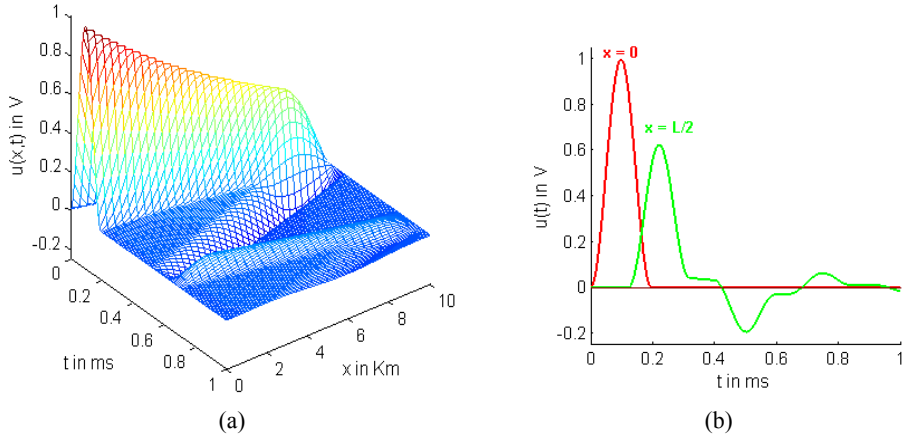


Figure 5. (a) Voltage simulation for a coaxial cable transmission line with short circuit at $x = L$; (b) boundary voltage in a coaxial cable at the source end and simulation results for $x = L/2$.

Once the results confirmed that the WDF model was producing correct and reliable results, the algorithm was used for simulations under different conditions. The voltage source function was changed to an alternating waveform with $\tau = 12T$:

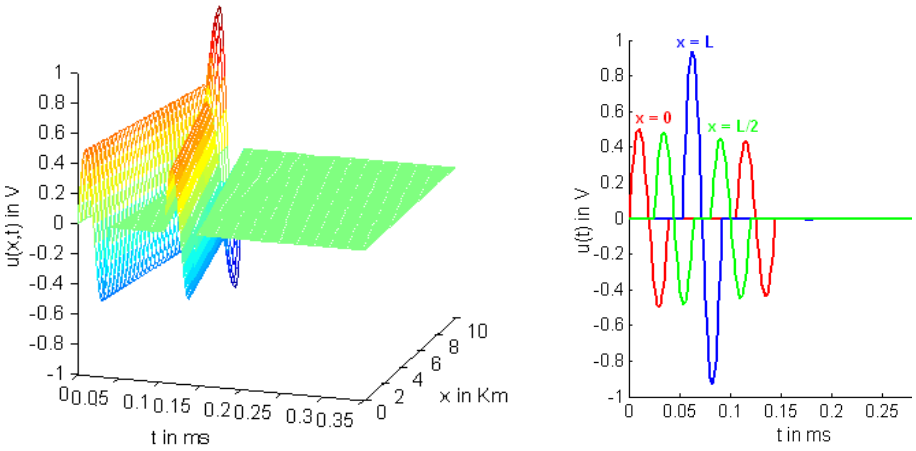
$$U_S(t) = \sin\left(\frac{\pi t}{\tau}\right) \cdot [\delta_{-1}(t) - \delta_{-1}(t-\tau)]. \qquad (23)$$

Data reported by Neri Vela [14] for a bifilar copper line served for the implementation of the new simulation. The parameters of this line are given as r = 2.74 mΩ/m, l = 1.02 μH/m, g = 34.35 pS/m and c = 27.33 pF/m. The propagation of voltage along this bifilar copper line with an open circuit at the load end is shown in Figure 6(a). It can be noticed the change in the response at the load end and also the change in amplitude of the reflected wave, with respect to the previous analysis for the coaxial cable. The voltage waveforms at the boundaries are plotted in Figure 6(b), together with the shape of the output for $x = L/2$.

Both the method and the implemented algorithm were also applied to the simulation of a third example: the propagation of voltage through a telephonic bifilar line with the load end in short circuit, whose parameters were taken from [15]. The voltage source is a unit step function assumed to be:

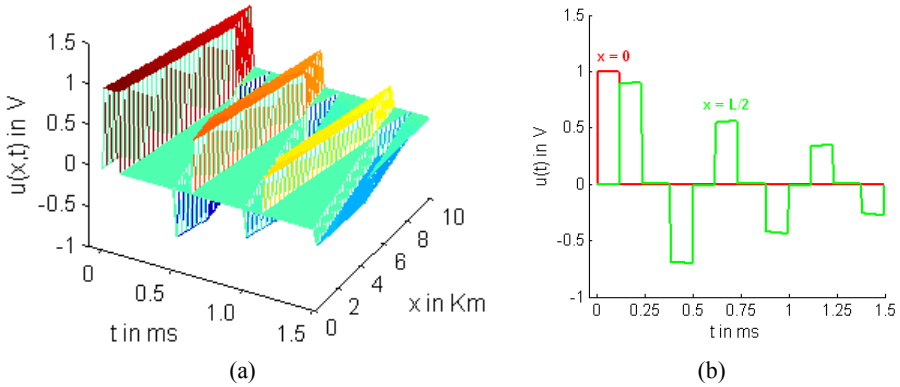
$$U_S(t) = \delta_{-1}(t) = \begin{cases} 1V, & 0 \le t \le 16T \\ 0V, & otherwise \end{cases} \quad \textbf{(24)}$$

The line is 10 km long and the parameters are r = 4.185 mΩ/m, l = 2.17 μH/m, g = 284 pS/m and c =5.4 pF/m. The results of the simulation are shown in Figure 7(a) for the propagation of the square pulse**.** Voltage waveforms at $x = 0$ and $x = L/2$ are presented in Figure 7(b).



(a) (b)

Figure 6. (a) Voltage simulation for the bifilar copper line with open circuit at $x = L$; (b) boundary values for u and simulation results at $x = L/2$ for the bifilar copper line.



(a) (b)

Figure 7. (a) Propagation of a square pulse in a telephone line; (b) voltage waveforms at $x = 0$ and $x = L/2$ for the telephonic line.

7 Conclusions

The method of multidimensional wave digital filters described in this paper is a powerful tool for solving PDEs and hence modelling physical systems. Throughout this report, we have tested an algorithm based in this method for the simulation of voltage propagation in transmission lines.

MD-WDF theory allows the modelling of physical behaviour for different types of transmission lines, as we have shown by applying our algorithm to several examples under different voltage sources, using different materials and under different boundary conditions.

Passivity of the modelled systems is achieved in the method thanks to the use of an inherent passive reference network. By using passive MD-WDFs, it can be ensured that the behaviour of the algorithm for numerical integration is as close as possible to the case of exact computations.

These inherent properties also lead to stability, allowing arbitrary variations in the parameters of the system, as it can be noticed from the examples presented. Robustness of the discrete model obtained, makes it very suitable not only for the simulation of voltage in transmission lines, but also for the numerical analysis of wave propagation and other linear problems in several fields of engineering [16, 17].

Hyperbolic problems, i.e., time dependent problems related to propagation of waves over finite distances in non-zero time, are the best suitable for the application of this method.

References

1. Fettweis, A.: Discrete passive modelling of physical systems described by PDEs. In: 6th European Signal Processing Conference, I, pp. 55—62, In: Signal processing VI, theories and applications, edited by J. Vandewalle, R. Boite, M. Moonen, A. Oosterling, Elsevier, Amsterdam (1992)

2. Fettweis, A.: Robust numerical integration using wave-digital concepts. Multidimensional Systems and Signal Processing, 17, 7—25 (2006)
3. Krauss, H., Rabenstein, R., Gerken, M.: Simulation of wave propagation by multidimensional digital filters. Simulation, practice and theory 4, 361—382 (1996)
4. García-Guzmán, J.: The modelling of physical systems using wave digital filters. MSc Thesis. The University of Warwick, Coventry, UK. (1999)
5. Ying-Hui Zhou, Li-Hua Shi, Cheng Gao, Bi-Hua Zhou, Bin Chen.: Combination of FDTD Method With Digital Filter in Analyzing the Field-To-Transmission Line Coupling. IEEE Transactions on Electromagnetic Compatibility 50-4, 1003—1007 (2008)
6. Li, W., Huang, W. P., Li, X.: Digital filter approach for Simulation of a complex integrated laser diode based on the traveling-wave model. IEEE Journal of Quantum Electronics. 40-5, 473—480 (2004)
7. Chien Hsun Tseng, Lawson, S.: Full parallel process for multidimensional wave digital filtering via multidimensional retiming technique. In: IEEE International Symposium on Circuits and Systems 3, pp. 209—212. Vancouver, Canada (2004)
8. Lorenz, P. Russer, P.: Discrete and Modal Source Modeling with Connection Networks for the Transmission Line Matrix (TLM) Method. In: IEEE/MTT-S International Microwave Symposium, pp. 1975—1978. Honolulu, Hawaii (2007)
9. Wanhammar, L., Soltanian, B., Gustafsson, O., Johansson, K.: Synthesis of bandpass circulator-tree wave digital filters. In: 15th IEEE International Conference on Electronics, Circuits and Systems, pp.834-837. Malta (2008)
10. Lawson, S.S., Guzmán, J.G.: On the modelling of the 2D wave equation using multidimensional wave digital filters. In: IEEE International Symposium on Circuits and Systems 2, pp. 377—380. Sydney, Australia (2001)
11. Krauss, H. and Rabenstein, R.: Application of multidimensional wave digital filters to boundary value problems. IEEE Signal Processing Letters 2-7, 138—140 (1995)
12. Rabenstein, R. and Krauss, H.: Discrete Simulation of Uniform Transmission Lines by Multidimensional Wave Digital Filters. International Journal of Numerical Modelling: Electronic Networks, Devices and Fields 9-3, 271—294 (1996)
13. Krauss, H.: Wave digital simulation of transmission lines with arbitrary initial potential and current distribution. In: IEEE Digital Signal Processing Workshop, pp. 195—198. Loen, Norway (1996)
14. Neri Vela, R.: Líneas de transmisión, 1st ed. McGraw-Hill/Interamericana Editores, México (1999)
15. Sosa-Pedroza, J., Ortega, L.: Líneas de transmisión y guías de onda, 1st ed. Limusa, México (1988)
16. García-Guzmán, J., González-López, E., Campillo, R.: Análisis de modos normales de vibración de una membrana mediante filtros digitales de ondas. In: CIINDET 6° Congreso Internacional sobre Innovación y Desarrollo Tecnológico. Cuernavaca, México (2008)
17. Välimäki, V., Pakarinen, J., Erkut, C., Karjalainen, M.: Discrete-time modelling of musical instruments. Reports on Progress in Physics. 69-1, 1—78 (2006)

On the stability of a RLC parametric oscillator

Luis Moreno-Ahedo*, Marisol Prado*, Roberto Carmona** and Armando Aguayo*

Unidad de Estudios de Posgrado e Investigación*
Tecnológico de Estudios Superiores de Coacalco
lmoreno@ieee.org, marisolprado@ieee.org, webmaster@tesco.edu.mx
Universidad Politécnica del Valle de México
rcarmona@ctrl.cinvestav.mx

Abstract. We analyze the stability of a RLC parametric oscillator when the frequency and amplitude (γ, Ω) of the excitation source varies, the oscillator is modeled as a parametrically excited system then Floquet theory approach is used to study the stability by stating the monodromy matrix which is approximated symbolically, the analysis show in the (γ, Ω)-plane the dominoes of stability and instability, a propose algorithm computes the transition curves that is the boundaries between the dominoes of stability, the analysis is supported by numerical simulations.

Keywords: Parametric oscillator, Arnold's tongues, monodromy matrix, parametrically excited systems, parametric resonance

1 Introduction

When the model of system the applied excitation source appears as a time-varying coefficient the system is called parametrically excited (PE) system [1], examples of such systems are: the variable length pendulum and the pendulum with a vertically oscillating pivot, torsional oscillators, cantilevers and translational oscillators to name a few ones.

Quite time ago Melde [2] and Faraday [3] recognized experimentally the parametric resonance phenomenon occurred in PE systems, through the work developed by [4], [5], [6], [7] and [8] now it is well-know that PE systems experience parametric resonance when driven at frequencies close to twice ω_0/n, where $n \leq 1$ and ω is the natural frequency, in many fields of engineering this phenomenon can occur, for example in bridges [9] and cause degradation and failure of mechanical structures, however some current applications exploiting parametric resonance for example mass sensing [10] and signal filtering [11].

In the study of PE systems there are well established results through three basic approaches: the perturbation theory [12], the harmony balance [13] and the Floquet theory [14]. However the study of PE system with varying frequency has received little attention namely, in [15] the resonance zones are determinate by approximation procedures and in [16] the harmony balance is used to analyse a PE system with combined frequencies.

The problem of studying a PE system with varying frequency can be seen as the frequency analysis of linear time periodic systems where the main results are [17] which

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 226-234
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

used an one-to-one map induced by geometrically periodic signals and [18] where a general harmony balance method is used to stablish the frequency response. We shall apply the algorithm developed in [19] to analysis the stability of a parametric oscillator when the frequency of the excitation source varies, the oscillator is implemented as the parametric forced RLC circuit shown in figure 1.

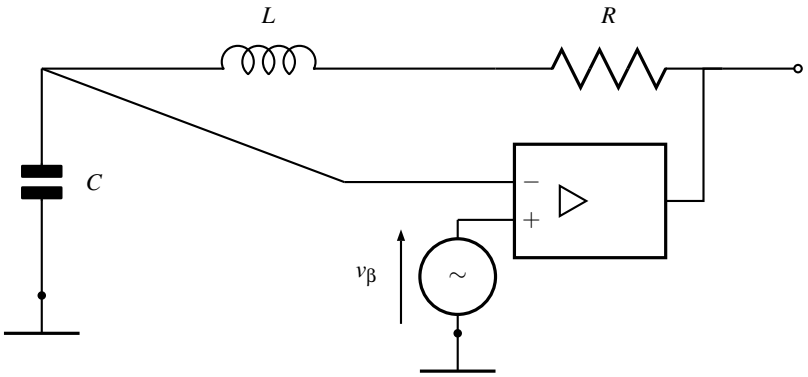


Fig. 1. Circuit of a RLC parametric oscillator, the box represent an analog multiplier

2 RLC parametric oscillator model

In this section the RLC parametric oscillator [20] is modeled. Applying the Kirschoff's voltage law to the electronic circuit shown in the figure 1 yields

$$v_C + v_R + v_L = k v_C v_\beta$$

where the voltages are given by $v_R = R\frac{dq}{dt}$, $v_C = \frac{q}{C}$, $v_L = L\frac{d^2q}{dt^2}$ and $v_\beta = \beta\cos\omega t$, substituting and ordering terms

$$\ddot{q} + \frac{R}{L}\dot{q} + \frac{1}{LC}(1 - k\beta\cos\omega t)q = 0$$

where q is charge, C capacitance, L inductance, R resistance, k the gain of the analog multiplier and the pair (γ, ω) are the amplitude and frequency of the excitation source v_β, without loss of generality can be written

$$\ddot{q} + \lambda\dot{q} + \omega_0^2(1 + k\beta\cos\omega t)q = 0$$

where $\omega_0^2 = \frac{1}{LC}$ and $\lambda = \frac{R}{L}$, the above equation using the dimensionless variable $\tau = \omega_0 t$ can be reduce to:

$$\frac{d^2q}{d\tau^2} + \mu\frac{dq}{d\tau} + (1 + \gamma\cos\Omega\tau)q = 0 \tag{1}$$

where $\Omega = \frac{\omega}{\omega_0}$, $\gamma = k\beta$ and $\mu = \frac{R/L}{\omega_0}$.

The change of variable, [21]:

$$q = xe^{\frac{1}{2}\mu}$$

transforms the equation (1) into:

$$\ddot{x} + \left(1 - \frac{1}{4}\mu^2 + \gamma\cos\Omega\tau\right)x = 0 \tag{2}$$

For a given dissipative term μ, we shall plot the stability chart in the (γ, Ω)-plane of Eq. (2).

The above equation corresponds to so-called Mathieu equation, [22]:

$$\ddot{x} + [\alpha + \beta\cos(t)]x = 0$$

which is the a special case of the Hill's equation, [22]:

$$\ddot{x} + [\alpha + \beta p(t)]x = 0$$

where $p(t+T) = p(t)$ and $\int_0^T p(t) = 0$

For both equations the stability is usually given in terms of the parameters as a stability chart in the (α, β)-plane as the figure 2 shows, the shaded regions are zones where the parametric resonance occurs such zones are known as resonance zones or Arnold's tongues [23], while the white regions are zones where the solution is bounded, the boundaries are called transition curves where the solution is periodic. To analyse this kind of equations it is used the Floquet theory given in following section.

3 Preliminaries

Consider the linear time periodic (LTP) system:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A(t)}\boldsymbol{x}(t) \tag{3}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. The Floquet theory states, [14]:

Theorem 1. (Floquet's theorem) *Any* $\boldsymbol{\Phi}$ *fundamental matrix of* (3) *can be written as:*

$$\boldsymbol{\Phi}(t) = \boldsymbol{P}(t)e^{\boldsymbol{R}t}$$

where $\boldsymbol{R}, \boldsymbol{P} \in \mathbb{R}^{n\times n}$ *is non-singular,* $\boldsymbol{P}(t) = \boldsymbol{P}(T+t)$ *and* T *is the minimal period* □

Floquet theory also shows that $\boldsymbol{\Phi}(t+T) = \boldsymbol{\Phi}(t)e^{\boldsymbol{R}T}$, if $t = 0$ then:

$$\boldsymbol{C} \triangleq e^{\boldsymbol{R}T} = \boldsymbol{\Phi}^{-1}(0)\boldsymbol{\Phi}(T)$$

The matrix $\boldsymbol{C}$ is the so-called monodromy matrix [13], its eigenvalues λ_i are called characteristic multipliers. The importance of the matrix is shown in the following corollary.

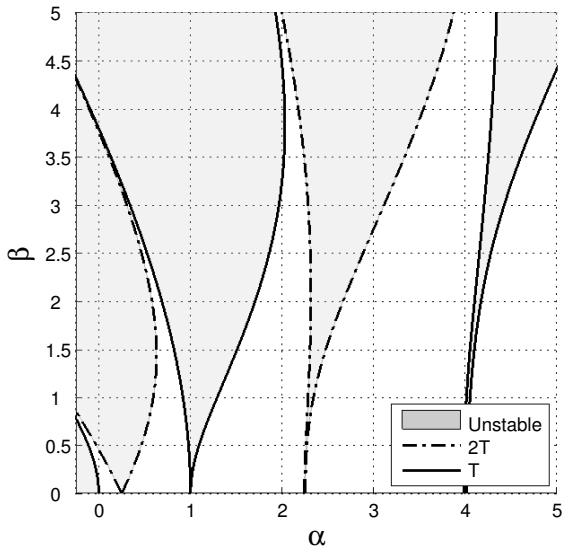


Fig. 2. Stability chart of the Mathieu eq. $\ddot{x}+(\alpha+\beta\cos t)x=0$

Corollary 1. *The solution* $\boldsymbol{x}(t)$ *of the system* (3) *satisfies*

$$\boldsymbol{x}(t+T)=\lambda\boldsymbol{x}(t)$$

if and only if λ *is a characteristic multipliers of* (3). □

From the above corollary we deduce the criterion shown in table 1.

The above result is elegant but computing the characteristic multiplier in practical cases is almost impossible however the Floquet theorem applied to the case of Hill's equation provides an attractive result as follows.

Solutions	Multipliers
Stability trivial solution	Inside or on the unit circle (simple multipliers)
Asymptotic	Inside unit circle
Instability of of trivial solution	At least one outside the unit circle or on the unit circle with multiple elementary divisor
Periodic solution	At least one multiplier equal to 1 or -1

Table 1. Multipiers

The Hill's equation written in state variable yields:

$$\dot{\boldsymbol{x}} = \begin{bmatrix} 0 & 1 \\ -\alpha - \beta p(t) & 0 \end{bmatrix} \boldsymbol{x} = \boldsymbol{A}(t)\boldsymbol{x} \tag{4}$$

Let be $\boldsymbol{\Phi}$ a fundamental matrix of the system (4) and initial conditions such that $\boldsymbol{\Phi}(0) = \boldsymbol{I}$. Then the monodromy matrix of (4) is $\boldsymbol{C} = \boldsymbol{\Phi}(T)$. The characteristic multipliers are given by $p(\lambda) = \lambda^2 - \mathrm{tr}[\boldsymbol{C}]\lambda + \det[\boldsymbol{C}]$ applying the Liouville's formula[1] yields $\det[\boldsymbol{C}] = \det[\boldsymbol{\Phi}(T)] = 1$ then $\lambda_i = \frac{1}{2}\left[\phi \pm \sqrt{\phi^2 - 4}\right]$ where $\phi = \mathrm{tr}[\boldsymbol{C}] = \mathrm{tr}[\boldsymbol{\Phi}(T)]$ is the trace of the monodromy matrix applying the results of table 1 we deduce the following criterion.

Criterion 1 *Let be* $tr[\boldsymbol{C}] = tr[\boldsymbol{\Phi}(T)]$ *the trace of the monodromy matrix and* $x(t)$ *the solution of* (4) *then, [25]:*

(i) If $|\mathrm{tr}[\boldsymbol{\Phi}(T)]| < 2$ *then* $x(t)$ *is bounded.*
(ii) If $|\mathrm{tr}[\boldsymbol{\Phi}(T)]| > 2$ *then* $x(t)$ *is unbounded.*
(iii) If $|\mathrm{tr}[\boldsymbol{\Phi}(T)]| = 2$ *then* $x(t)$ *is periodic* □

This criterion enable us to evaluate only the trace of the monodromy matrix $\mathrm{tr}[\boldsymbol{\Phi}(T)]$ instead of calculating the characteristic multipliers, which is practical because the trace $\mathrm{tr}[\boldsymbol{\Phi}(T)]$ can be approximated using the algorithm developed in [19] which approximates symbolically the monodromy matrix using the Taylor's method for ordinary differential equations [26], the algorithm is implemented in a computer algebra program, namely Mathematica©.

4 Computing transition curves

A pseudo-code of the algorithm to approximate the monodromy matrix is given in Algorithm 4.1.

The inputs of the algorithm are: the periodic matrix $A(t)$, the minimal period T , M is order Taylor's method, n is the number of divisions of the time interval and m is the matrix A dimension.

The first loop (i) in the algorithm computes the general step of the method, the second (ii) loop makes a copy of the general step to compute the vectors column solutions $\boldsymbol{x}_i$ of the monodromy matrix and approximates such solutions as a sequence of points, the line (iii) takes the last element of the sequences to form the monodromy matrix. In the algorithm $\boldsymbol{e}_i$ are the standard basis of the vector space $\mathbb{R}^m$.

[1] [24] Let bet $\boldsymbol{\Phi}(t)$ a fundamental matrix of $\dot{\boldsymbol{x}} = \boldsymbol{A}(t)\boldsymbol{x}$ then $\det|\boldsymbol{\Phi}(t)| = \det|\boldsymbol{\Phi}(0)|\, e^{\int_{t_o}^{t} \Sigma_{j=1}^{n} a_{jj}(s)ds}$

Algorithm 4.1: MONODROMYMATRIX($\boldsymbol{A}(t), T, M, n, m$)

$\boldsymbol{F} := \boldsymbol{A}(t)\boldsymbol{x}_j^i[t]; h := \frac{T}{n}; \boldsymbol{S} := \boldsymbol{x}_j^i[t]$
for $l := 1$ **to** M
do (i)
$\begin{cases} \boldsymbol{F} := \frac{d\boldsymbol{F}}{dt} \\ \boldsymbol{S} := \boldsymbol{S} + \frac{h^l}{l!}\boldsymbol{F} \end{cases}$ t:=0;
for $i := 1$ **to** m
do (ii)
$\begin{cases} \boldsymbol{S}_i := \boldsymbol{S} \\ \boldsymbol{x}_j^i[0] = \boldsymbol{e}_i \\ \textbf{for } j := 1 \textbf{ to } n \\ \quad \textbf{do} \\ \quad \left\{ \boldsymbol{x}_j^i[t+h] := \boldsymbol{S}_i[t] \right|_t \\ t := t + h \end{cases}$
$\boldsymbol{C} := \left[\boldsymbol{x}_n^1 \, \boldsymbol{x}_n^2 \cdots \boldsymbol{x}_n^m \right]$ (iii)

Writing the equation of the electronic parametric oscillator (2) in state variable yields:

$$\dot{\boldsymbol{x}} = \begin{bmatrix} 0 & 1 \\ -\left(1 - \frac{1}{4}\mu^2 + \gamma\cos\Omega\tau\right) & 0 \end{bmatrix} \boldsymbol{x} = \boldsymbol{A}(t)\boldsymbol{x}$$

Suppose that the dissipative term is $\mu = 1$ then we shall approximate the monodromy matrix $\boldsymbol{C}$ with the algorithm 4.1 for each frequency Ω between $(0.25, 1.5)$ with a step $h = 0.01$ considering γ as a parameter, for example considering $\Omega = 1$ the algorithm approximates the trace of the monodromy matrix $\phi = \mathrm{tr}[\boldsymbol{C}]$ as a polynomial

$$\begin{aligned} &-1.95 - (1.61 \times 10^{-15})\gamma + 1.08\gamma^2 - (3.5 \times 10^{-15})\gamma^3 + 6.44\gamma^4 \\ &+4.77\gamma^6 + (1.12 \times 10^{-14})\gamma^7 + 2.14\gamma^8 + (1.11 \times 10^{-15})\gamma^9 + 1.99\gamma^{10} \\ &\vdots \\ &-(1.44 \times 10^{-19})\gamma^{21} \cdots - (1.86 \times 10^{-84})\gamma^{67} + (6.36 \times 10^{-73})\gamma^{68} \\ &-(3.73 \times 10^{-80})\gamma^{70} - (4. \times 10^{-96})\gamma^{71} + 0.\gamma^{72} \end{aligned}$$

The γ roots of the polynomial $|\phi| = 2$ are the values for which the solution of the equation (2) is periodic according with the criterion 1.

In this way computing for each $\Omega \in (0.25, 1.5)$ the γ values for which the solution of (2) is periodic we shall get the transition curves in the (γ, Ω)-plane, this curves are the boundaries between stable an unstable dominoes, the result of this procedure is show in the figure 3 where the mark $\boldsymbol{U}$ is for unstable dominoes and the mark $\boldsymbol{S}$ is for stable dominoes.

In the stability chart of figure 3 the unstable dominoes are regions where the parametric resonance phenomenon occurs that is where the solution of the equation (2)

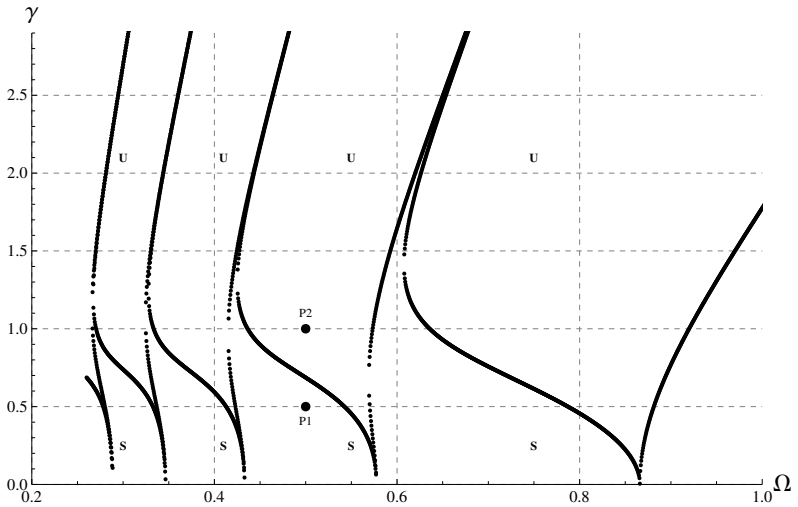
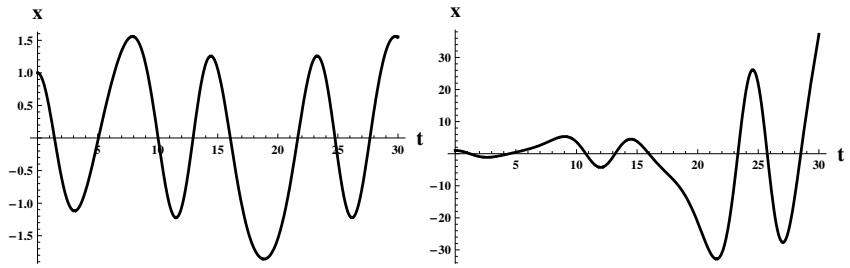


Fig. 3. Stabillity chart in the plane (γ,Ω)-plane of the parametric oscillator (2)



(a) $P_1 = (\gamma,\Omega) = (0.5, 0.5)$ (b) $P_2 = (\gamma,\Omega) = (1, 0.5)$

Fig. 4. Solution of the equation $\ddot{x} + \left(1 - \frac{1}{4}\mu^2 + \gamma\cos\Omega\tau\right)x = 0$ for $\mu = 1$

grows exponentially, for example consider the solution of the eq. (2) for the cases shown in chart 3, namely $P_1 = (\gamma,\Omega) = (0.5, 0.5)$ and $P_2 = (\gamma,\Omega) = (1, 0.5)$, the solutions are plotted in the figure 4 as can been seen the solution for P_1 is bounded while for P_2 grows exponentially.

As final remark the actual Mathematica code which computes the stability chart of the equation 2 is given in the appendix.

5 Conclusions

The stability analysis for a RLC parametric oscillator when the frequency and the amplitude (Ω,γ) of the excitation source varies is presented by approximating symbolically the monodromy matrix in terms of the parameter γ for each discrete step-size of the frequency Ω, the approximation enable to compute the transition curves in the (Ω,γ)-plane

these are the boundaries between the stable and unstable dominoes, the actual stability chart is given and supported by numerical simulations of the system's response.

Appendix

The following Mathematica© code have been developed to run with version 6.0.

```
In[9]:= Clear[S1, η, ω, β, M, S, m, F, α, f, γ, A, P, q, g, k, k1, L,
  k2, η, β, α, n, t, M, Periodo]

In[10]:= LISTA = {};

For[m = 0.5, m ≤ 10, {Periodo = 2 * Pi / m; p = 1 - 1/4 + γ * Cos[m * t];
  A = N[Expand[{{0, 1}, {-(p), 0}}]]; M = 6; n = 32; H = N[Periodo/n];
  F = A.{x1[t], x2[t]}; {{x1'[t], x2'[t]} = F, f = F, f = Integrate[f, t],
   S = {x1[t], x2[t]}};
  For[i = 1, i ≤ M, {f = Expand[D[f, t]],
    S = TrigReduce[H^i/Factorial[i] * f + S]}; i++];
  S1 = ReplaceAll[S, {x1[t] → x11[t], x2[t] → x22[t]}];
  {{x1[0], x2[0]} = {1, 0}, {x11[0], x22[0]} = {0, 1}}; T = 0;
  For[i = 1, i ≤ n, {{x1[H + T], x2[H + T]} = ExpandAll[ReplaceAll[S, {t → T}]],
    {x11[H + T], x22[H + T]} = ExpandAll[ReplaceAll[S1, {t → T}]], T = T + H},
   i++]; ϕ = Expand[x1[T] + x22[T]]; Blimite = 15; Fb = ϕ - 2;
  Ra1 = Solve[Fb == 0, γ, VerifySolutions→ True]; Fb = ϕ + 2;
  Ra2 = Solve[Fb == 0, γ, VerifySolutions→ True];
  Sol = Sort[Join[Ra1, Ra2]]; Ra = γ /. Sol;
  Ra = Select[Ra, (Im[#] == 0 && Re[#] < Blimite && Re[#] > 0) &];
  For[k = 1, k ≤ (Length[Ra]),
   {LISTA = AppendTo[LISTA, {m, Extract[Ra, k]}]}; k++],
  Print[m]}; m = m + 0.01]

In[16]:= P = ListPlot[LISTA, PlotRange -> {{-1, 5}, {0, 3}}]
```

References

1. Tondl, A., Nabergoj, R., Ruijgrok, T., Verhulst, F.: Autoparametric Resonance in Mechanical Systems. Cambridge University Press, USA (2000)

2. Melde, F.: Uber die erregung stehender wellen eines fadenformigen korpers. Ann. Phys. Chemie. **109** (1860) 193–215
3. Faraday, M.: On a peculiar class of acoustical figures; and on certain forms assumed by a group of particles upon vibrating elastic surfaces. Philosophical Transactions of the Royal Society (London) **121** (1831) 299–318
4. Rayleigh, L.: The Theory of Sound. Volume I. Dover (1945)
5. Mathieu, E.: M'emoire sur le mouvement vibratoire d une membrane de forme elliptique. Journal des Mathématiques Pures et Appliquées: (1868) 137–203
6. Hill, G.W.: On the part of the motion of lunar perigee which is a function of the mean motions of the sun and moon. Acta Math **8** (1886) 1–36
7. Raman, C.: Some remarkable cases of resonance. Physical Review **35** (1912) 449–458
8. van der Pol, B., Strutt, M.: On the stability of the solutions of Mathieu's equation. Philosophical Magazine and Journal of Science **5** (1928) 18–38
9. Billah, K., Scanlan, R.: Resonance, tacoma narrows bridge failure, and undergraduate physics textbooks. American Journal of Physics **59** (1991) 118–124
10. Zhang, W., Baskaran, R., Turner, K.L.: Effect of cubic nonlinearity on auto-parametrically amplified resonant mems mass sensor. Sensors and Actuators A: Physical **102** (2002) 139 – 150
11. Rhoads, J.F., Shaw, S.W., Turner, K.L., Baskaran, R.: Tunable microelectromechanical filters that exploit parametric resonance. Journal of Vibration and Acoustics **127** (2005) 423–430
12. Nayfeh, A.H.: Perturbation Methods. Wiley-Interscience (2000)
13. Jordan, D., Smith, P.: Nonlinear Ordinary Differential Equations. Oxford University Press, USA (2007)
14. Yakubovich, V.A., Starzhinskii, V.M.: Linear Differential Equations With Periodic Coefficients. Krieger Pub Co, Israel (1975)
15. Miyasar, A., Barr, A.: The linear oscillator under parametric excitation with fluctuating frequency. Journal of Sound and Vibration **124** (1988) 79 – 89
16. Szemplinska-Stupnicka, W.: The generalized harmonic balance method for determining the combination resonance in the parametric dynamic systems. Journal of Sound and Vibration **58** (1978) 347 – 361
17. Wereley, N., Hall, S.: Frequency response of linear time periodic systems. In: Decision and Control, 1990., Proceedings of the 29th IEEE Conference on. (1990) 3650 –3655 vol.6
18. Sandberg, H., Mollerstedt, E., Bernhardsson: Frequency-domain analysis of linear time-periodic systems. Automatic Control, IEEE Transactions on **50** (2005) 1971 – 1983
19. Moreno, L., Collado, J.: New scheme for symbolic computation of monodromy matrix. In: Proceedings of The European Control Conference. (2009) 1389–1394.
20. Bally, F.X., B.P.: Un modéle électronique simple et exact de l'oscillateur paramétrique. Bulletin de L'Union des Physiciens **86** (1992) 12671276
21. Bellman, R.: Stability Theory of Differential Equations. Mcgraw-Hill, USA (1953)
22. Winkler, W., Magnus, S.: Hill's Equation. John Wiley and Son Ltd, USA (1966)
23. Arnold, V.I.: Mathematical Methods of Classical Mechanics. 2nd edn. Springer, USA (1997)
24. Chicone, C.: Ordinary Differential Equations with Applications. Springer-Verlag, USA (1999)
25. Stoker, J.J.: Nonlinear Vibrations in Mechanical and Electrical Systems. Interscience Publishers, USA (1954)
26. Henrici, P.: Elements of Numerical Analysis. John Wiley & Sons Inc, USA (1966)

Development and Implementation of a Two Channel System to Measure the Response of Quartz Crystal Resonator Gas Sensors using an FPGA

José Lorenzo Muñoz-Mata[1], Severino Muñoz-Aguirre[1], Humberto González-Santos[1], Georgina Beltrán-Pérez[1], Juan Castillo-Mixcóatl[1],

[1] Facultad de Ciencias Físico-Matemáticas, Benemérita Universidad Autónoma de Puebla Av. San Claudio y Rio Verde, Col. San Manuel, CU. C.P. 72570, Puebla, Puebla, México

Abstract. The development of gas sensors based on quartz crystal resonators requires the use of a frequency counter to measure the sensor response. However, the commercial frequency counters have at most 2 channels and in odor sensory systems (Electronic Noses), there are used arrays of at least 8 sensors. Furthermore, it is also desirable that the signal processing and data analysis stage can be included in the same system. FPGAs can provide a solution to this kind of problems. The present work involves the development of a high resolution frequency counter using a FPGA for quartz crystal resonators, which is capable of processing and storage the data in 32-bit registers. The data transfer and acquisition was performed by a microcontroller. The virtual instrumentation software LabVIEW was used for the sensor response display and data storage. Tests were performed using gas sensors to obtain real data. The system was scaled to two channels verifying that it behaves in a correct and stable way.

Keywords: FPGA, Frequency Counter, Gas Sensor.

1 Introduction

Current technology on electronic noses has had substantial progress in the recent years [1]. Sensor arrays that respond to a wide range of compounds, as well as advanced pattern recognition and artificial intelligence techniques, which allow the user to extract relevant and reliable information, have been used. An electronic nose has been defined as a system that detects and identifies odors and vapors, typically by using chemical sensing with signal processing and pattern recognition subsystems [2].

The development and utilization of gas sensor arrays (commonly called Electronic Noses, EN) have received a high importance in the field of the scientific research due to the necessity of working on detection, recognition and discrimination of gases. This is a high interest area owing to the enormous variety of gas sensors as well as the diversity of feasible materials to be used as sensing films. Likewise, there are many applications of the system such as air quality monitoring, quality control in the food industry and beverages, cosmetology, biotechnology, etc. [3].

The quartz crystal microbalance (QCM) has been used in sensor applications such as gas-mass detectors and in recent years its applications have been extended since

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 235-244

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

scientists realized that it can be operated in contact with liquids and viscoelastic deposits [4]. Moreover, applications of QCM as gas sensors are widely used due to their high correlation with the human nose.

The QCM gas sensor operation principle is based on the fact that when the gas molecules interact with the sensing film, its mass increases and the resonance frequency decreases owing to the mass loading effect. This frequency shift is proportional to the amount of mass that interacts with the sensing film and it is known as the sensor response [5, 6].

The frequency shift is described by the Sauerbrey equation (Eq. 1) [7], which relates the mass changes in the sensing film per unit area at the QCM electrode surface with the observed change in the crystal oscillation frequency.

$$\Delta f = -2.3\times10^{-6} \cdot F^2 \frac{\Delta m}{A} \tag{1}$$

Where Δf is the frequency shift (Hz), -2.3×10^{-6} is a constant obtained from the quartz density (ρ_q) and the shear modulus(μ_q), Δm is the mass (g) of the adsorbed gas molecules, A is the coated area (cm^2) and F is the fundamental resonance frequency (Hz) of the crystal.

The use of a frequency counter is essential for the measurements of gas sensors based on quartz resonators, as it was mentioned above. Therefore, in order to characterize QCM sensors is necessary to observe the shift rate in the sensor frequency. Furthermore, coating the QCM with different types of sensing films it will be possible to obtain different characteristics and to observe the tendency of each sensor achieving more precise measurements for many gases or vapors [5]. Then it is very important the use of a frequency counter in order to measure the frequency variation of one or various sensors. Although some commercial frequency counters are commercially available, they only have one or at least two input channels; however in the electronic nose it could be necessary to use a large number of sensors. Therefore, it is necessary to develop frequency counters with more than two input channels, which must have the characteristics of commercial systems. Virtually any frequency counter commercially available can measure *5 Vpp* square wave frequencies with 50 Ω output impedance provided by the QCM [8].

The present work reports the development of a frequency counter, which is part of a characterization system to measure quartz crystal resonator gas sensors. The counter was designed in an FPGA and it has been connected with an acquisition card, which in turn was in communication with the computer. The virtual instrumentation software LabVIEW was used for the software development in order to display and storage information. Tests were performed using real sensors to obtain information of real measurements. The system was scaled to two channels verifying its stability for one as well as two channels. Finally as a future work the system will be upgraded at least to four channels in order to perform a characterization of various gas sensors, simultaneously.

2 Experimental Set-Up

The implemented experimental set-up is shown in Figure 1. The frequency counter was designed using an Altera Field Programmable Gate Array (FPGA) from the Cyclone II family [9] which has a 50 MHz clock inside. A function generator (AFG3102, Tektronix) was used to evaluate the counter performance. An oscilloscope (TDS-3034, Tektronix) was used to observe the measured signals and a data acquisition card (USB-DAQ) was used to send the information to a computer. In order to verify the frequency counter logic correct performance, it was used a logic analyzer 1582AD (Agilent) [10].

2.1 Frequency Counter Design

The direct frequency counting method was used for the frequency counter design. This method performs the frequency counting from the sensor and the oscillator circuit during a timebase period of one second, as shown in Figure 2.

To perform the frequency count Equation 2 was used.

$$Frequency = \frac{Pulse\ Number}{t_0}. \tag{2}$$

Therefore, if the timebase is one second, the frequency is simply equal to the counted pulse number.

The frequency counter firmware was developed using the software Quartus II. The implemented logic was designed using blocks in order to develop the frequency counter for one channel, which is shown in Figure 1. The blocks are described below.

Timebase. A state machine was designed, which is in charge of initialize the timebase of one second period. The state diagram implemented is shown in Figure 3a. Once the start operation signal is received from the computer via the microcontroller the enable signal is generated, which is also the frequency counting activation signal. Moreover, it is used to enable the data storage in the corresponding block. If the start operation signal is not present the state does not change, otherwise this signal will be automatically generated by this block each second starting another timebase period until a stop signal is received from the computer.

Frequency Counter. It was developed a circuit that verifies the arrival of the enable signal, which was produced by the timebase block. This means that when a timebase is initiated, the frequency counting starts. The counting is performed using a counter circuit, which consists of an array of 32 enable type flip-flops. The clock signal of the counter circuit is taken from the oscillation circuit, which would be the signal generated by the QCM sensor, as is shown in the dataflow diagram in Figure 3b. The stored data are deleted after they are transmitted to the microcontroller by a clear signal.

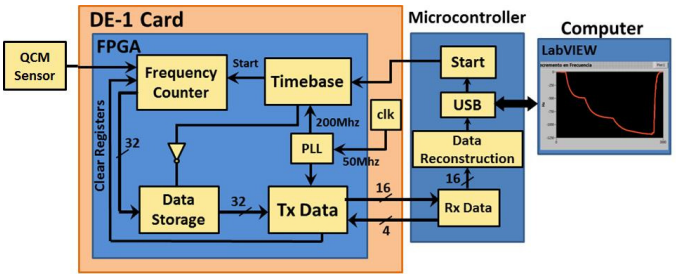


Fig. 1. Block diagram of the implemented system.

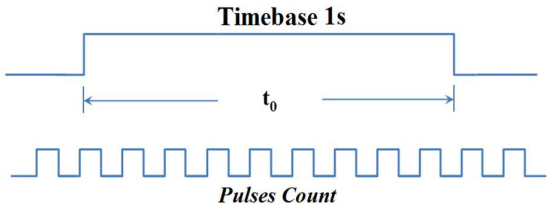


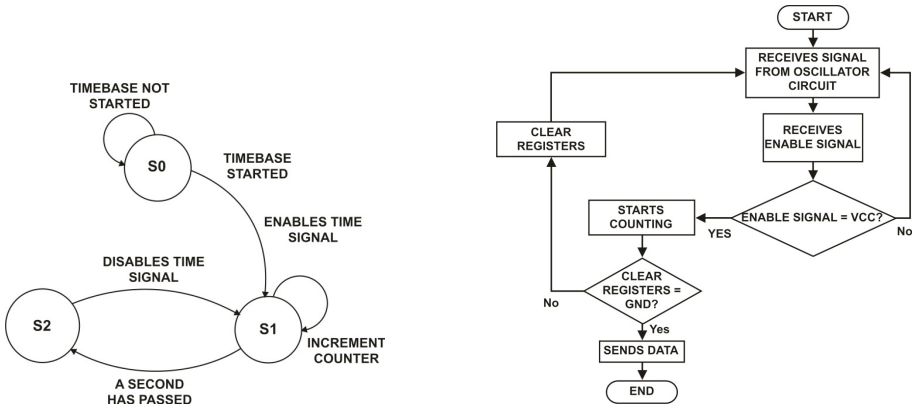
Fig. 2. Frequency counter operation principle.

Data Storage. The data storage block was designed using a circuit implemented with latch devices, which store the information sent by the frequency counter block when the timebase finishes. This means that one second has passed and then the enable signal is activated in order to capture the information (See Figure 4a). The obtained data from the sensor are stored in 32-bit registers in such way that they can be sent to the computer through an interface card via USB.

Data Transmission. For the data transmission, a strategy was designed to send the data from the FPGA in 16-bit packets. The data flow is controlled by a microcontroller using various control bits. When the microcontroller is ready to receive the data, it sends a control signal in order to ask for the first 16 bits packet, in this case the most significant bits (MSB). Once the microcontroller receives the first data packet, it sends the next control signal to receive the next 16-bit packet (LSB) as is shown in Figure 4b.

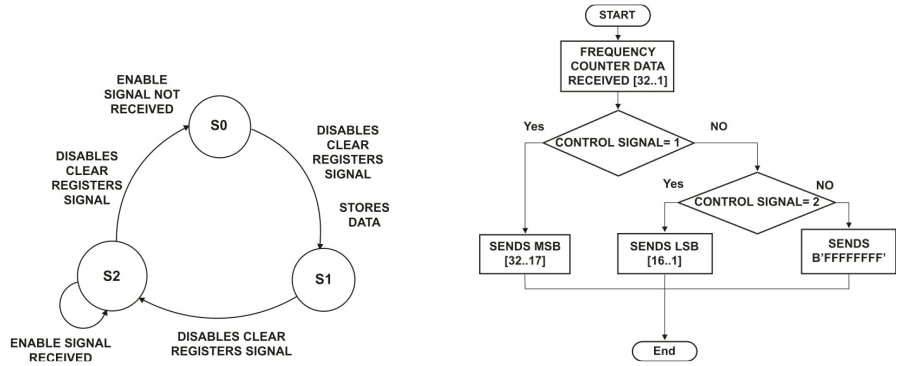
Acquisition Card Design. To perform the data acquisition from the FPGA, it was fabricated an acquisition card (USB-DAQ) using the microcontroller PIC18F4455 (Microchip) [11], which contains the Universal Serial Bus (USB) communication module. The printed circuit (PCB) was designed using the Altium Designer 6.0 software.

Data Acquisition Firmware Development. For the data acquisition from the FPGA to the computer it was developed the microcontroller firmware using PIC C Compiler 4.0 [12].



a) b)

Fig. 3. a) Timebase state diagram. b) Frequency counter flowchart. Both implemented in the FPGA



a) b)

Fig. 4. a) Data storage state diagram. b) Data transmission flowchart. Both implemented in the FPGA

First of all, the communication parameters setting is performed, i.e. to define the type of communication to be performed, in this case the bulk communication [13], the size of memory reserved for data transmission, the variables declaration as well as the in-out port configuration. The microcontroller USB module is initialized and it verifies if it is connected to the computer in order to start the device enumeration process. Then an endless loop begins in which the device waits for being enumerated by the computer. Once the device (in this case the microcontroller) has been enumerated by the computer, it checks if some data has been received. If this is true the data are stored in a buffer called *endpoint*, which will contain such information until the device decides to accept these data or data packet [14]. The Figure 5 shows the flowchart with the implemented algorithm for the data acquisition.

To be able to perform the data transfer control among the FPGA, the microcontroller and the computer, it was used a multiplexer, which receives the con-

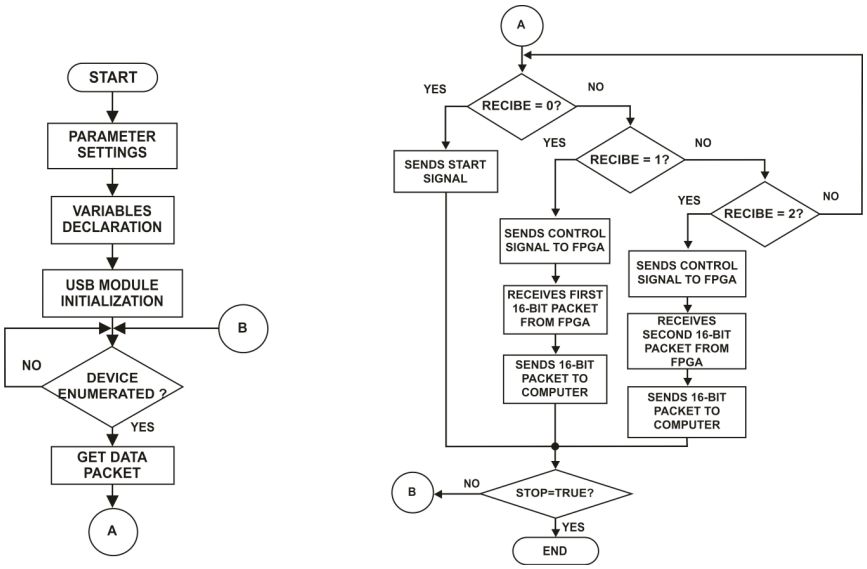


Fig. 5. Acquisition data flowchart of the implemented algorithm in the microcontroller.

trol signal from the computer. When the microcontroller receives the first control signal, it sends the start signal to the frequency counter timebase block inside the FPGA in order to start the frequency counting. The second control signal from the computer indicates to the microcontroller that it must send the corresponding control signal to the FPGA data transmission block to obtain the first 16 MSB data packet from the 32-bit register storage. The maximum number of bits available to receive in parallel by an input port in the microcontroller is 8 bits; therefore the 16-bit packet is received through two 8-bit ports to be able to achieve the acquisition in parallel. Once the frequency counter data are received, the information is recovered in a 16-bit register and these data are sent to the computer. The following 16 bits are sent to the computer where the complete information (32 bits) is reconstructed by software.

Software Development. The software development was carried out with the instrumental software LabVIEW v9.0 [15]. The flowchart for the implemented software is shown in Figure 6.

The first control signal is generated by the computer and it is sent to the microcontroller, which sends the start signal to the FPGA. Then the second control signal is sent to the microcontroller which, in turn, sends its respective control signal and receives the first data packet form the FPGA. After receiving the first data packet from the microcontroller, these are sent to the computer and it performs the same process for the second data packet. When the computer obtains both data packets they are reconstructed and displayed on an indicator, plotted and stored. The same process is repeated in a cycle of approximately three minutes, which is in order to wait for the sensor to attain stability. After that, the software asks if tree minutes have passed to be able to capture a data and to make the baseline zero and these data are displayed, plotted and stored, respectively. This process is performed to have a better understanding of the interaction mechanism between the gas molecules and the sen-

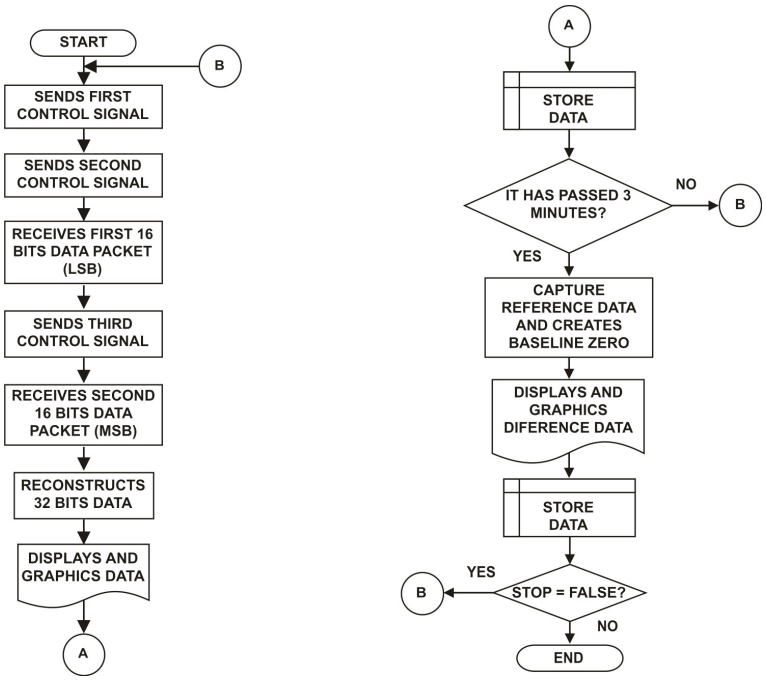


Fig. 6. Acquisition Data flowchart of the software algorithm in the computer.

sing film, which is related to the decrease in QCM resonant frequency by applying a gas to the sensor.

3 Experimental Results

Once the USB-DAQ card for the data acquisition, the frequency counter with the FPGA and the computer software were achieved, the connection of the whole system was performed. A test firmware was developed to produce an incremental data within the FPGA using the same information transmission strategy. The preliminary experimental set-up was implemented using a function generator, an oscilloscope, an Altera development board DE-1 connected with the acquisition card USB-DAQ for the purpose to verify that the frequency counter data were correctly obtained by the computer. The preliminary experimental set-up is shown in the Figure 7.

A 100 KHz square signal was applied to the system taking the function generator as pattern. The obtained information in the computer generated by the FPGA frequency counter corresponds quite well with the input frequency. Subsequent, higher frequencies (1 MHz, 10 MHz, 20 MHz and 50 MHz) were tested with satisfactory results and it can be said that the system has a stable behavior similar to the function generator (1ppm). Until then the software was only able to display the data, plots and store information. Then, it was performed a test using real sensors with controlled temperature. The experimental set-up is shown in Figure 8. In order to

perform such test, additionally it was used a refrigerated bath (RTE-10, Neslab) in order to keep the sensor temperature controlled. Also, it was used a measurement chamber made of Teflon which is capable of resisting many organic solvents (See Figure 8). The sensor was placed inside the measurement chamber. In this case a 12MHz sensor coated with an ethyl cellulose sensing film of 29 KHz (Δf) was used. The response was measured applying three 7600 ppm concentration steps of ethanol and the sensor response was measured and stored. For this experiment the entire software was used for a single channel where the response of the sensor can be visualized in the computer.

At the beginning of the test the sensor response is let to stabilize for approximately three minutes to capture the baseline data, while the sensor response is plotted and stored. Once the baseline data is captured, a difference is made between this value and the response of the sensor to obtain the absolute response (Δf); all these data are plotted and stored as is shown in Figure 9. In the left chart the sensor raw frequency data are shown, while in the right chart the frequency decrement of the sensor (Δf) is plotted. It could be observed the response of the sensor when the ethanol sample was injected as well as the frequency decrement of approximately 492 Hz after 15 minutes of stabilization. A second sample was injected, which provoked a frequency decrement of approximately 381 Hz for a total decrement of 873Hz after stabilization, 14 minutes after the second injection. The third sample was injected and the sensor had a frequency decrement of 303 Hz for a total decrement of 1176Hz after 17 minutes. Finally the measurement chamber was purged pumping air and the sensor response recovery can be observed. The test finishes after 50 minutes. All these data were displayed, plotted and stored into a file. The plotted response is quite stable and smooth since the sensor stability is approximately ±1Hz. Therefore, it can be said that the frequency counter was able to measure with such stability.

Finally, the system was scaled to two channels and it was tested using real sensors exposed to ethanol vapor at room temperature. The results are shown in Figure 10, where can be observed that the system can register the frequency response of both sensors at the same time. The system can register frequency variations that take place in a time period of one second.

4 Conclusions

A frequency counter using an FPGA was developed to measure the response of gas sensors based on quartz crystal resonators. Tests on the communication, among the FPGA, the data acquisition card and the computer were performed, generating an internal data within of the FPGA and verifying the effectiveness of the system. There were performed tests on the whole system integrated using as input data a function generator and a stability similar to this instrument (1ppm) was obtained. It was achieved a software upgrade in order to process the information gathered from the sensors, which means to obtain the raw frequency of the sensor, as well as the frequency decrement (Δf). In order to verify the system performance, there were performed tests using a gas sensor and applying three concentration steps of ethanol vapor. The obtained results showed a smooth behavior of the response with a stability of approximately ±1Hz for 12 Mhz sensors. The system was scaled to two channels

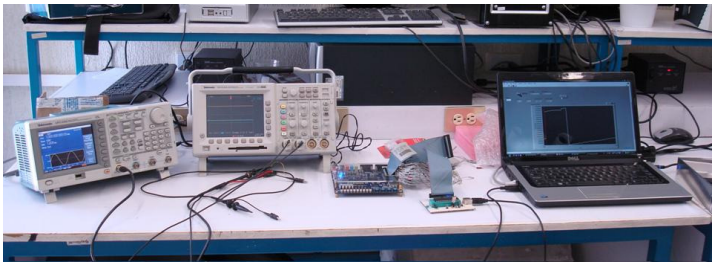


Fig. 7. Experimental set-up to measure frequency with a function generator.

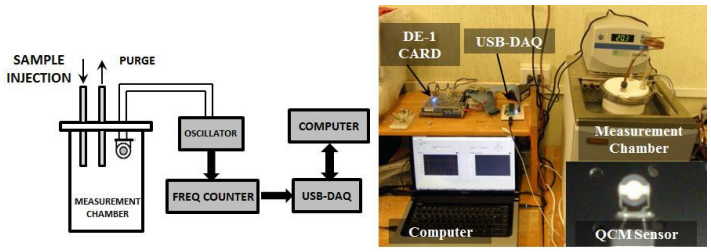


Fig. 8. Experimental set-up using gas sensors at controlled temperature.

using a preliminary software version and performing tests for two gas sensor exposed to ethanol gas at room temperature.

Currently a work under progress is on the information process to achieve a system robust enough in order to obtain a good sensor characterization and be able to automate the whole system, as well as a four channels system expansion.

Acknowledgments

This work was financially supported by the project Conacyt Jovenes Investigadores 61126.

References

1. H. Troy Nagle & R. Gutierrez-Osuna, S. S. Schiffman. The How and Why of Electronic Noses. IEEE Spectrum, September (1998)
2. J. W. Gardner, P. N. Bartlett, Electronic Noses – Principles and Aplications. Oxford Science Publications, (1999)
3. S. Muñoz-Aguirre, T. Nakamoto, A. Yoshino, T. Moriizumi, Odor Approximation of fruit flavors using a QCM odor sensing system. Sensor an Actuators B 105 (2005) 144-149
4. A. Arnau, Piezoelectric Transducers and Applications, Springer, (2004)
5. S. Muñoz Aguirre, "Study of sensing films structure controlled by LB method for QCM odor/gas sensor", Doctoral Thesis, Department of Electrical Engineering, Tokyo Institute of Technology. June (1999)

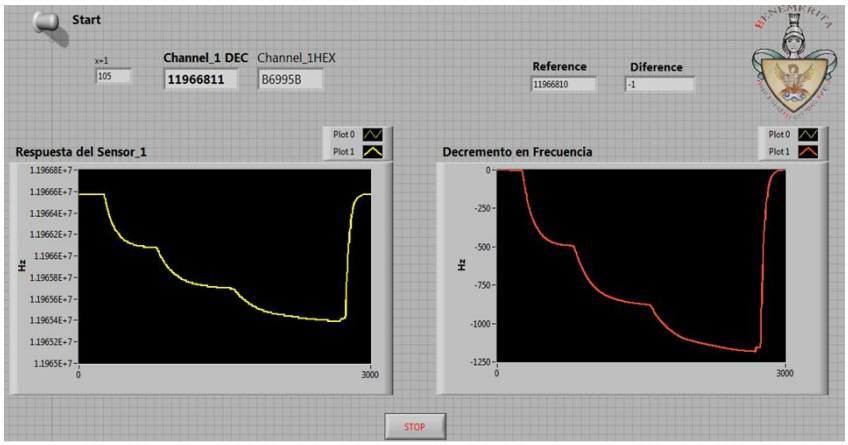


Fig. 9. Sensor response plotted as raw data (left) and frequency decrement (Δf)

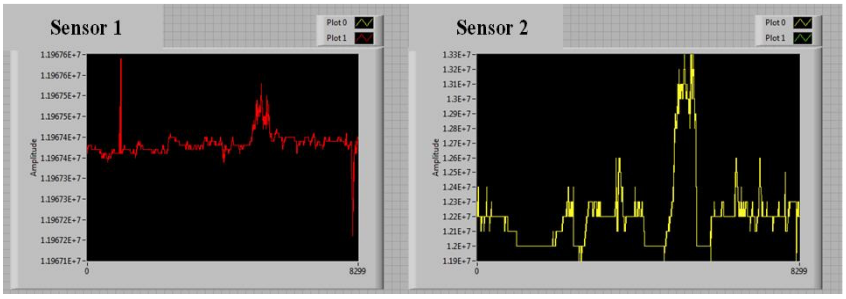


Fig. 10. Data obtained using two sensors tested at room temperature with the two-channel frequency counter.

6. S. Muñoz-Aguirre, TakamichiNakamoto, ToyosakaMoriizumi. Study of deposition of gas sensing films of quartz crystal microbalance using an ultrasonic atomizer. Sensors an Actuators B 105 (2005) 144-149
7. W. H. King. Jr., Piezoelectric Sorption Detector, Analytical Research Division, Esso Research and Engineering Co., Linden N.J. Vo.36, No. 9 Aug (1964)
8. SRS Application Notes, QCM- Quartz Crystal Microbalance Theory and Calibration. Stanford Research Systems. http://www.thinksrs.com/downloads/PDFs/ApplicationNotes/QCMTheoryapp.pdf
9. http://www.home.agilent.com/
10. http://www.terasic.com.tw/cgi-bin/page/archive.pl?Language=English&CategoryNo=53&No=83&PartNo=2#section
11. Microchip. 28/40/44 Pin High Performance, Enhanced Flash USB Microcontrollers DS39617A, (2003)
12. Custon Computer Services. C Compiler Reference Manual. (2009)
13. Eric Lopez Pérez. Ingeniería en Microcontroladores, Protocolo USB (Universal Serial Bus)
14. Derek Liu Asia & Pacific Operation Microcontroller Division. USB Device Development with the MC9S08JM60, Freescale Semiconductors Application Note (2008)
15. http://digital.ni.com/

Hybrid Routing Protocol for Wireless Ad-Hoc Networks

Aldo Méndez[1], Marco Panduro[1], Omar Elizarraras[1], Manuel Munguía Macario[2],

[1] Universidad Autónoma de Tamaulipas, UAMRR, Carr. Reynosa-San Fernando S/N, Colonia Arcoiris, 88779, Reynosa-Tamaulipas, México
{almendez, mamendoza}@uat.edu.mx
[2] Universidad Autónoma de Nuevo León, FIME, Av. Cd. Universitaria s/n, 66451, San Nicolás de Los Garza, Nuevo León, México
manuel.munguiamc@uanl.edu.mx

Abstract. Wireless ad-hoc network is a decentralized network of autonomous nodes able to communicate with each other over wireless links. A fundamental assumption in ad-hoc networks is that any node can be used to forward packets between arbitrary sources and destinations. Some sort of routing protocol is needed to make the routing decisions. In addition, in a wireless environment the ad-hoc network presents many problems such as mobility and limited bandwidth which makes routing difficult. Therefore, this paper presents a routing protocol for wireless ad-hoc networks. The proposed protocol is hybrid type where each node maintains routing information only to those that are within its zone, and information regarding only its neighboring zones. This proposed protocol is evaluated in effectiveness and advantages it can offer with respect to HARP and FSR protocols.

Keywords: HARP protocol, FSR protocol, routing, ad-hoc network, hops.

1 Introduction

Wireless ad-hoc networks are autonomous systems of nodes forming network in the absence of any centralized support. By routing packets cooperatively among the nodes, these nodes can communicate with each other without any infrastructure where each node itself acts as a router for forwarding and receiving packets to/from other nodes [1]. The problem in the ad-hoc networks is the change in network topology due to the node mobility. In addition to, the design of network protocols for ad-hoc networks is a complex issue. Ad hoc wireless network routing protocols are usually divided into two groups: proactive (table driven) and reactive (on-demand) routing [2]. Proactive protocols exchange routing information periodically between hosts and maintain a set of available routes in the network [3], [4]. In contrast, reactive protocols, such as [5] and [6], delay route discovery until a particular route is required, and propagate routing information only on demand. The hybrid protocols combine proactive and reactive routing strategies [7], [8].

To determine the network organization and routing in the ad-hoc networks we need establish efficient distributed algorithms [9]. In a decentralized network the message

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 245-254
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

routing is not a well-defined problem because the network topology changes. Therefore, it is necessary to design routing protocols where any node can be used to forward packets between source and destination [10]. A wireless ad-hoc environment presents many problems such as mobility and limited bandwidth which makes routing difficult. Therefore, this paper presents a novel routing protocol for wireless ad-hoc networks. The proposed protocol is a hybrid type where each node maintains routing information only to those that are within its zone, and information regarding only its neighboring zones. This proposed protocol is evaluated in effectiveness and advantages it can offer with respect to the routing protocols: Hybrid Ad hoc Routing Protocol (HARP) [8] and Fisheye State Routing (FSR) [11]. Where FSR is a proactive routing protocol and it fisheye technique to reduce information required to represent graphical data, to reduce routing overhead. FSR maintains the accurate distance and path quality information about the immediate neighboring nodes, and with the progressive detail as the distance increase. In FSR, link state packets are not flooded. In addition to, nodes exchange link state information only with the neighboring nodes to maintain up-to-date topology information. Through this exchange process, the table entries with larger sequence numbers replace the ones with smaller sequence numbers. The complete topology information of the network is maintained at every node and the desired shortest paths are computed as required. The topology information exchange takes place periodically rather than being driven by an event. This is because instability of the wireless links may cause excessive control overhead when event-driven updates are employed. Moreover, HARP is a hybrid routing scheme, which exploits a two-level zone based hierarchical network structure. Different routing approaches are utilized in two levels, for intra-zone routing and inter-zone routing, respectively. The Distributed Dynamic Routing (DDR) algorithm is exploited by HARP to provide underlying supports. In DDR, nodes periodically exchange topology messages with their neighbors. A forest is constructed from the network topology by DDR in a distributed way. Each tree of the forest forms a zone. Therefore, the network is divided into a set of non-overlapping dynamic zones. A mobile node keeps routing information for all other nodes in the same zone. The nodes belonging to different zones but are within the direct transmission range are defined as gateway nodes. Gateway nodes have the responsibility forwarding packets to neighboring zones. In addition to routing information for nodes in the local zone, each node also maintains those of neighboring zones [12].

The remainder of this paper is organized as follows. In section 2, we present the operation of the proposed routing protocol. Our proposed routing protocol is simulated and compared with FSR and HARP in section 3 and the conclusions of this work are provided in the last section.

2 Proposed Routing Protocol

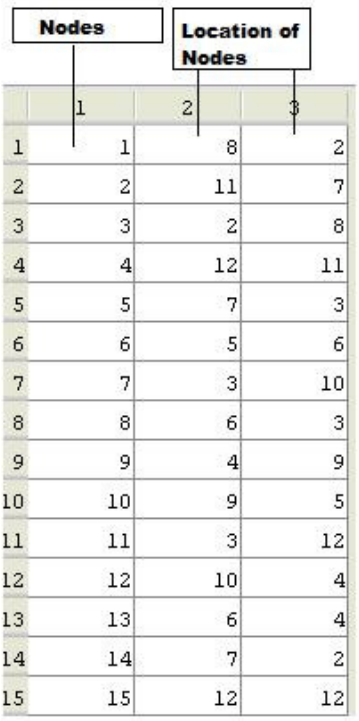
The objective of the proposed routing protocol is that communication between nodes is efficient taking into account the cost, route, distance, bandwidth used, and delay. Where the cost is the value assigned to each node to form the route path, the route is

the path from the source node to the destination node, and the distance is the value obtained from the path of the route.

The structure of the proposed protocol is based on several features, including: routing table, network structure, sectors, intra-zones, and preferred neighbor node [8], [11].

The routing table in the proposed protocol is used to store the information of nodes that are within of the network. Figure 1 shows an example of a routing table.

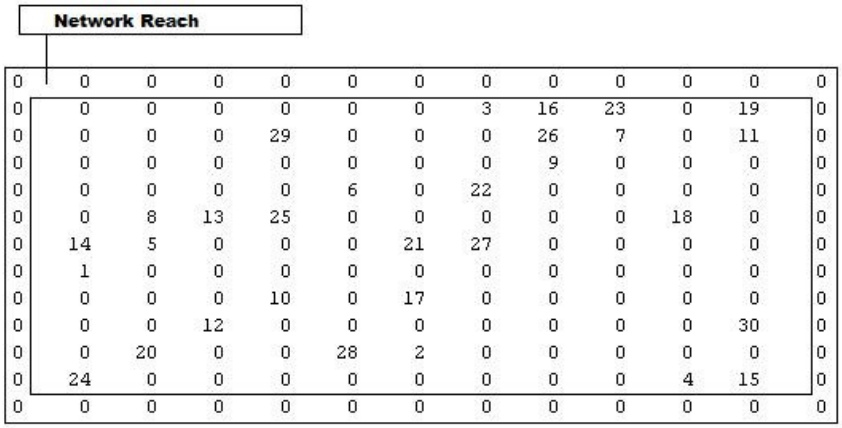
The network structure in the proposed protocol is used to place the available nodes to transmit. A value 0 indicates that position is available for a node, except the 0s that are in the border of the network and these serve to delimit it (see figure 2).



	1	2	3
1	1	8	2
2	2	11	7
3	3	2	8
4	4	12	11
5	5	7	3
6	6	5	6
7	7	3	10
8	8	6	3
9	9	4	9
10	10	9	5
11	11	3	12
12	12	10	4
13	13	6	4
14	14	7	2
15	15	12	12

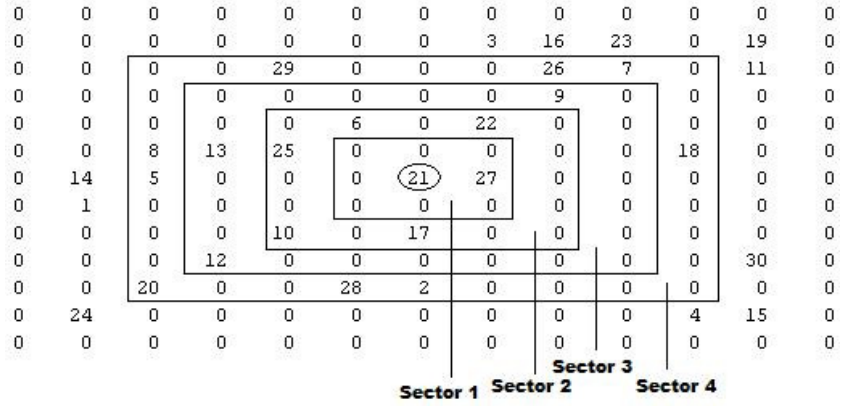
Fig. 1. Routing table of the proposed routing protocol.

The sectors are determined by levels according to network matrix taking the node into account that needs to transmit the message. The first sector is composed of neighboring nodes of the destination node. The second sector is formed by the nodes around the first sector, and the third sector will be conformed by the nodes that are around the second sector, as shown in figure 3.



0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	3	16	23	0	19	0
0	0	0	0	29	0	0	0	26	7	0	11	0
0	0	0	0	0	0	0	0	9	0	0	0	0
0	0	0	0	0	6	0	22	0	0	0	0	0
0	0	8	13	25	0	0	0	0	0	18	0	0
0	14	5	0	0	0	21	27	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	10	0	17	0	0	0	0	0	0
0	0	0	12	0	0	0	0	0	0	0	30	0
0	0	20	0	0	28	2	0	0	0	0	0	0
0	24	0	0	0	0	0	0	0	0	4	15	0
0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 2. Network structure for the proposed routing protocol.



0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	3	16	23	0	19	0
0	0	0	0	29	0	0	0	26	7	0	11	0
0	0	0	0	0	0	0	0	9	0	0	0	0
0	0	0	0	0	6	0	22	0	0	0	0	0
0	0	8	13	25	0	0	0	0	0	18	0	0
0	14	5	0	0	0	21	27	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	10	0	17	0	0	0	0	0	0
0	0	0	12	0	0	0	0	0	0	0	30	0
0	0	20	0	0	28	2	0	0	0	0	0	0
0	24	0	0	0	0	0	0	0	0	4	15	0
0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 3. Sectors location for the proposed routing protocol.

The intra-zones are located within a sector as shown in Figure 4. This is a difference of proposed protocol with respect to the HARP protocol, because the hops number between transmitting node and receiving node is less due it moves according to the sector in which it is.

The preferred neighbor node is determined by taking the shortest distance between nodes of one sector to another (see figure 5), i.e. in order to reduce the delay time of transmission of the message. This is a difference between proposed protocol and FSR.

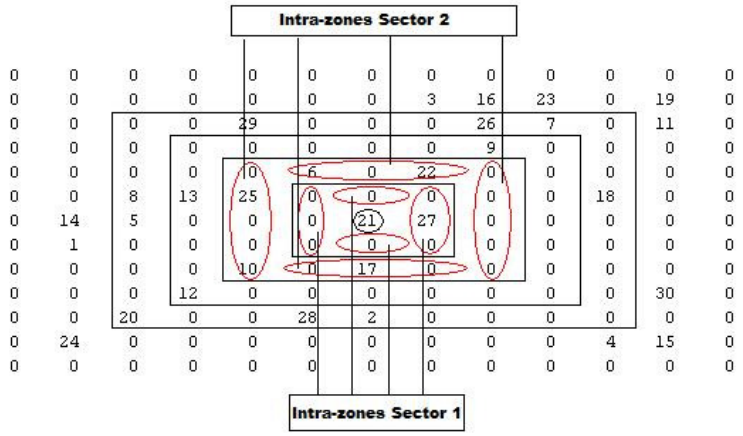


Fig. 4. Intra-zones belonging to each network sector.

2.1 Operation

Step1. When a node of the network needs to transmit a message, it is situated in the center of the network. In addition to, the proposed protocol establishes the sectors and intra-zones and these determine the reach of each node with their neighboring nodes (see figure 6).

Step 2. Having defined the node transmitter and receiver, the proposed protocol analyzes the first sector. In this case, it verifies the routing table of each intra-zone and if there is no information about the destination node, then it goes to other sectors taking the preferred node into account (see figure 7).

Step 3. The destination node in sector 1 is searched, but in case of not receiving affirmative answer then the search continues into other sectors. For this, locate the preferred neighbor node which is determined based on the nearest node of the new sector, as shown in figure 7. The result of this would be a decrease in the transmission delay. When the preferred neighbor node is found the search can continue at sector 2 and performs the same operation as step 2.

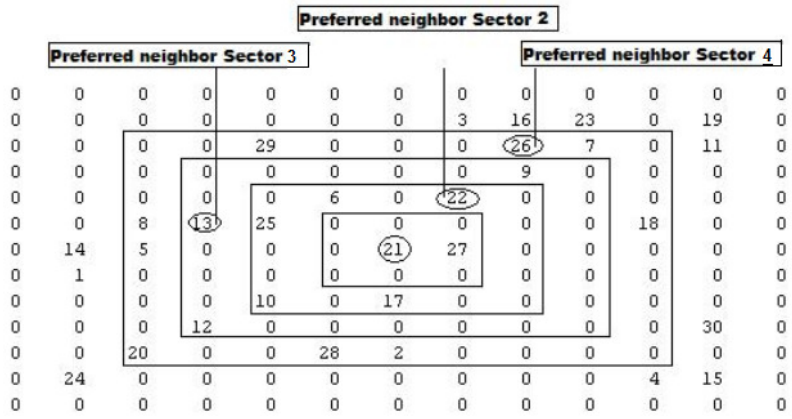


Fig. 5. Preferred neighbors of network sector.

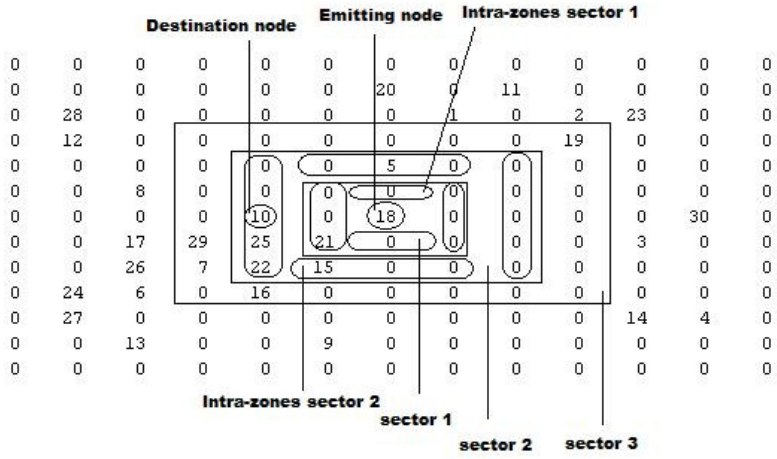


Fig. 6. Network structure in the transmission of the message.

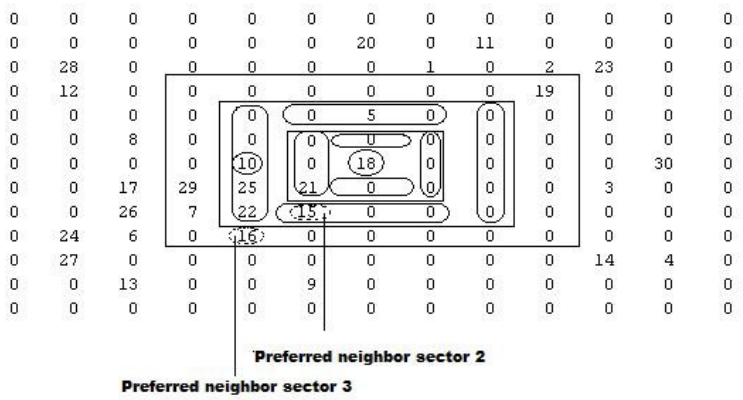


Fig. 7. Selecting the preferred neighbor.

Step 4. At the time receiving response of an intra-zone of which destination node is in that sector, the proposed protocol begins to analyze the possible routes and using Dijkstra selects the best route [13].

Step 5. The proposed protocol calculates distance, hops number and delay once selected the route.

The evaluation results of the proposed routing protocol are presented in the next section.

3 Performance Evaluation

The proposed protocol is evaluated and compared with respect to HARP [8] and FSR [11] by a simulation process, and it was made through the design of a program in Matlab, on a personal computer, which considers the aspects of modeling of proposed routing protocol mentioned in section 2, as well as the properties and the behavior of the system. We have considered a network with 105 nodes, where these are located randomly. The number of point-to-point links in a transmission path is the hops count. Within the network each node is assigned a value and it is called as cost. Distance is obtained value of the path from source node to destination node. Delay is the average time duration of packet transmitting in the network form a source node to the destination node. The simulation results are the following.

a) Hops number. Figure 8 shows that proposed protocol achieves an improvement of 12% and 35% as compared to the existing FSR and HARP protocols, respectively. On the other hand, this improvement is attributed to that the proposed protocol chooses the nearest node, and protocol FSR takes a neighbor node but it does not verify if it is the closest. The protocol HARP does not have complete information of all the nodes of the network, and it has to go asking for zone and intra-zone to reach the destination node.

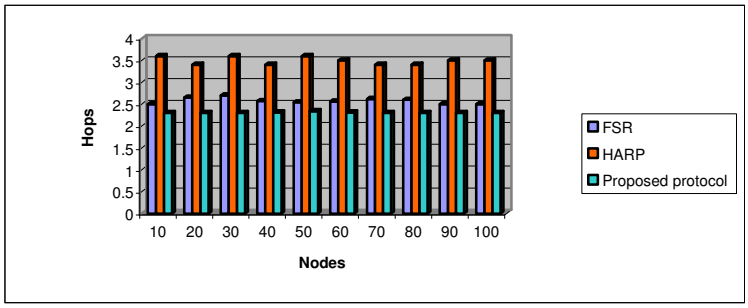


Fig. 8. Hops number behavior under FSR, HARP and proposed protocol.

b) Distance. When compared to the FSR and HARP with our proposed protocol illustrated in figure 9, improves in 14% to FSR and 27% to HARP. This improvement, is because the route selection in HARP is based on stability with the destination node. With respect to FSR, it does not take into account the preferred neighbor node to consult to the other sectors on destination node information, when not doing this the protocol takes a random node but not always it is nearest.

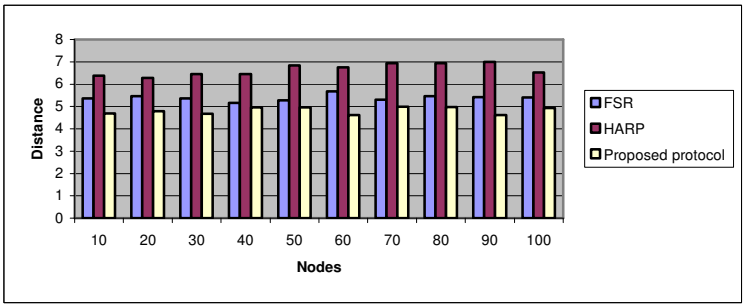


Fig. 9. Distance comparison of FSR, HARP and proposed protocol.

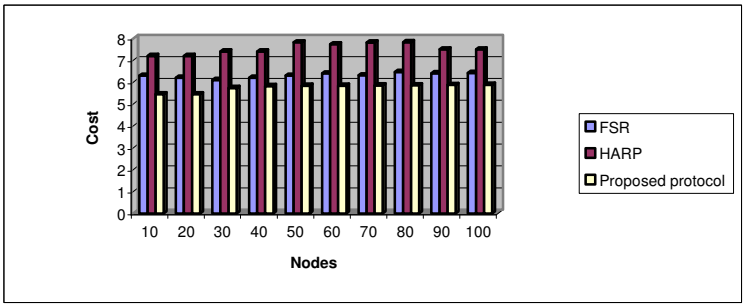


Fig. 10. Simulation result to cost.

c) Cost. Figure 10 shows the behavior of FSR, HARP, and proposed protocol. We observe that the proposed protocol has a lower cost. For the case HARP, having a

greater distance it goes through a larger nodes number requesting information of the destination node. The protocol FSR does not have a preferred neighbor node, so it can pass through nodes that are not required to obtain information of the destination node.

We have seen the advantages of our proposed protocol in terms of hops, distance and cost. We are now going to analyze the behavior of another very important system parameter, which is the delay.

d) Delay. The delay result is illustrated in figure 11. In this figure we observe that the proposed protocol has less delay that HARP and FSR because if the node is within the sector then it can be easily localized with the table of the intra-zones, in case it is not in the sector then it chooses the preferred neighbor, thereby ensuring communication with the other area is optimal. Moreover, the protocol HARP searches the information of the destination node in each intra-zone and therefore a greater delay in the transmission. Additionally, the FSR protocol has no preferred neighbor property; therefore the selected node is not the most appropriated. In addition, as FSR does not use intra-zones then has greater delay because the protocol FSR must consult the routing table of all the sector and when not finding information it takes a node of the following sector without evaluating if it is nearest.

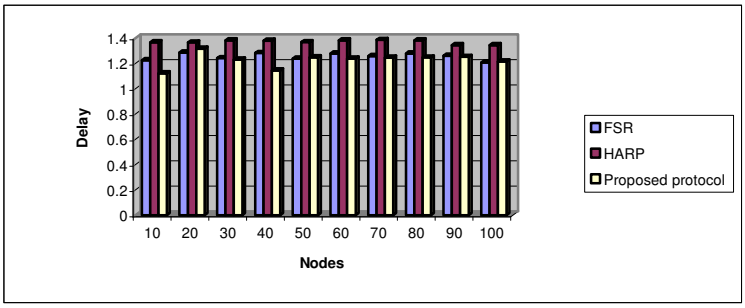


Fig. 11. Simulation result to delay.

4 Conclusions

By according to the results we conclude that the proposed protocol is more efficient and dynamic as it manages the bandwidth efficiently to the node or nodes that need to transmit, selects the shortest route by analyzing each node and obtaining a lower cost. Because the industry of the telecommunications is growing, there is a good option for the implementation of an ad-hoc network under proposed protocol since it would guarantee a very good service to the users and it would stay stable according to the amount of users who connect themselves to the network.

5 References

1. Mishra, M.R., Panda, M.R., Bisoyi,S.K.: Topographical Automation of MANET using Reactive Routing Protocols. International Journal of Computer & Communication Technology. 2, 24-29 (2010)
2. Royer, E., Toh, C-K.: A Review of Current Routing Protocols for Ad-Hoc Mobile Wireless Networks. IEEE Personal Communications. 6, 46-55 (1999)
3. Perkins, C., Bhagwat, P.: Highly Dynamic Destination Sequenced Distance Vector Routing for Mobile Computer. In: Proceedings of ACM SIGCOMM. ACM, New York (1994).
4. Jacquet, P., Muhlethaler, P., Clausen, T., Laouiti, A., Qayyum, A., Viennot, L.: Optimized Link State Routing Protocol for Ad-Hoc Networks. In: Proceedings IEEE INMIC 2001, pp. 62-68. IEEE Press, Lahore (2001)
5. Royer, E.M., Perkins C.E.: Ad-Hoc On-Demand Distance Vector Routing. In: Proceedings 2nd IEEE Workshop on Mobile Computing Systems and Applications, pp. 90-100. IEEE Press, New Orleans (1999).
6. Johnson, D.B., Maltz, D.A., Broch, J.: DSR: The Dynamic Source Routing Protocol for Multi-Hop Wireless Ad-Hoc Network. In: Perkins, C.E (ed.). In Ad-Hoc Networking, Chapter 5, pp. 139-172, Addison-Wesley (2001)
7. Lemus, A., Mendez, A.: Performance Improvement for Ad-Hoc Networks with ZRP using Route Maintenance. WSEAS Transactions on Computers. 3, 831-838 (2004)
8. Nikaein, N., Bonnet, C., Nikaein, N.: Hybrid Ad Hoc Routing Protocol- HARP. In: Proceedings of IST 2001. http:// www.cs.cornell.edu/people/egs/615/harp.pdf
9. Hoebeke, J., Moerman, I., Dhoedt, B., Demeester, P.: An Overview of Mobile Ad Hoc Networks: Applications and Challenges. Journal of the Communications Network. 3, 60-66 (2004).
10. Murthy, C.S.R., Manoj, B.S.: Ad-Hoc Wireless Networks: Architecture and Protocol, Prentice Hall PTR, New York (2004)
11. Gerla, M., Hong, X., Pei, G.: Fisheye State Routing Protocol (FSR) for Ad-Hoc Networks. IETF MANET Working Group, INTERNET DRAFT, June 17, 2002. http://tools.ietf.org/html/draft-ietf-manet-fsr-03
12. Liu, C., Kaiser J.: A Survey of Mobile Ad Hoc Network Routing Protocols. TR-4, MINEMA, TR-4, University of Magdeburg (2005).
13. Dijkstra, E.W.: A note on two problems in connection with graphs. Numerische Mathematik. 1, 269-271(1959)

Performance of the IEEE 802.15.4a access protocol under preamble-based clear channel assessment

Paúl Medina, Jaime Sánchez, José-Rosario Gallardo

Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE)
Carretera Ensenada-Tijuana No. 3918, Zona Playitas, Ensenada, B. C. México. C.P. 22860
pmedina@cicese.mx, jasan@cicese.mx, gallardo@site.uottawa.ca

Abstract. This work reviews the clear channel assessment methods available for the different physical layer options of IEEE 802.15.4 networks, which are used by the CSMA-CA access algorithm at the MAC sublayer in order to determine the channel availability. Then it focuses on the preamble-based channel assessment mechanism, used exclusively by the Time Hopping Ultra Wide Band physical layer, and analyzes the effectiveness of this mechanism in reaching the ultimate goal of the access algorithm, which is to avoid frame collisions. Owing to the fact that preamble-based clear channel assessment require of searching preamble symbols, the paper goes deep into the preamble format and synchronization algorithms.

Keywords: IEEE 802.15.4a, Time Hopping Ultra Wide Band (TH-UWB), Clear Channel Assessment (CCA), preamble sensing, Energy Detection (ED).

1 Introduction

In a typical IEEE 802.15.4 network, the devices transmit frames to their coordinator node, which is in charge of organizing the communications with its children nodes, among other responsibilities. Two kinds of communications are distinguished: beacon-enabled and non-beacon-enabled [1]. In the beacon-enabled option, the coordinator periodically transmits a beacon frame, which serves as a reference to create a superframe structure. The first portion of the superframe is assigned for random access and uses the slotted CSMA-CA protocol. Optionally, the coordinator can create a contention-free access section and an inactivity period within the same superframe. In the non-beacon-enabled option, there is only a random access mode, controlled by the unslotted CSMA-CA protocol, whose operation is strongly linked to the carrier sense mechanism used to determine the channel availability, this mechanism is also known as Clear Channel Assessment (CCA) [2].

All of the physical layer options considered in the original IEEE 802.15.4 standard consist of some kind of continuous-wave modulation, for which it is relatively easy to determine the channel availability. The original standard specifies three different CCA-options that indicate a busy medium whenever: 1) the energy in the band is above a fixed threshold, 2) there is a signal in the band compliant with the specifications, and 3) the combination of the two previous options. In the standard,

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 255-264

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

these options are named CCA-1, 2, and 3, respectively [1].

The IEEE 802.15.4a standard, released in 2007 [3] as an amendment to the original IEEE 802.15.4 specifications, adds two new options to the physical layer of these networks, one based on Chirp Spread Spectrum and another based on TH-UWB (Time Hopping Ultra Wide Band). With respect to the medium access control, the amendment keeps the same MAC sublayer but appends three new ways of CCA for the exclusive use of TH-UWB physical layer; namely CCA-4, 5, and 6. These new CCA options take into consideration the low duty cycle of TH-UWB signals, which hampers the common ways of performing the CCA in wireless environments.

In this work we study how these modifications in the CCA mechanism affect the performance of the MAC sublayer. Specifically, we are interested in investigating the performance of unslotted CSMA-CA protocol when the CCA-5 mechanism is used as the carrier-sense criterion at the physical layer. Important to mention that even though lot of research has been already done about the CCA and CSMA-CA, all of them rely on the continuous wave CCA. The originality of this work is to deal with the TH-UWB CCAs, which up to the authors' knowledge has not been explored yet.

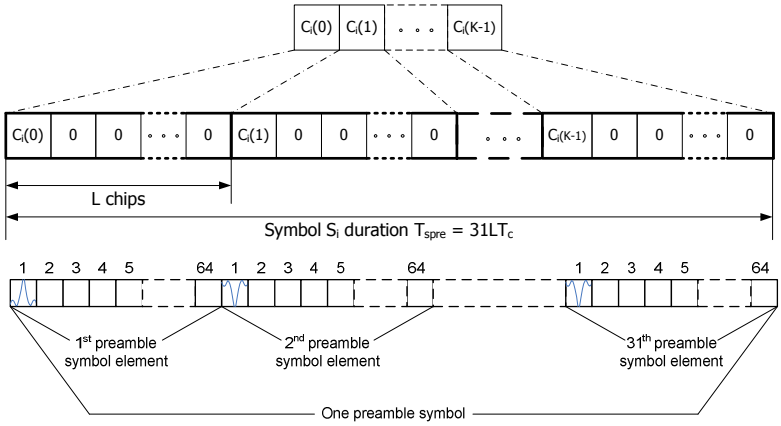
The organization of the paper is as follows. Section 2 describes the preamble format of the IEEE 802.15.4a specifications, then Section 3 describes the new CCA-options which are based on this preamble format. Later, Section 4 studies a representative scenario evaluated trough simulations, and finally Section 5 concludes the document.

2 IEEE 802.15.4a preamble format

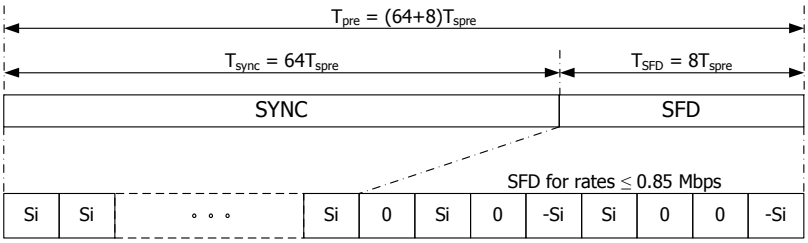
The IEEE 802.15.4a standard defines a physical layer based on pulses transmissions in the 0-10 GHz band. It specifies 16 FDM channels in this band, 12 of them are 500 MHz wide (-3dB bandwidth) and 4 of 1 GHz. The exact pulse shape is not defined, but it instead should fill some requirements in terms of the correlation with a reference pulse. For instance, in a 500 MHz channel the reference signal is a root raised cosine pulse with roll-off parameter $\beta = 0.6$ and 2 ns wide.

The preamble consists of N_{sync} repetitions of the preamble symbol S_i, where S_i is a sequence of pulses and blank spaces codified using a preamble code C_i. The standard specifies 8 preamble codes of 31 elements ($C_i, i = 1, 2, \ldots 8$), each of them consists of a sequence of ternary elements $c_{ij} \in \{-1, 0, 1\}$. These 8 codes were selected for its use in the preamble symbols of IR-UWB thanks to their properties of perfect autocorrelation in both coherent and non-coherent detection. They are also balanced sequences because all of them have almost the same amount of zero and non-zero elements (15 zero and 16 non-zero).

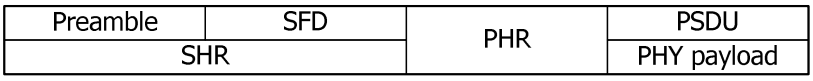
In Fig. 1a it is shown how a preamble symbol S_i is constructed from its corresponding preamble code C_i. It is observed that S_i constitutes an L length spreading of C_i, this means that after a pulse of polarity set by C_i (or the absence of the pulse in the case where $C_i(k) = 0$) there are inserted $L - 1$ empty chips. As an example the figure also shows a scheme of a preamble symbol of $L = 64$. Fig. 1b illustrates the synchronization header (SHR), which starts with a long sequence of repetitions of the preamble symbol and finishes with the start frame delimiter (SFD). The SFD is a predetermined sequence of full, inverted, or empty preamble symbols. Later, on Fig 1c, it is observed that after the SHR transmission follows the physical



a) Structure of the preamble symbol



b) Preamble format



c) Physical layer frame format

Fig. 1. Preamble.

layer header (PHR) of 19 bits length, followed by the payload (PSDU) of 127 bytes maximum length.

Due to the scope of this work, the format of the data symbols is not shown; however it is important to highlight two of the key differences between the data and preamble symbols: 1) the preamble does not use a time-hopping sequence, then the pulse position is not randomized as in the data symbols; 2) in the data symbols the information is transmitted in pulses bursts, in the preamble symbol the pulses are transmitted isolated. It is also important to emphasize that all the nodes use the same time-hopping sequence in the data symbols and the same code in the preamble, so that the multiple access relies on the carrier sense, the signal low duty cycle, and the non-synchronism in the start of frame transmissions.

3 The new clear channel assessment (CCA) options

As was mentioned in Section 1, the CCA is a critical component of the unslotted CSMA-CA algorithm, used as the access protocol by IEEE 802.15.4 networks. Although some ways of performing the CCA has been stated in the original standard, the TH-UWB physical layer, because of its impulsive nature, is unable to follow them, and then it has to specify its own CCA-strategies. Continuing the numbered sequence the new standard enumerates the new CCA options as CCA-4, 5, and 6. In the following paragraphs is described how these options work.

CCA-4 option consists of always indicating the availability of the wireless channel, turning the CSMA-CA access into ALOHA. The CCA options 5 and 6 try to decrease the collision among frames through the use of preamble sensing. Both of them are based on the fact that preamble and data symbols follow a different modulation format [3], with preamble symbols being easier to detect than data symbols when the receiver is not synchronized to the time-hopping sequence. In this case, data symbols seem more like random noise than preamble symbols because of the random position of pulses within the data symbols given by the time-hopping sequence being used.

The CCA-5 method consists of indicating a busy channel upon detection of a preamble symbol, and it continues indicating a busy channel for T_{max} seconds after the last preamble symbol has been detected, where T_{max} is the time necessary for the transmission of a maximum-length frame plus its acknowledgment. T_{max} only depends on the transmission scheme and is always the same value even if maximum-length frames are never transmitted and MAC frame acknowledgments are never used (frame acknowledgment is optional in all access schemes of the standard). Fig. 2 shows just an example that illustrates the behavior of CCA-5.

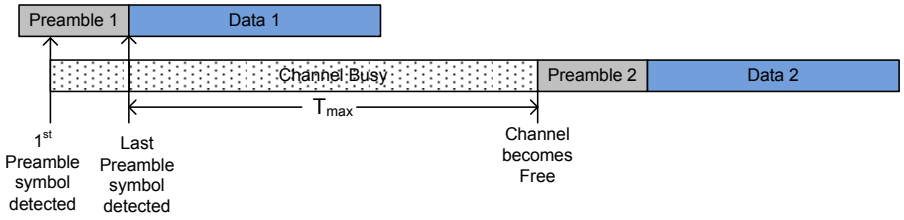


Fig. 2. CCA-5 operation example

The CCA-6 method, on the other hand, requires the insertion of preamble symbols in between the data symbols, enabling the neighbors to detect the current transmission in a similar way to that done by the continuous-carrier schemes. It is important to mention that the CCA-6 method is able to learn the channel availability without the long delays introduced by the CCA-5 method, at the cost of increasing the time required for a frame transmission.

In this work we study how these new CCA mechanisms affect the performance of the MAC sublayer. Specifically, we are interested in investigating the performance of unslotted CSMA-CA protocol when the CCA-5 mechanism is used as the carrier-sense criterion at the physical layer. Our interest in CCA-5 comes out because of, as a compensation of being unable to determine the channel availability immediately, it

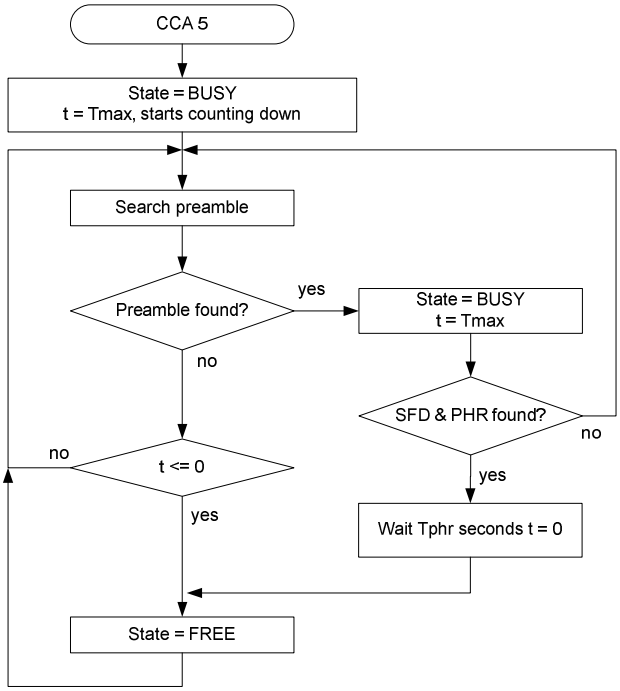
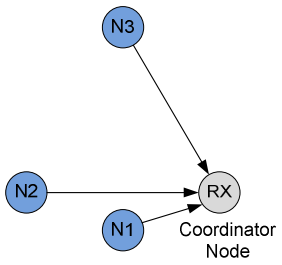


Fig. 3. CCA-5 Flowchart

has to introduce the use of a timer for the periods where it is not possible to perform a physical preamble sensing.

For a better understanding, Fig. 3 shows a flowchart of CCA-5 operation. When the MAC sublayer requests to the PHY layer to evaluate the medium availability with the CCA-5, it triggers the algorithm shown in Fig. 3, indicating a busy medium at the beginning and looking for preamble symbols over the air. It must pass T_{max} seconds without detecting any preamble symbol for changing the channel state to free. Once the channel is declared free it remains free until a preamble symbol is detected and then the state is changed to busy. Again it should pass T_{max} seconds with no preamble detected for the status to return back to free, and so on. It is observed in the flow chart that the T_{max} countdown can be shortened if the SFD is detected and the PHR is decoded. When that happens, the node can read the duration field contained in the PHR, so it can know when the frame will end and declare a free channel before T_{max} countdown concludes. This is known as 'virtual sensing' in IEEE 802.11 networks [5], however in TH-UWB it is done in the physical layer while in 802.11 it is performed at the MAC sublayer.



Relative link powers in dB

	N1	N2	N3	RX
N1	-	0	-3	0
N2	0	-	-3	-3
N3	-3	-3	-	-3
RX	0	-3	-3	-

Fig. 4. Scenario of study

4 Case of study and results

Fig. 4 shows the scenario considered for the performance study of the CCA-mechanisms. In such scenario there are one receptor node (RX) and three transmitter nodes (N1, N2, and N3). The traffic source of each transmitter node generates maximum length frames following a Poisson process with a rate of 100 frames/second. The transmitter nodes contend for the transmission of each frame following the access rules of unslotted CSMA-CA with the CCA-4 or 5. Observe that the distances between each pair of nodes are different and each link is received with different power. Fig. 4 also includes a table that shows the relative power in dB of each node link, where it is observed that:

- N1 signal reaches RX with more power (0 dB)
- N2 and N3 signals reach RX with the same power (-3 dB, the half of N1 signal)
- The power of link N1-N2 (0 dB) is higher than the link powers N1-N3 and N2-N3 (-3 dB both)
- Node RX could be seen as the coordinator node with which the children nodes (transmitters) have to communicate.

The scenario presented was evaluated through computer simulations in Matlab. It is considered that all the links use the data rate of 0.85 Mbps with a mean peak repetition frequency of 3.9 MHz, which means 4 pulses per data burst are used in the data symbol. For a detailed explanation about these parameters check [6]. Table 1 lists the main parameters used at the physical layer, where, for obvious reasons, is made emphasis in the preamble related parameters. The channel model used was the CM1 [7], which is an UWB channel model for indoors with line of sight.

At the receiver side, a non-coherent scheme based in energy detection is used, with an integration time of 8 ns [8]. The energy detector simply integrates over a period of time the square of the received signal. This scheme was adopted because of its simplicity, and also because it constitutes the most feasible low-cost scheme with the current technology. The synchronization method used was the one employed by [9] for this kind of reception, and the same synchronization parameters were also picked. Broadly speaking, in this method the energy samples are correlated with a binary mask constructed following the code preamble. The correlation samples are organized in vectors, where the number of elements in each of them corresponds to one preamble symbol. When the maximum value lies in the same position over several consecutive vectors the detection of a preamble symbol is declared. Later, only in the RX node, it is performed the fine synchronization algorithm, the looking for the start frame delimiter, the data and the Reed-Solomon decoding.

Figs. 5 and 6 show the main results of this research work for the explained scenario of study. For comparison purpose, both figures show at the left side the results when the CCA-5 is used as the channel sense mechanism, and in the right side the corresponding results when it is used the CCA-4, which means that the unslotted CSMA-CA access turns to ALOHA.

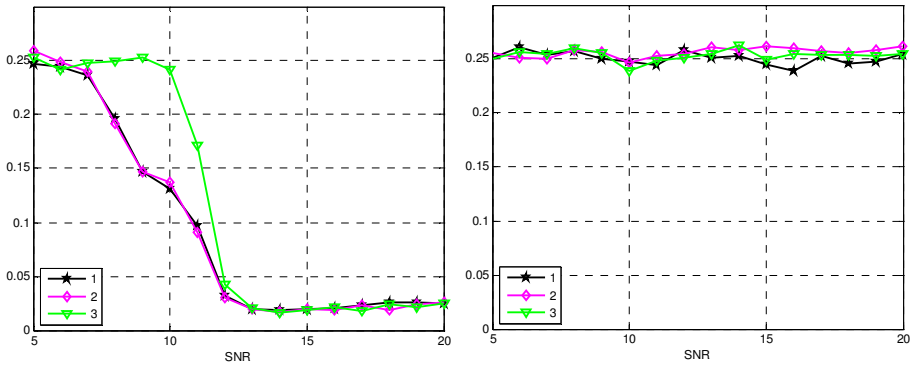
The figure of merit sketched in Fig. 5 is the Overlapping Frame Rate (OFR) per node, obtained for each transmitter node as the number of overlapped frames divided into the total transmitted frames:

$$\mathrm{OFR(node\ i)} = \frac{\textit{Overlapped Frames}(\mathrm{node\ i})}{\textit{Total Transmitted Frames}(\mathrm{node\ i})} \qquad (1)$$

Observe that we preferred the word 'Overlap' instead of 'Collision', due that as we will see later not all the frame overlaps turn into a destructive collision, as happens in the continuous wave transmission. Turning back our attention to Fig. 5a, we can see how the effectiveness of the CCA-5 in avoiding the frame overlap is proportional to the SNR with whom the pulses of the potential interferer are detected. It is also seen

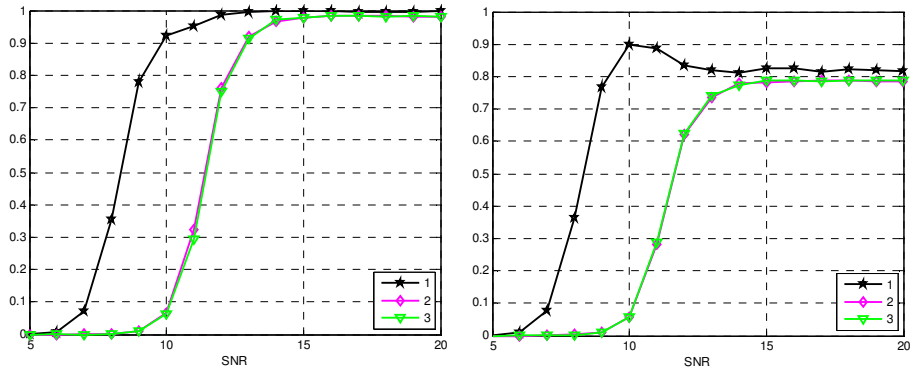
Table 1. Physical layer specifications.

Data rate	0.85 Mbps
Chip Period (Tc)	2 ns
Data Symbol Length	512 chip periods
Preamble code spreading (L)	64 chips
Preamble length	64 preamble symbols
SFD length	8 preamble symbols
Preamble code number	5th in all nodes
Channel model	CM1 [7]
Tmax	1.72 ms
Reed-Solomon decoding	yes
Receptor	Based on Energy Detection
Integration Time (Tint)	8 ns (4 chips)



a) Unslotted CSMA with CCA-5 **b)** Unslotted CSMA with CCA-4 = ALOHA

Fig. 5. Overlapping Frame Rate: Number of overlapped divided by total of transmitted frames, per node.



a) Unslotted CSMA with CCA-5 **b)** Unslotted CSMA with CCA-4 = ALOHA

Fig. 6. Successful synchronization rate in the node RX: Number of frames successfully synchronized divided by the total of transmitted frames, per node.

that transmitters N1 and N2 are more efficient in avoiding the frame overlap since the link N1-N2 is shorter than the links between these nodes and N3, then N3 requires higher SNR in order to reduce the overlapping. A floor level is observed in the overlapping, that comes from the vulnerability period of the CCA-5 in which it is impossible to detect a transmission that just begun.

Fig. 5b only confirms that in the CCA-4 nothing is done for avoid the overlapping (ALOHA), in fact they remains at the same level matter neither the SNR nor the node position. It is well-known that this level only depends on the network load and the frame size [10]. Notice as well that, when the SNR is not higher enough the CCA-5 cannot detect the current transmissions and has a similar behavior to CCA-4.

Fig. 6 sketches the Successful Synchronization Rate (SSR) for each node, obtained in the RX for each transmitting node making the following division:

$$\mathrm{SSR}(\text{node i}) = \frac{\textit{Successfull Syncronized Frames}(\text{node i})}{\textit{Total Transmitted Frames}(\text{node i})} \qquad (2)$$

It can be seen in Fig. 6 that, in both CCA schemes, node N1 gets a higher synchronization probability because of it is closer to the RX node than N2 and N3, and those nodes have exactly the same trend because their links to RX have exactly the same distance. As we previously mentioned, in Fig. 6a can be observed that not all overlapping is destructive, because the synchronization efficiency is close to one; even though, as seen in Fig. 5a, overlapping is not completely eliminated by CCA-5. Fig. 6a shows that CCA-4 cannot reach the synchronization rate achieved by CCA-5. Again, the level shown is surely dependent on the network load.

It is noticeable that the figures of merit are both dependent on the received signal strength, however the overlapping is dependent on how transmitter 'hears' to the nodes it is contending with, and synchronization depends in how the intended receiver 'hears' the transmitter.

5 Conclusions

The performance of the access protocol with the CCA-5 is evaluated and compared to that in which the CCA-4 is used (i.e. no carrier sense), both when energy detection receivers are employed. As expected, the use of CCA-5 brings a better performance to the access protocol, which can be seen in the plots of overlapping and synchronization rates against SNR (signal-to-noise ratio) achieved by each channel assessment method.

The effectiveness on CCA-5 in the avoidance of frame collision was found superior to CCA-4, however the level of implementation penalties introduced at the physical layer have not yet been analyzed. It was observed in both schemes a kind of capture effect, where in overlapped transmissions the receptor synchronizes with one of the frames. Just like in phase modulation of continuous carrier, in TH-UWB the receiver also tends to synchronize with the strongest signal. However, unlike the narrowband schemes, in TH-UWB if the receiver is already synchronized to a weak signal it will never synchronize to a stronger signal that comes after.

At this point of our research we made our discussion based on the results obtained by computer simulations of different network scenarios. A mathematical analysis that gives us a better understanding of the system behavior is our next step in the research path.

References

1. IEEE 802.15.4-2006. "Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs)." IEEE Standard 802.15.4, 2006 Edition.

2. I. Ramachandran, S. Roy, "Clear Channel Assessment in Energy-constrained Wideband Wireless Networks," IEEE Wireless Communications Magazine, 2007, Vol. 14, No.
3. IEEE 802.15.4a-2007. "Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs). Amendment 1: Add alternate PHYs." IEEE Standard 802.15.4a, 2007 Edition.
4. F. A. Molish, P. Orlik, Z. Sahinoglu, J. Zhang, "UWB-based sensor networks and the IEEE 802.15.4a standard – a tutorial," First International Conference on Communications and Networking in China, 2006 (ChinaCom '06).
5. Gast, M. 2002. "802.11 Wireless Networks: The Definitive Guide". O´Reilly. First Ed. Sebastopol, California.
6. P. Medina, J. R. Gallardo, J. Sánchez, F. Ramírez Mireles, "Impact of Delay Spread on IEEE 802.15.4a Networks with Energy Detection Receivers", Journal of Applied Research and Technology, Vol. 8, No. 3., 2011.
7. Molish F. A., Balakrishnan K., Cassioli D., Chong C., Emami S., Fort A., Karedal J., Kunisch J., Schantz H., Schuster U. & Siwiak K., "IEEE 802.15.4a channel model – final report," 2004.
8. H. Urkowitz, "Energy detection of unknown deterministic signals," Proc. of IEEE, vol. 55, pp. 523-531, Abril 1967.
9. M. Flury, R. Merz, J.-Y. Le Boudec & J. Zory, Performance Evaluation of an IEEE 802.15.4a Physical Layer with Energy Detection and Multi-User Interference, IEEE International Conference on Ultra-Wideband (ICUWB 2007), September 2007.
10. N. Abramson, "Development of the AlohaNet," IEEE Trans. Info. Theory, vol. IT-31, no. 2, Mar. 1985, pp. 119–23

Real-Time Communication Protocol for Broadcast Based Networks

Carlos Franco[1], Luis Gutierrez[1], Raul Jacinto[2]

[1] Universidad de Guadalajara, CUCEA, Periférico Norte 799, C.P.44140,
Zapopan, Jalisco, Mexico
[2] DAVTI, Francisco Rojas Gonzalez 232, C.P.44650,
Guadalajara, Jalisco, Mexico
carlos.franco@cucea.udg.mx, luis.gutierrez@redudg.udg.mx, raul.jacinto@gmail.com

Abstract. One of the main challenges in the design of real-time distributed systems is the definition of the communications scheme that is required to perform all system tasks in the network in such a way that all time constraints are met. In this work is presented a real-time communications scheme based in broadcast networks. This proposal is the result of the study of other schemes such as Token bus, FIP, Profibus and CAN. It is presented a mechanism for media access control based on arbitrated message contention according to priority, which is given by a master plan. The proposed scheme considers periodic and aperiodic message delivery with static schedule and is based on common characteristics found in hard real-time systems (HRTS), includes a closed task set with time constraints, where critical tasks are defined as periodic tasks and it takes advantage of the broadcast nature of most networks found in real-time distributed systems. This work includes also a simulation scheme of the proposal.

Keywords: Real-time systems, real-time communication, broadcast networks, arbitrated message contention, message scheduling.

1 Introduction

Broadcast networks are present in almost all today's network environments, and bus topology [1] is widely used because of its low cost and ease of administration.
In the literature [7][9][11] are discussed several factors that contribute on the delay of message delivery in the communication process: queuing, packeting, switching and propagation. These factors are present in different stages of message transmission. There are in particular, six delay moments, which are represented in figure 1. These delays, which are identified as d1, d2, … d6, are presented in the different OSI model layers.

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 265-275

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

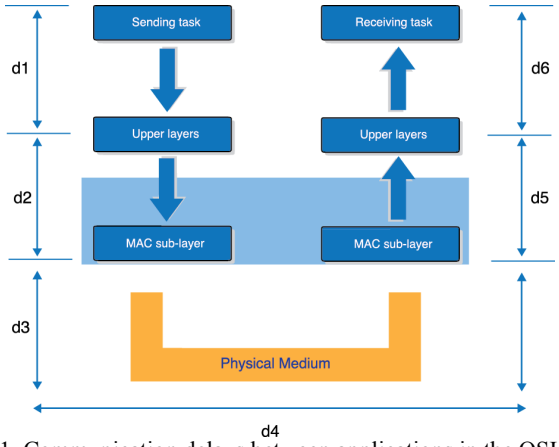


Figure 1. Communication delays between applications in the OSI model.

Delays d1 and d6 are presented when communication process is taking place in the upper layers of OSI model (layers 3 to 7). This is when the sending task and the receiving tasks are executed. Delays d2 and d5 are generated within layer 2 and generally are due to physical medium access control; however, d5 is significantly shorter than d2 because in the receiving task, there is no contention actually for physical medium. Delay d3 emerges when data is put into the physical medium and it is considered a queuing delay. Delay d4 is due message propagating in physical medium. As we can see, d2 is the hardest delay to deal with (it occurs in the medium access MAC sub-layer) because it is required to develop admission control mechanisms and packet scheduling schemes. Additionally, some techniques and communication models allow to shape traffic and to evaluate quality of service (QoS) requirements [5] for a particular application.

2 Problem Definition

As it is common in several hard real-time systems, execution plan is known in advance. In centralized or single-processor environments, the execution depends on a single entity -a dispatcher or a network referee- that defines which task is executed next. In other environments, such as Profibus [2], the execution control is distributed, where a token grants access to the network to its possessor. Both approaches have their own advantages but also disadvantages [6].

On one side, for the approach that uses a token, real-time execution can only be guaranteed to the node that holds token. On the other hand, the FIB approach, based in a bus referee turns out to be a very rigid scheme. CAN [3] is also based in the node priority, not in the priority of the task, situation that can lead to a problem in the real-time execution of the system. In this paper, an execution scheme based on distribution of control in the whole network is presented.

3 General Description of the Communications Scheme

It is assumed that there exist a closed number of participant nodes and there is already a master execution plan, where all tasks have been assigned to their corresponding processing entity. This plan is feasible and all deadlines are met. Each node has an instance of the global execution plan and it is assumed that each node has available all resources required to perform all the assigned tasks.
Assigned tasks to each node are not necessarily communication tasks in all cases, so it is possible to schedule tasks that do not require delivery of messages. It is very important to identify communication tasks from the others. Trough broadcast we can guarantee a minimum synchronization level between all participant nodes, because each sent message would be known and listened by the entire network. Global plan allows each node to know when a message is going to be sent, the order of sent messages and therefore, identify when and which message send each time.
Regarding to the transmitting node, when it sends a message, this is received by all network members (broadcast) but only the node that the message is directed to will process it. In that moment, as the received message has the address of the next node allowed for transmitting, it is assured an accurate synchronization in messages delivery. The authorized node sends its message according to the described procedure, which is repeated until the execution plan is completed. Figure 2 presents a general view of the communications network.

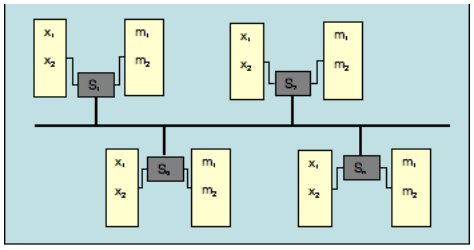


Figure 2. General view of the communications network

In the figure can be seen: the instance of the global plan, represented by the set X, the set of participant nodes $S = \{s_1, s_2, ...s_n\}$ and the schedulable set of messages assigned to each node, denoted by $M_{(S_i)} = \{m_1, m_2, ...m_n\}$. The proposed communications scheme allows delivery of periodic and aperiodic messages [8] with static schedule [10].

4 Protocol Description

4.1 Network topology

It is considered a bus network topology, where medium access is controlled with the implementation of a "logical ring" within the network to avoid collisions. The ring will work with a circulating token that will become the transmitting permission, so each node will send all required messages. Only one node will have the token at a time, so only one node can transmit at the same time. With this strategy, collisions are avoided and it is guaranteed that messages are delivered according to the schedule previously established.

4.2 Medium access control

Access to medium is serialized because the message delivery is developed according to global plan. When network is initializing, one node will randomly be designated as the master node and will verify the following:

- All nodes must have been assigned with all necessary tasks for the operation of the system
- Identify the sequence of messages to be delivered among all nodes
- Circulate a start message for the network initialization, in such a way that all nodes have an instance of the schedule, the messages and tasks assignment within the system
- Create the initial token for the network

Once the initialization token is circulated, each and every node will know which tasks must perform, what messages is going to receive, what messages are required to be sent and at what time these actions are going to take place. Therefore, the network will have a pre-established token circulation, and there will be guarantees that system feasibility can be accomplished.
At this point, all nodes have a copy of the schedule that is going to be performed and they must identify if there is aperiodic traffic to transmit.
Then, the master node sends the first token to initiate the normal execution of the system. This token will consider that in this moment, the highest priority for sending aperiodic traffic is for the master node. This is only for the initialization of the network operation.

4.3 Periodic and Aperiodic Traffic

There exist a set of periodic messages, a set of aperiodic messages (represented both by communication tasks) and other tasks that are not communication tasks. Periodic messages have very strict time constraints because they represent critical communication messages (real-time) whereas aperiodic messages have more relaxed time constraints because they do not represent critical communication tasks.

All messages have fixed length and the same structure: message ID, data field and next station to transmit node ID (token).

4.4 Message Scheduling

Periodic messages will be sent according to system's global plan and aperiodic messages will be locally sorted in every node according to its deadline. A message with closest deadline will have higher priority compared to a message with a later deadline. With this, messages will be delivered in the right order and system requirements will be satisfied. Provided the fact that broadcast based technology [4] (Ethernet, for example) is well known in their data propagation times and that now it already has a medium access control mechanism that avoids collisions, it is possible to accurately calculate if the delivery of a set of messages is schedulable in the network according to its deadline.

5 Formalization and Evaluation of the Protocol

The proposed protocol requires certain initial conditions for its operation:
C1: It is assumed a closed set of sites or participant nodes S
C2: Each node is assigned a set M of nodes and tasks that is schedulable
C3: Each node has a copy of the global message schedule, known as execution plan X.
C4: Messages have fixed length and a three-field structure: message identifier, data field, and next transmitting station or node identifier.
C5: Message transmission time δ is negligible
C6: All nodes receive the same message at the same time, including the origin node.

Let $S = \{s_1, s_2, ...s_n\}$ be the set of sites or participant nodes, $M_{(S_i)} = \{m_1, m_2, ...m_n\}$ the set of schedulable messages assigned to node S_i, such as $\forall m_j \in M$, we have that $m_j = (i, data, token)$, where i is the index or identifier of the message in the execution plan X, [redacted] represents the information that is going to be transmitted and $token$ represents the permission for the next node can actually transmit a message.
$X = \{x_1, x_2, ...x_n\}$ is the set of messages that conform the execution plan, where $\forall x_i \in X$, we have that $x_i = (s_{origen}, m_j, s_{destino})$ where s_{origen} is the node that sends the message m_j to node $s_{destino}$.
We have that $\forall s_n \in S, \exists p_n$, such as p_n is a local administrator that manages communication activities in each node and is responsible to send, receive and deliver a message m_j from node S_{origen} to node $S_{destino}$ by using the functions $create_i(m_j)$, $send_i(m_j)$, $receive_i(m_j)$ and $deliver_i(m_j)$.
Functions $create_i(m_j)$ and $deliver_i(m_j)$, are higher lever functions than sending or receiving the message m_j at the nodes s_{origen} or $s_{destino}$. Reception of a message does not imply the immediate delivery of the message because delivery is conditioned by

timing parameters.

$send_i(m_j)$ is the function that assures the sending of the message m_j from node s_{origen} to node $S_{destino}$

$receive_i(m_j)$ is the function that assures the reception of the message m_j from transmission medium to node $S_{destino}$.

$create_i(m_j)$ is the function that communicates to the application level to generate messages that eventually are going to be send to the communication medium.

$deliver_i(m_j)$ is the function that assures the delivery of the message m_j to the node $S_{destino}$.

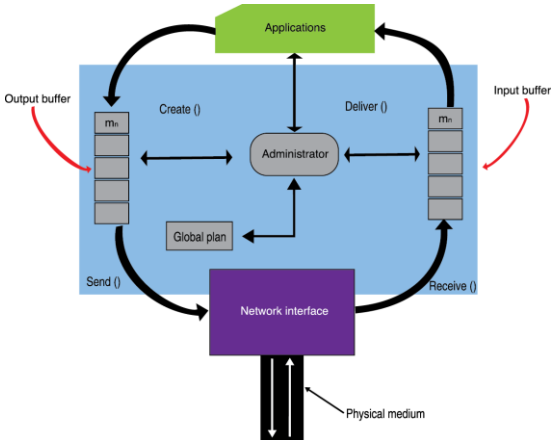


Figure 3. Architecture of a communication node within the network.

It is necessary to perform a simulation process in order to verify the average arrival time of messages in several executions of the global plan and also to verify that there is no contention for the physical medium.

Next, are presented the considerations taken into account for the simulation:

1. It is assumed that there is a closed set of participant nodes
2. There exists a previous fixed schedule for periodic messages
3. It is assumed that the periodic messages set that is going to be send is feasible
4. Every node that comprises the network has a copy of the execution plan (to avoid medium access contention) and therefore, knows which are the messages to send.
5. Nodes do not require an explicit synchronization, because having a copy of the global schedule is an intrinsic mechanism of synchronization.
6. Messages are sent according to the sequence established in the execution plan.
7. In each message, a token is transmitted. If the token is released, the next site or node in the list can send its message.
8. The message has a three field structure:
    a. Message identifier
    b. Data

c. Token

The token acts as an identifier for the next node to transmit.

Based on the simulation previously described, it is expected:

1. Know the average arrival time of messages in every node
2. Verify that no exists contention for the medium.

For the simulation of aperiodic tasks there are included two proposals, the first (proposal 1) consists in applying the criteria that the node that has the token, is the node that send the aperiodic traffic and only in case this node has no aperiodic messages to send, the token goes to the next node according to global plan.
The second (proposal 2) applies the criteria of circulating the token (to send aperiodic traffic) in a predefined order (which is defined for example at the moment of the network initialization). It is convenient to remark that aperiodic traffic will be sent only after sending all periodic traffic (real-time traffic) trying to take advantage of network idle times.

6 Results

For the simulation, it was performed a work that consisted in the evaluation and the comparison of a couple of simulation scenarios described in the previous section. In order to perform those comparisons it was required to design a simulation environment that could allow creating the network conditions to simulate. This implied the development of the following activities:

1. Determine the number of processors in the network
2. Generate a set of schedulable tasks for each node in the network, with the processing load in each task chosen for each node.
3. Determine the load of periodic messages present in each node.
4. Generate a global schedulable plan that includes the established communication tasks
5. Determine aperiodic messages load present in each node
6. Simulate the behavior of the communications network
7. Get the results
8. Interpret the results
9. Conclusions

6.1 Number of processors.

Simulation environments were developed with 5 and 10 processors. In this work are only presented simulation results with 5 processors because results obtained with 10 processors are very similar in both cases.

6.2 Generation of a set of schedulable tasks.

There were created random sets of tasks for each of the network nodes, which are completely schedulable in each one of the processors. The simulation presents two cases, when the processor load is 60% and when it is 80%. This means that in each node we had a maximum load of tasks for each processor equivalent to 60% or 80% and that set of tasks was schedulable.

6.3 Determination of the periodic load in each participant node.

From the total number of tasks assigned to each node, not all of them are communication tasks, that is, some tasks that do not require communicating with other nodes and do not require information exchange to perform correctly. This means that in the simulation scenarios it is required to define from the total number of tasks or messages assigned to each node, how many messages are periodic. For this simulation, it was considered that 10% of the tasks assigned to each node are communication tasks.

6.4 Generation of an execution plan in each node: Global plan.

As each set of tasks is schedulable in each node, now it is required that all sets of tasks are schedulable when they combine. This is, if the sets of tasks are schedulable in the local environments, there is no guarantee that all local plans can be schedulable when it comes about a global schedule. The simulation scenario then creates local plans that are also schedulable in the global level. The simulation performed did not detect any problem regarding to collisions or missed deadlines in the system messages or tasks.

6.5 Aperiodic messages load.

From the total of communication tasks present in each node, most are periodic tasks, but there were considered different levels of aperiodic tasks load, to analyze the behavior of the network to these variations.
For the task processing load of 60%, it was simulated a message load of 10% and an aperiodic messages load of 5%, 10%, 15% 20% and 25%.
For the task processing load of 80%, it was simulated a message load of 10% and an aperiodic messages load of 5%, 10%, 15% 20% and 25%.

6.6 Simulate behavior of the communications network.

To simulate the behavior of the communications network, it is required to clear up that it was considered a 2 time units delay for the minimal transmission and propagation time (communication time between two nodes, the closest) and a 5 time units delay for the maximum transmission and propagation time (communication time

between two nodes, the farthest). Tasks will have established periods between 100, 50 and 40 units. The execution units, the period and deadlines of each task or message are generated randomly. Precedence relationships are established randomly and after processes have been generated for each processor.

6.7 Obtained results.

Obtained results have to do with the communications network behavior and consist in the following:

- Response time for periodic messages
- Response time for aperiodic tasks

Results are shown in tables 1 and 2.

Task load: 60% of processor capacity					
Periodic messages load: 10%					
Aperiodic message load	**5%**	**10%**	**15%**	**20%**	**25%**
Proposal 1	2.96	3.1	3.52	3.67	3.98
Proposal 2	2.8	2.96	3.41	3.56	3.78

Table1. Simulation results for 60% of tasks load

Task load: 80% of processor capacity					
Periodic messages load: 10%					
Aperiodic message load	**5%**	**10%**	**15%**	**20%**	**25%**
Proposal 1	2.56	3.01	3.38	3.71	3.95
Proposal 2	2.52	2.91	3.31	3.66	3.92

Table2. Simulation results for 80% of tasks load

The response time for periodic messages is constant, because its attention is not in risk because it is assigned in fixed intervals. Results were obtained in the format in what the simulation tool delivers them, however it was necessary to treat them so they could be interpreted and plotted.

6.8 Interpretation of results.

As it can be seen in tables 1 and 2, there is a slight variation in the response times for the delivery of aperiodic traffic when the load percentage of processor capacity is increased in each node. This can also be noticed in figures 4 and 5.

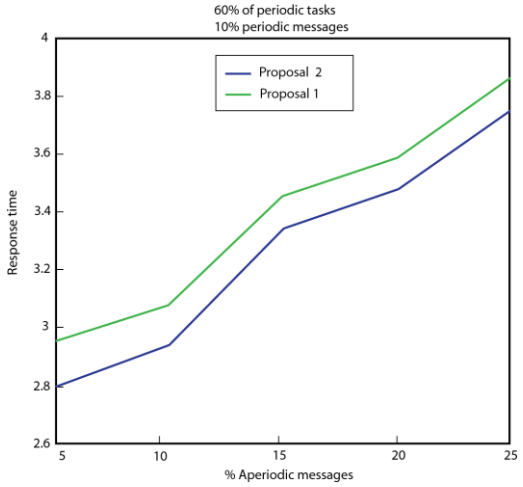


Figure 4. Comparing response time for aperiodic messages with 60% of processor load.

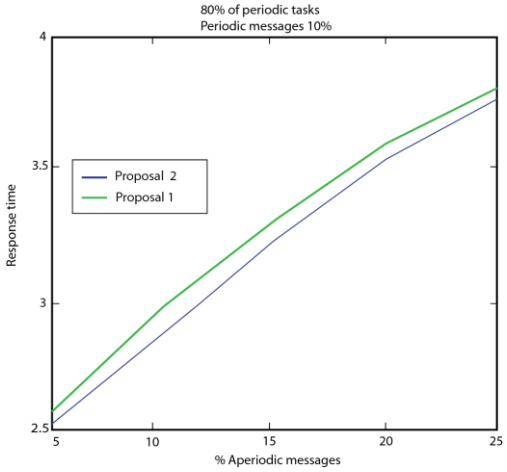


Figure 5. Comparing response time for aperiodic messages with 80% of processor load.

In figure 4, we can appreciate the behavior of the average propagation and transmission times when tasks load is 60% of the processors capacity. In figure 5, we can appreciate the behavior of the average propagation and transmission times when tasks load is 80% of the processors capacity. In both cases are considered the two proposals described previously for dealing with aperiodic traffic.

6.9 Conclusions

According to behavior presented in graphics and tables, we can conclude that the first proposal offers better response times. However, as the processor load is increasing, the second proposal tends to a more stable behavior and to offer similar response times, whereas the firs proposal as the processor load is increasing tends to increase also its response times. This behavior is consistent with the idea that the first scheme

is unfair under the perspective of how many nodes can send aperiodic traffic. That is the reason why it has better performance when there is not much load in the processor. Nevertheless, the second scheme, as it is more balanced (fair) it tends to improve the performance when processor load increases.

The first proposal is more efficient for cases when processor load is not so demanding and in cases where aperiodic messages is kept within the 2%, referred in the literature by Stankovic [12]. The second proposal is better for cases where aperiodic messages load is above average. Even this scenario is hard to find in practical situations, it is interesting for analysis. According to these results, it could be evaluated a third proposal that have a medium between both presented proposals regarding the amount of aperiodic messages it can handle and possibly it can be obtained a better scenario between response times and stability.

References

1. Weaver and C. Summers, “The IEEE Token Bus-A Performance Bound on GM MAP” IEEE transactions on Industrial Electronics, Volume 35, Issue 1, feb 1988.
2. E. Tovar and F. Vasques. “Real-time fieldbus communications using Profibus networks”. IEEE Transactions on Industrial Electronics. Volume 46, Issue 6, Dec 1999 Page(s): 1241 – 1251.
3. H. Kaschel, E. Pinto. “Análisis protocolar del bus de campo CAN” Reporte de trabajo. Facultad de Ingeniería, Depto. de Ingeniería Eléctrica. Universidad de Santiago de Chile. Chile, 2002.
4. G. LeLann and N. Rivierre, “Real-Time communications over broadcast networks: The CSMA-DCR and the DOD-CSMA-CD Protocols”. INRIA Report RR1863, 1993.
5. P.Ferguson and G. Huston. “Quality of Service: delivering QoS on the Internet and in corporate networks”. John Wiley & Sons press. USA, 2000.
6. C. Aras, J. Kurose, D. Reeves and H. Schulzrinne, "Real-Time Communication in Packet-Switched Networks", Proceedings of the IEEE, Vol. 82, No. 1, pp. 122-139. Enero, 1994.
7. Cottet, Francis, Joêlle Delacroix, Claude Kaiser, & Mammeri, Zoubir, “Scheduling in real-time systems”. England: John Wiley and sons press. USA, 2002.
8. Y. Atif and B. Hamidzadeh, “A Scalable Scheduling Algorithm for Real-Time Distributed Systems”, Proceedings of the 18th International Conference on Distributed Computing Systems, May 26-29 1998, pp. 352-359
9. D. Verma, H. Zhang and D. Ferrari. “Delay jitter control for Real-Time communication in a packet switching network”. In proceedings of TriComm. 1991.
10. Buttazzo, Giorgio C. “Hard real-time computing systems: predictable scheduling algorithms and applications”. Kluwer Academic Publisher. Boston, 1997.
11. Cheng, Albert M.K. “Real-time systems: Scheduling, analysis, and verification”. John Wiley and sons press. New Jersey, 2002.
12. Stankovic, John A., Spuri, Marco, Ramamritham, Krithi, & Buttazo, Giorgio. “Deadline scheduling for real-time systems: EDF and related algorithms”. Kluwer Academic Publisher Boston, 1998.

Design of Multi-beam Circular Antenna Arrays using CORPS and Differential Evolution

Marco A Panduro, Aldo Mendez, and Leopoldo Garza,

Unidad Académica Multidisciplinaria Reynosa-Rodhe, Universidad Autónoma de Tamaulipas (UAT) Carretera Reynosa-San Fernando, Reynosa, Tamaulipas, 88779 México
Phone: (52) 899.921.3300, Fax: (52) 899.921.3301,
E-mail: [mamendoza, almendez]@uat.edu.mx; pandurom@yahoo.com

Abstract. An innovative way to analyze the design of beam-forming networks (BFN) for scannable multi-beam circular antenna arrays using the CORPS (Coherently Radiating Periodic Structures) concept is introduced. This design of CORPS-BFN considers the optimization of the complex inputs of the feeding network by using the Differential Evolution (DE) method. Simulation results for different configurations of CORPS-BFN for a scannable circular array are presented. The results shown in this paper illustrate certain interesting characteristics in the behavior of the array factor for the scannable circular array. The most significant aspect that is unique to this proposal is the simplification of the feeding network based on CORPS

Keywords: Coherently radiating periodic structures, circular antenna array, differential evolution method.

1. Introduction

The flexibility and re-configurability are two important features of the present and future antenna systems. These two properties could be very easily defined combining smartly different independent beams or signals of the same antenna system. These systems, which are capable of managing independently different beams, are usually referred to as multi-beam systems.

The CORPS [1]-[5] concept starts from the idea to reproduce the behavior of the human eye and to apply its detecting strategies to the antenna field. The main idea behind CORPS is to try to find solutions to the common trade-offs in antenna arrays systems design, such as angular resolution, signal-to-noise ratio, coupling and grating lobes. For instance, one of the most interesting procedures used by the human eye to obtain the information corresponding with every spot or pixel of the image is the fact that this information is received effectively for many cones thanks to the coherent coupling that exists between the photo-detectors. The procedure is performed by small chained chemical reactions firstly originated by the excitation of one of the photo-detectors. This coupling generates high overlapped radiation zones, making the human eye capable to generate simultaneously high directive beams very close in the angular space, obtaining an impressive resolution.

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 276-285

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

In our case, it will introduce an innovative way to analyze the design of beam-forming networks (BFN) for scannable multi-beam circular antenna arrays using the CORPS concept. The main objective of this paper is to combine the technology based on CORPS to define the BFN and the Differential Evolution (DE) [6]-[11] method to look for optimal excitations, in order to generate a scannable multi-beam circular array. The contribution of this paper is to present a perspective of the design of CORPS-BFN considering scannable multi-beam circular arrays.

2. Behavior of the CORPS-BFN

A schematic representation of a CORPS-BFN of *n* inputs, *N* outputs and 3 layers is presented in Figure 1. As shown in Fig. 1, a CORPS-BFN is conformed by a mesh interconnected by means of Split (S)-nodes and Recombination (R)-nodes. The CORPS-BFN works as follows. The signal entered by one input port is divided in two and added with the arriving signals of the neighboring input ports. Following the path of each signal, we will find something like an inverted triangle which has the lower vertex at the input port. The opposite side of this vertex will define the output ports receiving some information from this input port, or in other words, the effective radiating area from which every input signal (or orthogonal beam) will be radiated. Since the isolation between the input ports is ensured and the spreading of the signal inside the structure is controlled, the CORPS-BFN is able to handle simultaneously several orthogonal beams without any problem. In the outermost branches, the inputs that are not used are finished with a matched load in order to avoid reflections.

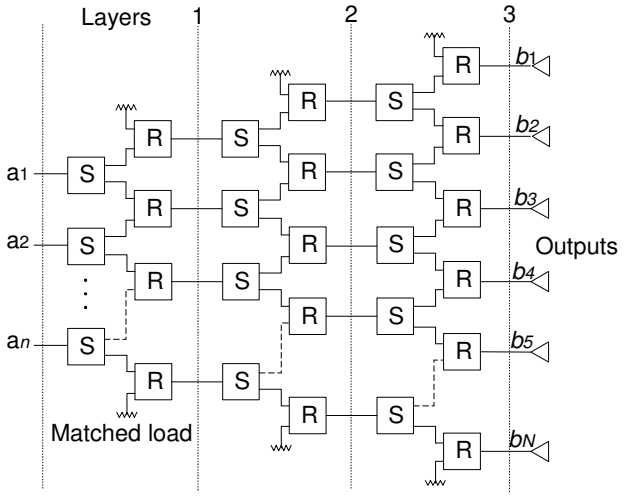


Figure 1. Schematic representation of a CORPS-BFN with S and R nodes

From [1] and [3] the Unitary Cell Scattering matrix that represents the behavior of an S-node could be extracted as follows:

$$[S] = \begin{bmatrix} 0 & j/\sqrt{2} & j/\sqrt{2} \\ j/\sqrt{2} & 0 & 0 \\ j/\sqrt{2} & 0 & 0 \end{bmatrix} \tag{1}$$

It is also shown in this paper that an S-node can act also like an R-node. In the same way, in order to evaluate the fields after an S-node or R-node the next expression can be used

$$V^- = S\, V^+ . \tag{2}$$

In (2) *S* is the Scattering Matrix of an S-node and V^+ is the Amplitude and Phase of the field at input ports of an S-node. Using (2) and the schematic representation of a CORPS-BFN (Fig. 1) it is possible to establish an iterative code (i.e. with MATLAB) that represents the propagation of signal throughout a general configured CORPS-BFN.

It is possible to establish different configurations for the CORPS-BFN with different number of inputs, outputs and layers. In this case, several orthogonal beams could be generated simultaneously by intercalating or interleaving the inputs of the CORPS-BFN, i.e., a group of different inputs will generate the beam # 1 and another group of inputs could generate the beam # 2. Following the philosophy of CORPS, each group of inputs must be established in a strategic way in order to have the capability to control electronically the corresponding beam pattern (over a scanning range) with a smaller number of complex inputs with respect to the number of antenna elements employed. Several configurations for the CORPS-BFN could be evaluated and studied. To set an example, the next configurations could be of interest.

1) For a system of 26 radiators and 25 input ports (i.e., a CORPS-BFN of one layer) two orthogonal beams could be generated simultaneously by intercalating the inputs of the CORPS-BFN, as shown in the Fig. 2. The interesting aspect of this case is that the group of 13 inputs (that generates the beam # 1) could control the 26 radiators of the array, and the remaining 12 (used for the beam # 2) could control to 24 of them.

2) For the system of 26 radiators, we could use a CORPS-BFN of two layers with 24 input ports. Two orthogonal beams could be generated simultaneously by intercalating the inputs of the CORPS-BFN by pairs as illustrated in the Figure 3. In the case of the beam # 1, 12 of 24 input ports could control 24 radiators of the array. For the beam # 2, the remaining 12 input ports control to 24 of 26 radiators.

For a set of complex inputs [**a**] feeding the CORPS-BFN, as shown in Fig. 1, the characteristics of Directivity (*D*) and Side Lobe Level (*SLL*) for each beam pattern can be calculated using the equation of the array factor as [12], [8]

$$AF(\phi, \mathbf{a}) = \sum_{n=1}^{N} b_n \exp\left[jkr\left(\cos\left(\phi - \Delta\phi_n\right) - \cos\left(\phi_0 - \Delta\phi_n\right)\right)\right] \tag{3}$$

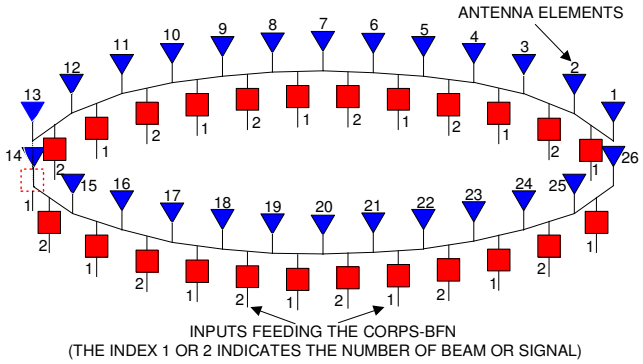


Figure 2. System of 26 radiators and 25 input ports.

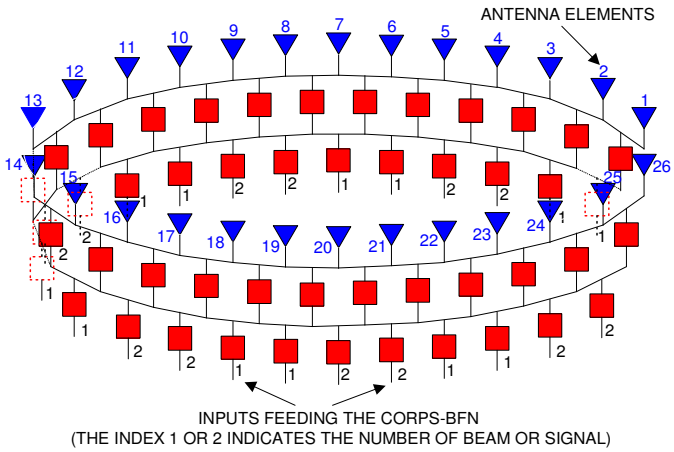


Figure 3. System of 26 radiators and 24 input ports

where b_i represents the complex excitation of the ith antenna element of the array, $\Delta\phi_n=2\pi(n\text{-}1)/N$ for n=1,2, …, N is the angular position of the nth element on the x-y plane, $kr=Nd$, i.e., $r=Nd\lambda/2\pi$, ϕ_0 is the direction of maximum radiation and ϕ is the angle of incidence of the plane wave.

In order to include the effect of mutual coupling for the circular array, the method of induced electro-motive force (EMF) [12] for thin and finite dipoles is considered. In this case, it is considered the side-by-side configuration and dipole lengths $l=\lambda/2$.

Next, the objective function and the evolutionary optimization technique used to optimize the complex inputs of the CORPS-BFN are described.

3. Objective function and the technique used

One of the latest evolutionary computational techniques is the DE algorithm, in which some individuals are randomly extracted from the solution population and geometrically manipulated [6], avoiding the destructive mutation of Genetic Algorithms (GA) [13]-[18]. The main advantage of DE is its low computation time compared to that of GA. DE is an alternative to speed up the GA.

First an initial population is formed in which the individuals have a Gaussian distribution. For each vector or solution (amplitude and phase of the complex inputs feeding the CORPS-BFN) of the population (N_p) X_i, i=1, 2, …, N_p of the G_{th} iteration, two new trial members, ε_{t1} and ε_{t2}, are generated as follows:

$$\varepsilon_{t1} = \varepsilon_{r1}^{(G)} + F\left(X_i^{(G)} - \varepsilon_{r2}^{(G)} \right) \tag{4}$$

$$\varepsilon_{t2} = \varepsilon_{r1}^{(G)} + F\left(X_i^{(G)} - \varepsilon_{r3}^{(G)} \right) \tag{5}$$

where $F \in [0, 2]$ is a real constant factor range suggested in [9], which controls the amplification of the differential variation, and the integers r_1, r_2, $r_3 \in [1, N_p]$ are randomly chosen such that $r_1 \neq r_2 \neq r_3$.

In this case each individual generates an array factor of certain characteristics of *SLL* and *D*. Therefore, the design problem is formulated as minimize the next objective function

$$Obj\text{-}fun=(|AF(\phi_{SLL}, \mathbf{a})| / max|AF(\phi, \mathbf{a})|)+(1/D(\phi, \mathbf{a})) \tag{6}$$

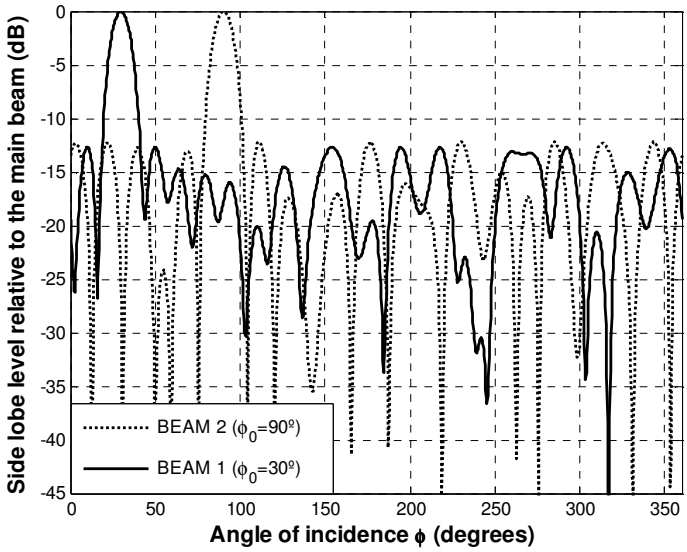
where ϕ_{SLL} is the angle where the maximum side lobe is attained. In this case both objectives (*SLL* and *D*) are uniformly weighted in the cost function.

After the objective function evaluation, the best solution in the set $\{\varepsilon_i, \varepsilon_{t1}, \varepsilon_{t2}\}$ becomes the new member for the next iteration, ε_i^{G+1}. Some individuals in the new population occasionally generate array factors which are not physically realizable, and an adjusting process is needed [7]. Taking the best solution into account, a termination criterion is proposed by fixing a number of iterations without an improvement over this solution. In [7], it is explained the procedure of DE in detail.

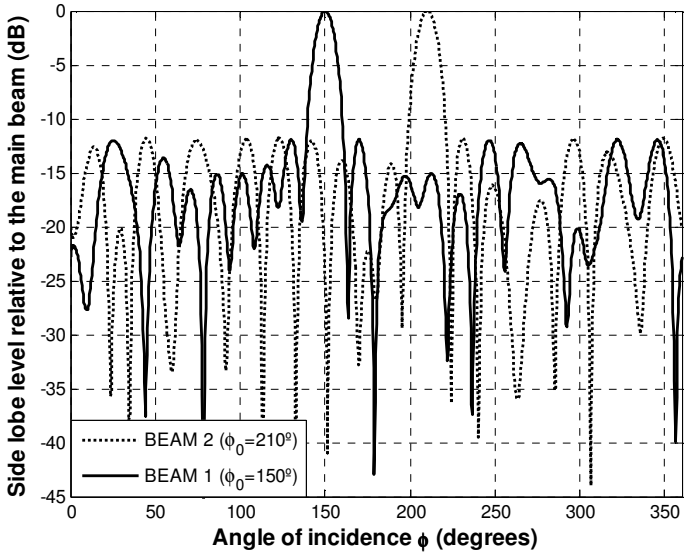
The simulation results of using this evolutionary algorithm for the optimization of the complex inputs of the feeding network are presented in the next section.

4. Simulation results

The DE algorithm was implemented to study the behavior of the array factor generated by the configurations shown in the Section II. The experiments parameters were set as follows: maximum number of generations r_{max}=500, population size N_p=200, and F=0.5 [6]. Figures 4-5 illustrate the behavior of the array factor generated by the configurations shown in Fig. 2 and Fig. 3.

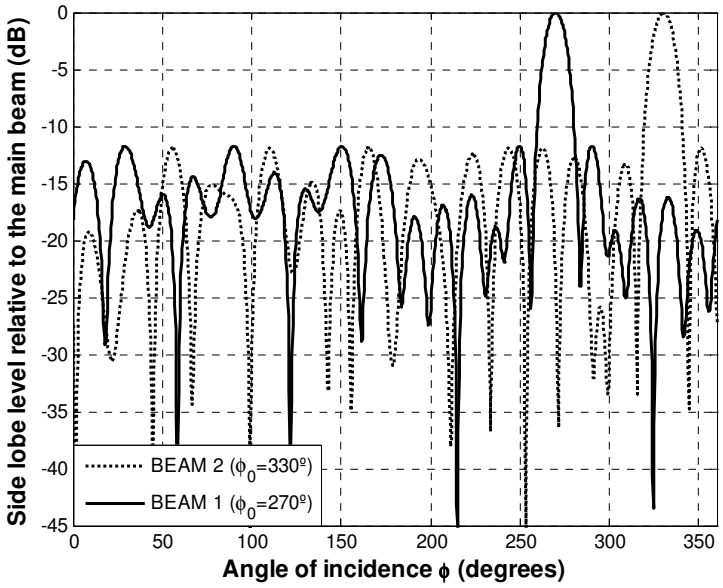


a)



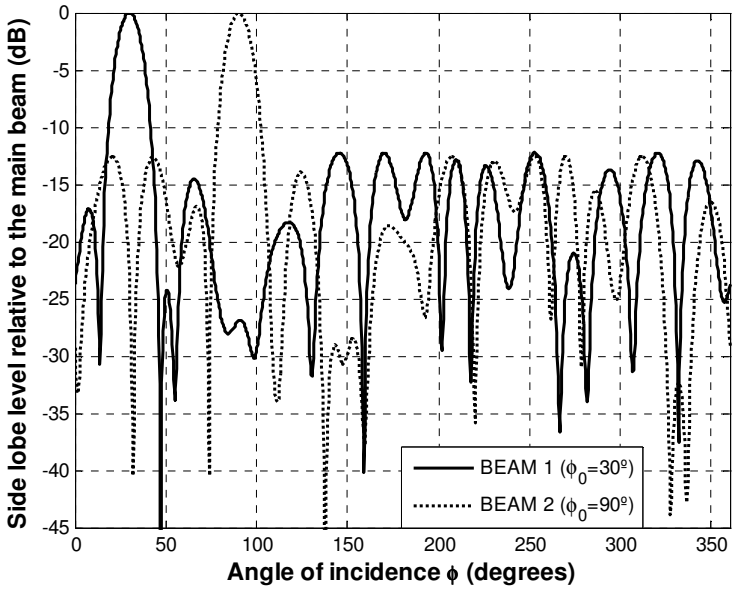
b)

Figure 4. Array factor generated by the configuration 1 shown in Fig. 2, a) ϕ_0=30º for beam # 1 and ϕ_0=90º for beam # 2, b) ϕ_0=150º for beam # 1 and ϕ_0=210º for beam # 2, c) ϕ_0=270º for beam # 1 and ϕ_0 =330º for beam # 2.



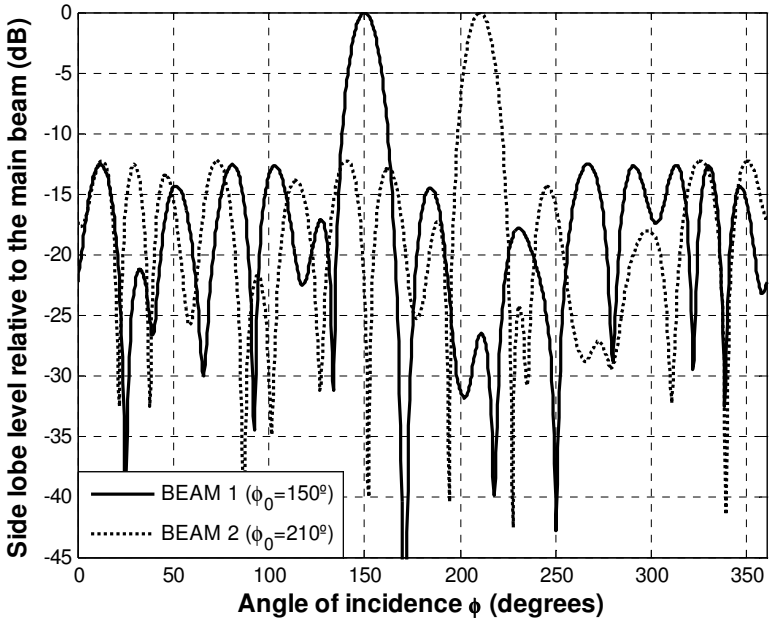
c)

Figure 4. (continued)

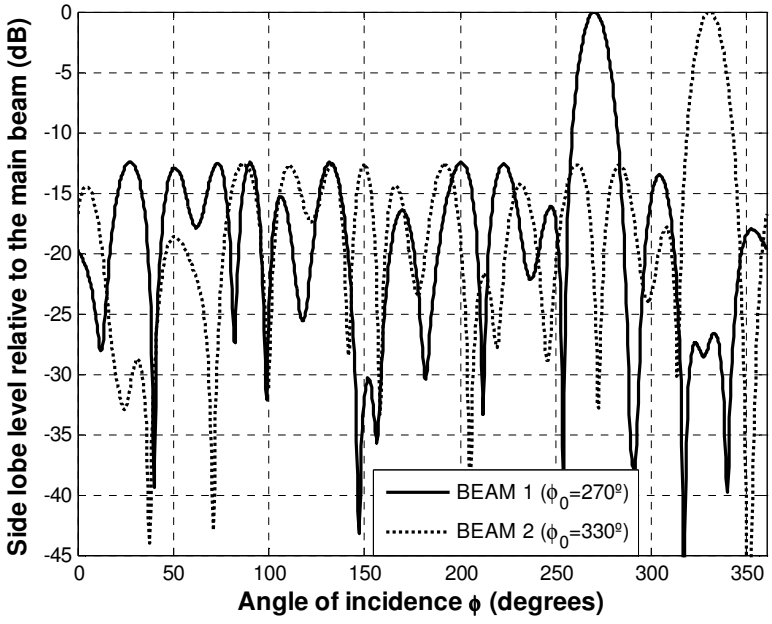


a)

Figure 5. Array factor generated by the configuration 2 illustrated in the Figure 3, a) ϕ_0 =30º for beam # 1 and ϕ_0 =90º for beam # 2, b) ϕ_0 =150º for beam # 1 and ϕ_0 =210º for beam # 2, c) ϕ_0 =270º for beam # 1 and ϕ_0 =330º for beam # 2.



b)



c)

Figure 5. (continued)

If it is considered the beam 1 from the Fig. 4, it is obtained a *SLL*=-12.66 dB and *D*=13.62 dB for ϕ_0=30º, *SLL*=-11.95 dB and *D*=13.55 dB for ϕ_0=150º, and *SLL*=-11.71 dB and *D*=13.54 dB for ϕ_0=270º. For the beam 2, it is obtained a *SLL*=-12.21 dB and *D*=13.57 dB for ϕ_0=90º, *SLL*=-11.8 dB and *D*=13.4 dB for ϕ_0=210º, and *SLL*=-11.92 dB and *D*=13.51 dB for ϕ_0=330º. Considering the beam 1 from the Fig. 5, it is obtained a *SLL*=-12.22 dB and *D*=13.44 dB for ϕ_0=30º, *SLL*=-12.58 dB *D*=13.44 dB for ϕ_0=150º, and *SLL*=-12.47 dB and *D*=13.49 dB for ϕ_0=270º. For the beam 2, it is obtained a *SLL*=-12.53 dB and *D*=13.49 dB for ϕ_0=90º, *SLL*=-12.23 dB *D*=13.46 dB for ϕ_0=210º, and *SLL*=-12.63 dB and *D*=13.45 dB for ϕ_0 =330º.
If these results are compared with respect to the uniform excitation case with conventional progressive phase excitation for *N*=26 (*SLL*=-6.92, dB and *D*=13.21 dB), we have a very significant performance improvement in terms of the side lobe level and a very significant simplification of the feeding network, i.e., the interesting aspect is that these two scannable beams are generated with *N*/2 complex inputs.
In this paper the idea was to demonstrate the possibilities of simplifying the feeding network for multi-beam circular antenna arrays by using CORPS. Although it was presented the case to generate two scannable beams, it is perfectly possible to define independently the number of input ports (defined by the number of orthogonal beams to be used simultaneously) and the number of radiating elements.

5. Conclusions

The design of beam-forming networks for scannable multi-beam circular antenna arrays using CORPS has been presented. Simulation results reveal that the design of CORPS-BFN optimizing the complex inputs with the DE algorithm could generate scannable multiple beams with a significant simplification of the feeding network. The behavior of the array factor for different configurations of CORPS-BFN for a scannable multibeam circular array was studied and analyzed.
Future work will deal with the design of CORPS-BFN for scannable multibeam planar (bi-dimensional) arrays and the study of new structures for designing BFN for multiple beam antenna systems.

Acknowledgments. This work was supported by the Mexican National Science and Technology Council, CONACyT, under grant 127919 and the Science and Technology Council of Tamaulipas Mexico, COTACyT, under grant 108166.

References

1. D. Betancourt and C. del Río. A Novel Methodology to Feed Phased Array Antennas. IEEE Transactions on Antennas and Propagation, Vol. 55, Issue 9, September, 2007, pp. 2489-2494.
2. M. A. Panduro and C. del Rio Bocio. Design of beam-forming networks for scannable multi-beam antenna arrays using CORPS. Progress In Electromagnetics Research, Vol. 84, 173-188, 2008.
3. Marco A Panduro and C. del Río. Design of Beamforming Networks using CORPS and

Evolutionary Optimization. AEU International Journal of Electronics Communications, Vol. 63, Issue 5, 2009, pp. 353-365.

4. D. Betancourt and C. del Rio Bocio, A Beamforming Network For Multibeam Antenna Arrays Based On Coherent Radiating Periodic Structures. Antennas and Propagation 2007, EuCAP 2007, Second European Conference, pp. 1-4, Nov. 2007.
5. M. A. Panduro and C. del Rio Bocio. Beam-forming Networks for Scannable Multi-beam Antenna Arrays using CORPS and Differential Evolution. Antennas and Propagation 2009 EuCAP 2009, 3rd European Conference, pp. 23-27, March 2009.
6. V. Feortisov and S. Janaqui. Generalization of the strategies in differential evolution. Proceedings of the IEEE Conference Evolutionary Computation, 1996.
7. S. Yang, A. Qing and Y. B. Gan. Synthesis of low side lobe antenna arrays using the differential evolution algorithm. IEEE Transactions on Antennas and Propagation Conference, 2003, pp. 1-22.
8. Marco A. Panduro, Carlos A. Brizuela, Luz I. Balderas y Diana A. Acosta, A Comparison of Genetic Algorithms, Particle Swarm Optimization and the Differential Evolution Method for the Design of Scannable Circular Antenna Arrays. Progress In Electromagnetics Research B, EMW Publishing Estados Unidos de América, Vol. 13, pp. 171-186, 2009.
9. Kurup, D., Himdi, M., and Rydberg, A.: Synthesis of uniform amplitude unequally spaced antenna arrays using the differential algorithm, IEEE Trans. Antennas Propagation, 2003, Vol. 51, pp. 2210-2217.
10. Parsopoulos K.E., Tasoulis D.K., Pavlidis N.G., Plagianakos V.P., Vrahatis M.N. Vector evaluated differential evolution for multi-objective optimization. IEEE Congress on Evolutionary Computation, 2004, pp. 19-23.
11. Storn, R., and Price, K. Minimizing the real functions of the ICEC'96 contest by differential evolution. In Proceedings of the IEEE Conference Evolutionary Computation 1996.
12. C. Balanis, Antenna Theory-Analysis and Design, 2nd Ed. New York: Wiley, 1997.
13. Y. Rahmat-Samii and E. Michielssen. Electromagnetic optimization by genetic algorithms. Wiley & Sons, New York, 1999.
14. Haupt, R. Thinned arrays using genetic algorithms. IEEE Transactions on Antennas and Propagation, 1994, Vol. 42, pp. 993-999.
15. Golberg, D. E. Genetic algorithms in search, optimization, and machine learning. (Addison-Wesley, Massachusetts, 1989.
16. Panduro, M. A. Design of coherently radiating structures in a linear array geometry using genetic algorithms. AEU International Journal of Electronics and Communications, 2007, Vol. 61, Issue 8, pp. 515-520.
17. Panduro M. A., Mendez A. L., Dominguez R. and Romero G. Design of Non-uniform Circular Antenna Arrays for Side Lobe Reduction using the Method of Genetic Algorithms. AEU International Journal of Electronics and Communications, 2006, Vol. 60, Issue 10, pp. 713-717.
18. F. J. Ares-Pena, J. A. Rodriguez-Gonzalez, E. Villanueva-Lopez, S. R. Rengarajan. Genetic algorithms in the design and optimization of antenna array patterns. IEEE Transactions on Antennas and Propagation 47 (1999), 506-510.

Near Optimal Antenna Selection Model for MIMO Systems

Rafael Aguilar González, Ulises Pineda Rico, Enrique Stevens-Navarro, and Juan Francisco Castillo León

Facultad de Ciencias, Universidad Autónoma de San Luis Potosí,
Av. Salvador Nava Mtz. S/N, Zona Universitaria, 78290,
San Luis Potosí, SLP, México.

Abstract. Recently it has been shown that Multiple Input Multiple Output (MIMO) Systems offer great advantages over Simple Input Simple Output (SISO) systems particularly in terms of capacity. By increasing the number of antennas at both transmitter and receiver. Unfortunately, when the MIMO System are physically deployed, it has the disadvantages of increasing the implementation costs and the complexity involved due the increase of RF chains. However, there are several algorithms proposed to overcome such challenges. One of them consists in the selection of a sub-set of antennas with the main objective of maintaining the benefits of MIMO at an affordable trade-off among complexity and implementation cost. This article proposes a novel model for antenna selection which is able to achieve similar results to the optimal selection in terms of performance and significantly surpassing other algorithms in both performance and capacity, low cost complexity.

Keywords: Multiple Input Multiple Output (MIMO), Simple Input Simple Output (SISO), Radio Frequency (RF), Correlation Based on the Method (CBM), Improvement Correlation Based on the Method (ICBM), Line of Sight (LOS).

1 Introduction

There are several characteristics that make MIMO wireless systems very attractive. Among them find: near optimal capacity, increase in coverage area, growth of data rate, etc. Such characteristics are possible due the spatial-diversity exploitation directly size of the antenna array. Unfortunately, if the MIMO array is too large, it generates a considerable increase in the implementation cost. Thus, in the simplest case, when transmitting a signal, we require an antenna (or more if MIMO) and a RF chain. The RF chain is mainly composed of communications hardware such as analog to digital converters, filters, amplifiers and some other circuits that do not follow Moore's law [1] and hence increasing the implementation cost.

One clever way to improve the benefits of the MIMO systems is to increase the number of antennas at the receiver, providing a greater space diversity, while keeping the number of RF chains lower than the number of the antennas. Furthermore, we can make use of an *antenna selection algorithm* which selects a sub-set of antennas (i.e., the ones with better channel conditions) according to the same number of RF chains

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 286-296

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

that exist at the receiver. Thus, we take advantage of the space diversity while keeping the low cost of the MIMO system in terms of antennas and RF chains.

Ideally, the optimal method of antenna selection makes an exhaustive search to find the best L antennas of the N in an array of size $\binom{N}{L}$. The increase in capacity and performance is substantial compared to MIMO array antennas without increasing the array at the receiver [2], [3]. However, since the calculation of this combination of antennas is too complex and computationally prohibitive [4], [5], there are several algorithms that have been proposed to face this problem [6], [7] that will be briefly described in the following sections.

On this paper, we present a novel model for antenna selection in MIMO systems. Different from the existing models, it is able to achieve similar results than the optimal selection model in terms of performance and significantly surpassing other algorithms in both performance and capacity. Performance tests show that our model is able to achieve satisfactory results under Rician fading channel and Correlation channel which are very unfavorable for MIMO systems.

This paper is organized as follows. Section 2 describes a few antenna selection algorithms and channels used along this document. In Section 3.1 is described how the optimal selection works and in Section 3.2 the proposed model is presented and described, Section 4 explain how was calculated the complexity cost. Finally, the performance results are shown in Section 5 including valuable conclusions presented in Section 6.

2 MIMO Antenna Selection and Channels

2.1 Algorithms

Correlation Based on the Method (CBM) This algorithm presented in [6] assumes full knowledge of the channel by the receiver and it works as follows: if the channel matrix $\mathbf{H}$ has two lines equal, one of them must be eliminated because if one of them is not in the matrix there is no loss of information. If they have different powers, the selected row shall be the one with the highest power (the square of the norm of the vector). When any row is not identical, the correlation is applied among them eliminating the row yielding the highest value of this operation. The objective of this algorithm is to create a channel matrix with unique vectors and with the higher value of power.

Improvement Correlation Based on the Method (ICBM) A substantial improvement to the CBM algorithm was developed in [7]. Here, the algorithm calculates the correlation and non-correlation. The channel matrix is filled incrementally and does not require to know all the signals. This difference respect CBM allows to ICBM to achieve a faster antenna selection process since does not require to make the correlation with all the vectors of the channel.

CBM and ICBM were designed to work under slow Rayleigh fading channel conditions. Thus, when they work in different or faster channels, CBM and ICMB tend to decrease their effectiveness and in consequence there is a substantial loss of performance and capacity.

2.2 The Fading Channel

Rayleigh fading channel is rich in scattering and does not have a line of sight (LOS) path between transmitter and receiver, something that suits to the MIMO systems. However, there are other channels that the MIMO systems may encounter, environments that are not suitable for them but very similar to everyday wireless communications situations, e.g. absence of multi path (Rician) and sameness in the received signal (correlation).

Therefore, for sake of simplicity and generalization, we define the MIMO channel that will be used thorough this paper as a Rayleigh fading represented by a $N_R \times N_T$ antenna number at R_x and T_x respectively, $\mathbf{H}$ matrix, $\mathbf{H} = [\mathbf{h}^{(1)}, \ldots, \mathbf{h}^{(N_R)}]^T$, $\mathbf{h}^{(k)} = [h_{k,1}, \ldots, h_{k,N_T}]$, unless specified otherwise. Additionally, indexes $(\cdot)^T$ and $(\cdot)^H$ stand for transpose and conjugated transpose respectively.

Rician fading Channel When the transmitter and receiver have a LOS path, the channel follows the Rice distribution. In this case the spread is decreasing, affecting the performance of MIMO systems. This degree of affectation is determined by the K Rician Factor (KF) which is the radius of the power of the LOS component of the channel to the power in the fading component [8].

Thus, for the presence of LOS between the transmitter and receiver, the MIMO channel can be modeled as the sum of a fixed component and a scattered component, thus we have

$$\mathbf{H} = \sqrt{\frac{KF}{1+KF}}\mathbf{H}_{LOS} + \sqrt{\frac{1}{1+KF}}\mathbf{H}_w \tag{1}$$

where $\sqrt{\frac{KF}{1+KF}}\mathbf{H}_{LOS} = \mathrm{E}\{\mathbf{H}\}$ is the LOS component and the faded component is calculated as $\sqrt{\frac{1}{1+KF}}\mathbf{H}_w$, which assumes uncorrelated fading. Therefore, when $KF = 0$ corresponds to a Rayleigh fading channel while $KF = \infty$ corresponds to a non-fading channel.

Correlated Channel Another real-world phenomena that affects the MIMO systems is the correlation. This problem occurs when the separation between the base station antennas is not sufficient to allow the signals to take different paths. The order of this separation is in centimeters while the distance between base station(BS) and mobile device is the order of kilometers. Thus, when the base station antennas receive a signal from a mobile device, and if it is very close to the BS, the result is the presence of the correlation effect. A model that is widely used to define a correlation matrix is defined by

$$\sum \mathbf{H} = \sum \mathbf{R} \otimes \sum \mathbf{T}, \tag{2}$$

where $\otimes$ represents the Kronecker's product, as described in [9]. $\sum T$ is the transmit-correlation matrix corresponding to the case when each R_x antenna is the same,

$$\sum \mathbf{T} = \mathrm{E}\{\mathbf{h}_i \mathbf{h}_i^H\} \quad \forall i, \quad \mathbf{h}_i = [h_{i,1}, \ldots, h_{i,N_T}]^T, \tag{3}$$

and for the case when each T_x antenna is the same, the receive-correlation matrix is given by

$$\sum \mathbf{R} = \mathrm{E}\{\mathbf{h}_j \mathbf{h}_j^H\} \quad \forall j, \quad \mathbf{h}_j = [h_{1,j}, \ldots, h_{N_R,j}]^T. \tag{4}$$

The degree of sameness determine the correlation coefficient with 1 as a maximum correlation and 0 as no correlation.

3 Model Description

The antenna selection algorithms aim to improve the features and benefits that can be provided by a MIMO antenna array physically implemented. To make use of them, it is necessary to increase the number of antennas at the receiver in order to increase the diversity. Among all received signals, a selection is required based on the parameters that increase performance and capacity of MIMO systems. This process must be done without adding additional RF chains, therefore the implementation of the system that we are referring to is shown in Figure 1.

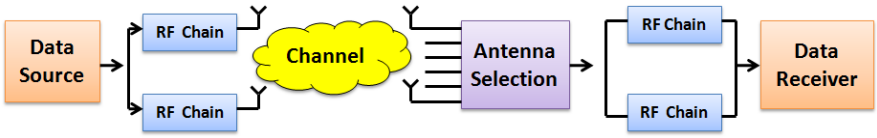


Fig. 1. Block diagram of the antenna selection model for MIMO systems.

3.1 Optimal Selection

The Optimal Selection of getting a sub-set of antennas is done by calculating $\binom{N}{L}$, which is considering all the possible combinations according to the number of antennas that exists physically and according to the number of RF chains available. Clearly, we can see that the greater number of antennas, the better results. However, notice that this operation is computationally expensive. The optimal choice is made according to the flow chart shown in Figure 2.

First we modeled the channel **H** and note that the random variables can be Rayleigh, Rician or Correlated. The channel capacity is calculated by: equation (5), **I** stands for identity matrix and ρ is the signal to noise ratio [10].

$$C = \log_2 \left[\det \left(\mathbf{I}_{N_R} + \frac{\rho}{N_T} \mathbf{H}\mathbf{H}^H \mathbf{H}\mathbf{H} \right) \right] \text{bit/s/Hz}. \tag{5}$$

In order to calculate the capacity is assigned a value from the matrix **H** and is placed temporarily in the **HH** variable which will be used exclusively for this operation and will have a size of 2×2. Since the simulation assumes that there are two RF chains at

both the transmitter and receiver, the second value is placed in **HH** after the first and will change accordingly to the results obtained in the building process from the total combinations available. All vectors will be used and will be part of the Optimum Selection from all those vectors that produced the maximum capacity. Using this selection we can get the maximum capacity and performance that can be achieved by the system. However, because this algorithm is too complex is not advisable to use it, especially if you are in channels that affect MIMO systems, e.g. the Rician fading Channel and Correlated Channel.

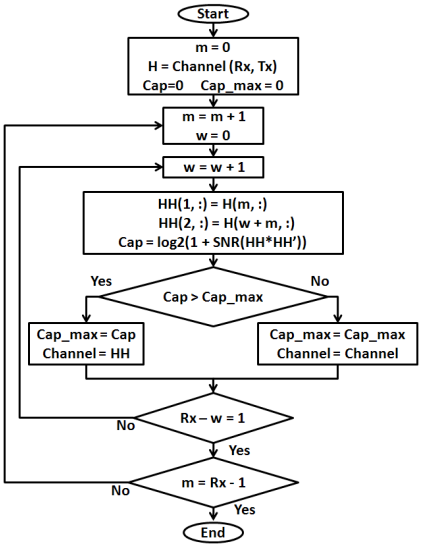


Fig. 2. Optimal Antenna Selection Model for MIMO systems.

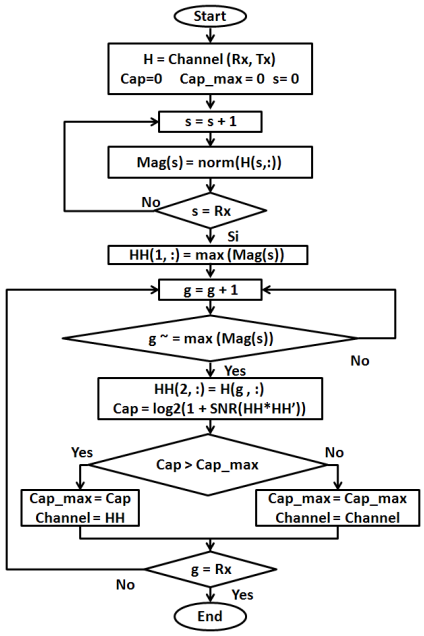


Fig. 3. Proposed Antenna Selection Model for MIMO systems.

3.2 Proposed Model

The variables generated by this algorithm have the same dimensions as the optimal selection. Random samples contained in **H** can also be Rayleigh, Rician or Correlated. Unlike other algorithms that do not handle the last two distributions. The Proposed Model algorithm begins by calculating the power of each vector of the matrix **H** by the square of the norm of the vector. Needless to say, we are after the vector (or set of vectors) that will offer the maximum power.

The capacity is also calculated using the equation (5). The vector with the maximum power is placed permanently in the first row of the matrix **HH** and has the same function

as the optimal selection. This vector, will be placed consecutively in the second row of the matrix **HH** with all different vectors (sorted in terms of maximum power) and next the capacity is calculated. Finally, will be selected the second line which has obtained the highest capacity. The antenna choice of our proposed model is made according to the flow chart shown in Figure 3.

The main advantage of this algorithm is that forms the matrix of the vector channel has better power characteristics and the calculation of capacity does combinations of values of the matrix, because it assigned a specific location for the first selected signal, and only test the other options in conjunction with the first. As will be shown in the following section, this algorithm stands out because its results are very close in performance to the Optimal Selection but does the selection process with a much smaller number of operations to the Optimal Selection, as shown in the results, making this model suitable for implementation in mobile systems by the low battery consumption that will have, as well as to be implemented in fast channels.

This method is similar to the algorithms shown in Section 2, because you get a similar way the first selected signal, but also makes the calculation of capacity as it holds the optimal selection, but may seem complex at base part of their selection on ability, but having filled an area of the channel matrix with the vector of maximum power, drastically reducing the number of operations performed, showing that this model is adequate for obtaining good results with a reduced cost of complexity

4 Measuring Complexity

Through the theory of computational complexity can be aware of the resources necessary to carry out an algorithm, the more elaborate and extensive as this will require a greater number of operations which is reflected in resources in the case of a device Mobile means more battery consumption.

Using the function Floating point Operations Per Second (Flops) of Matlab [11], which measures performance based on the number and type of operations performed by the algorithm. As cited in [12], consumption of flops that have some operations is reflected in Table 1 and Table 2 . Where the degree of complexity is based on the characteristics that have the operation to do, whether scalar, vectors or matrices, real or complex.

Operation	if $\mathbf{c} \in \mathbb{R}$ $O(\cdot)$	if $\mathbf{c} \in \mathbb{C}$ $O(\cdot)$	
$\mathbf{a} + \mathbf{b} = \mathbf{c}$	$n+8$	$n+9$	$flops$
$\mathbf{a} - \mathbf{b} = \mathbf{c}$	$n+8$	$2n+9$	$flops$
$\mathbf{a} * \mathbf{b}^{\mathbf{T}} = \mathrm{c}$	$6n+10$	$2(2n+7)$	$flops$

Table 1. Elementary vector operations measured in flops

Operation	if $\mathbf{C} \in \mathbb{R}$ $O(\cdot)$	if $\mathbf{C} \in \mathbb{C}$ $O(\cdot)$	
$\mathbf{A} + \mathbf{B} = \mathbf{C}$	$n+8$	$n+9$	$flops$
$\mathbf{A} - \mathbf{B} = \mathbf{C}$	$n+8$	$2n+9$	$flops$
$\mathbf{A} * \mathbf{B}^T = \mathbf{C}$	$6n+10$	$2(2n+7)$	$flops$

Table 2. Elementary matrix operations measured in flops

5 Results

The simulations performed are developed using Monte Carlo method in Matlab. It is included for comparison purposes a MIMO system with two antennas each at the T_x and the R_x, which is called Non Selection. The MIMO system that implemented to make use of antenna selection algorithms is an array of RF chains of 2×2, where the T_x has two antennas and the R_x has eight antennas from which only two will be selected. The algorithms considered are ICBM, the Optimal Selection and the Proposed Model.

CBM algorithm was used as reference by the good results obtained in time using it, in the Figures 4, 5 and 6, do not show these results to avoid congestion at the figures and it was decided to begin comparisons with ICBM.

Figure 4 top shows that the proposed antenna selection model is very close to optimal model performance and far superior to the improvements made to CBM and especially at 2×2 MIMO system. Note that although, note in Figure 4 that the proposed model cannot fully achieve the same capacity as the Optimal Selection nevertheless it achieves a higher capacity than ICBM in a Rayleigh channel (the channel where ICBM performs better). Note the capacity obtained with ICBM and the proposed model are similar, however, if we also compare them in terms of Bit Error Rate, as shown in 4 our proposed method is better.

Figures 5 and 6 do not include the 2×2 MIMO system, or ICBM, because the proposed model have better performance than those methods, also to be subject to drastically different channels lose their properties, hence the interest focuses only on the proposed model and the optimal selection.

Figure 5 shows the performance and capacity that can have both the Optimal Selection and the Proposed Model under a Rician fading Channel. It can be seen that the Rician channel clearly affects the performance of both techniques. The Optimal Selection performs better but the proposed model stays very close. At this point, we have to emphasize the fact that the Proposed Model remains very close to the optimum signal showing robustness and ence against the absence of multi path, something that does not happen with CBM or ICBM.

The correlated channel also affects the performance of both algorithms, but the most notable change comes from the correlation coefficient of 0.5 and onwards. Figure 6 shows that although the proposed model calculates the signal strength based on mathematical correlation, this estimation is not affected by the correlated channel while keeping a close performance in comparison with the optimal. Again, the proposed model shows better characteristics against correlated channels, particularly against its counterparts CBM and ICBM (which effectiveness lie upon low correlation channels).

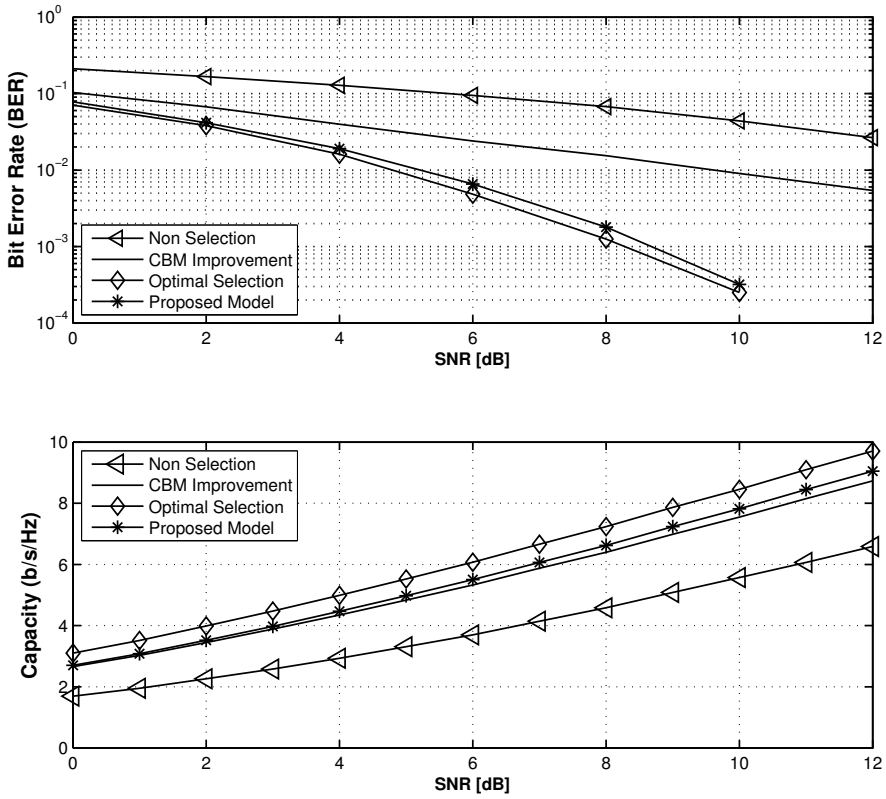


Fig. 4. Performance in Bit Error Rate (top) and Capacity (bottom) in b/s/Hz achieved by the non selection model, ICBM, optimal model and the proposed model under a Rayleigh channel conditions.

Figure 7 makes use of the function to measure the complexity Flops. The algorithms CBM, ICBM, the Proposed Model and the Optimal Selection are simulated in a Rayleigh channel. Clearly we can see that ICBM, and the Proposed Model are less than half the computational cost of the Optimal Selection, the Proposed Model stands out as making use of the ability to obtain a reasonable performance close to that of the selection optimal, but with a much lower complexity.

6 Conclusions

In this paper, we presented a new model of antenna selection for MIMO systems which is able to get better results than other algorithms that fulfill the same function. The antenna selection algorithms are designed to slow channels, but our proposed model shows more flexibility since can work in Rayleigh channels, Rician channels and under highly correlated conditions while maintaining its properties. Despite the fact of the

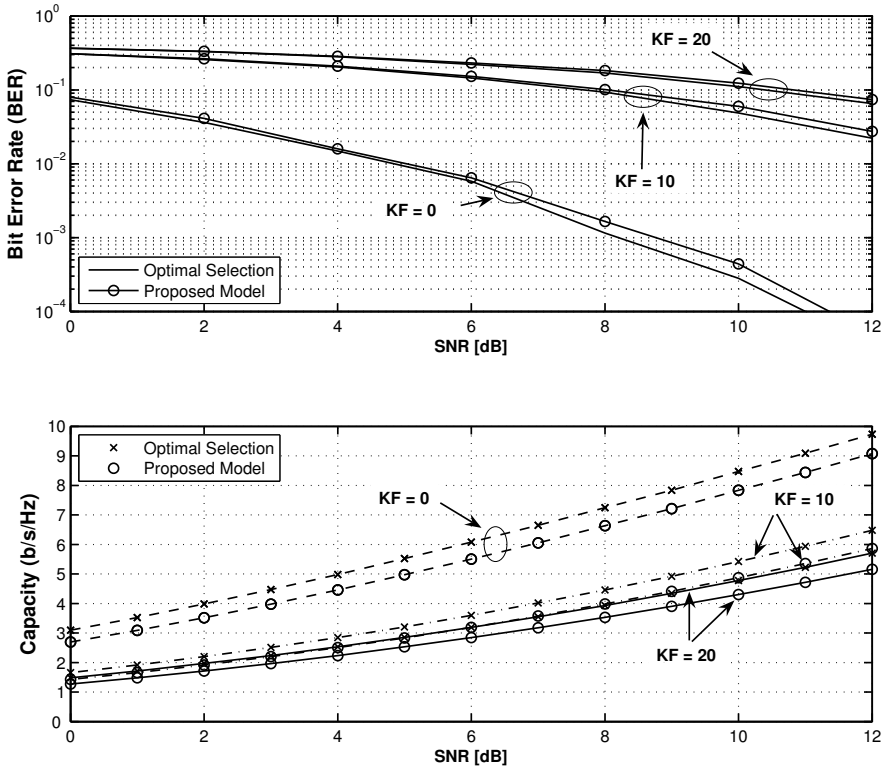


Fig. 5. Performance in Bit Error Rate (top) and Capacity (bottom) in b/s/Hz achieved by the optimal model and the proposed model under a Rician-fading channel.

influence of the channel, the proposed model shows that can be able to keep similar capacity levels as the optimal thanks to a clever choice of antennas. The analysis of the proposed model in this paper is conducted comparing the characteristics of the different antenna selection algorithms, showing the advantages of the new model as well as the favorable results very close to the best possible selection (Optimal).

The contribution made by the proposed model is to maintain the performance and capacity values similar to those obtained by the optimal model, but without making the same number of operations it. The proposed model can be implemented on mobile devices without posing a problem for battery consumption.

Acknowledgements

This work was suppoted under grants PROMEP/103.5/10/4520, PROMEP/103.5 /09/4245 and FAI: C10-FAI-05-10.38.

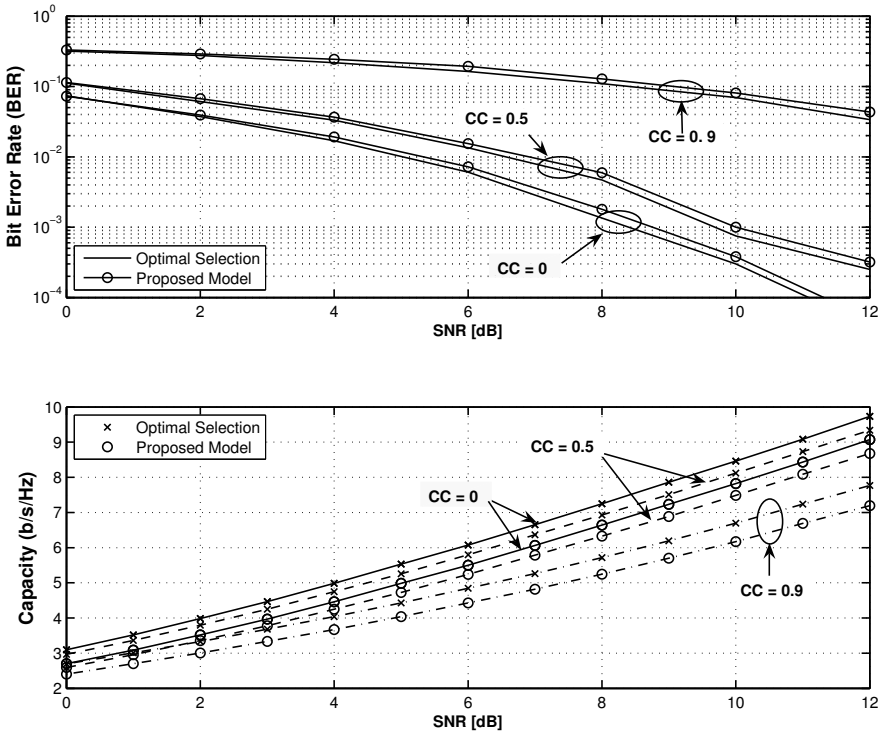


Fig. 6. Performance in Bit Error Rate (top) and Capacity (bottom) in b/s/Hz achieved by the optimal model and the proposed model under a Correlated channel.

References

1. R.R. Schaller, "Moore's law: past, present and future", IEEE spectrum, 1997.
2. Shahab Sanayei and Aria Nosratinia, "Capacity in MIMO Channels with Antenna Selection", IEEE Transactions On Information Theory, vol. 53, No. 11, November 2007.
3. Andreas F. Molisch, Moe Z. Win, Yang-Seok Choi and Jack H. Winters, "Capacity of MIMO Systems with Antenna Selection", IEEE Transactions On Wireless Communications, vol. 4, No. 4, July 2005.
4. Shahab Sanayei and Aria Nosratinia, "Antenna Selection in MIMO Systems", Adaptive Aantennas and MIMO Systems for Wireless Communications, IEEE Communications Magazine, October 2004.
5. A. Gorokhov, "Antenna selection algorithms for mea transmission systems", in Proc. Conf. Acouustics, Speech and Signal Processing 2002, pp. 2857-2860, 2002.
6. Yang-Seok Choi, Andreas F. Molisch, Moe Z. Win and Jack H. Winters, "Fast algorithms for antenna selection in MIMO systems", in Proc VTC, vol. 3, pp.1733-1737, October 2003.
7. Joo-Seok Park and Dong-Jo Park, "A New Antenna Selection Algorithm with Low Complexity for MIMO Wireless Systems", IEEE International Conference on Communications, August 2005.
8. Arogyaswami Paulraj, Rohit Nabar, and Dhananjay Gore, "Introduction to Space-Time Wireless Communications", Cambridge University Press, 2003.

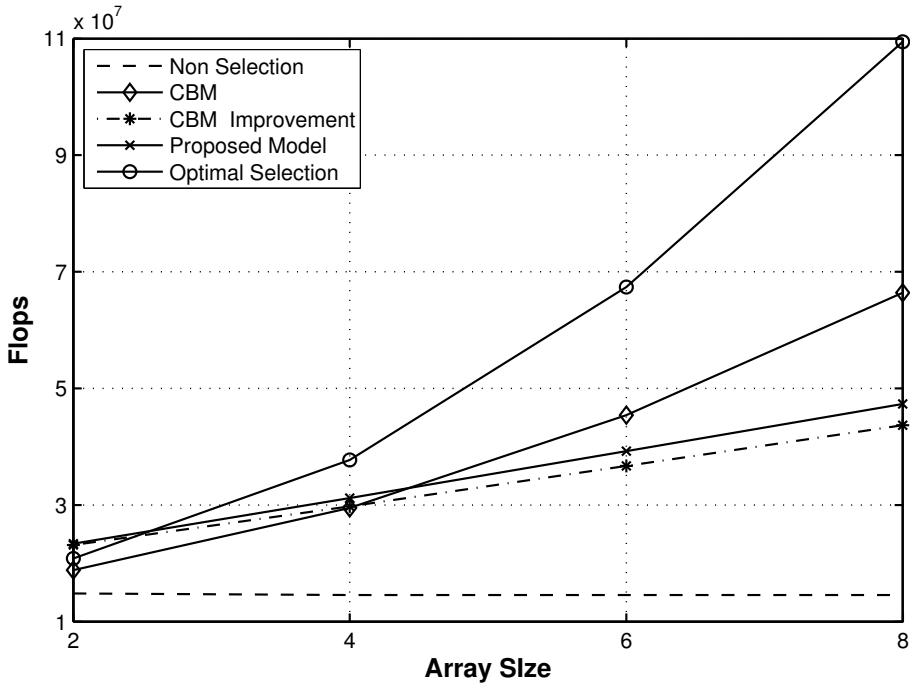


Fig. 7. Measurement of operations performed by Flops function as a function of increase diversity in the receptor achieved by a non selection model, the improved CBM, the optimal model, and the proposed model.

9. J.P. Kermoal, L. Schumacher, K.I. Pedersen, P.E. Mogensen, and F. Frederiksen, "A stochastic MIMO radio channel model with experimental validation", IEEE Journal on Selected Areas in Communications, 20(6):1211-1226, Aug 2002.
10. J. H. Winters, "On the capacity of radio communications systems with diversity in Rayleigh fading environments", IEEE J. Selected Areas Comm., vol. 5, pp. 871-878, June 1987.
11. http://icl.cs.utk.edu/papi/; http://www.mathworks.com/.
12. Pineda Rico Ulises, "Link Optimisation for MIMO Communication Systems", Lambert Academic Publishing, 2010.

Coverage Maps of 3G Cellular Networks Using Geographic Information Systems

Jazmín Ponce Rojas, Sergio Vidal-Beltrán, Iván Zamudio Castro, Federico Felipe Durán.

Maestría en Ciencias en Ingeniería de Telecomunicaciones, SEPI, ESIME Zacatenco, Instituto Politécnico Nacional, Unidad Profesional Adolfo López Mateos, Lindavista, México D.F., ponce_jaz7@hotmail.com

Abstract. This paper shows the procedure to obtain a continuous coverage map based on a collection of power level measurements through the combination of different models of Geographic Information Systems, using satellite photographs of the area under analysis, and a group of punctual samples of the power level of the Common Pilot Channel (CPICH); which is used to estimate the radio communications channel conditions, those measurement were taken at different positions and distances. A spectrum analyzer was used to obtain georeferenced measurements, and by a technique of statistical prediction, as the Krige Method, generate continuous coverage maps, making it possible to know the distribution of power level, and therefore understanding the behavior and configuration of a Base Station, which in third generation cellular systems is also called Node B.

Keywords: GIS, Coverage Maps, Cellular Network, Krige

1 Introduction

In the cell phone as all services must comply with the quality offered to users. That is why designers and administrators of wireless networks require an experimental analysis to determine the performance of networks.

In a cellular scheme, the analysis is done in the coverage area (cell) that is assigned to the Base Station –also called Node B-, it is precisely in this area where the service provider ensures that the Node B perform properly the transmission and reception of radio, filtering of the signal, amplification, modulation y demodulation of the signal, besides being the interface to the Controller of Radio Network (RNC). Normally a Node B has a total average power transmission of 20 w (43 dBm) [1].

Both the uplink (User Station to Base Station) as the downlink (Base Station to User Station) are implemented with WCDMA (*Wideband Code Division Multiple Access*) which is the technique of media access in third generation cellular networks (3G), through which multiple users access to a channel at the same time, but with a unique code that identifies (*Scrambling Code SC*).

The third generation cellular system operates in a unique center frequency at which power is the parameter of network control, and therefore the variable to analyze;

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 297-307

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

specifically examine the power of the Common Pilot Channel (CPICH), which transmits a carrier used to estimate the communication channel parameters. It is the physical reference for other channels, and is used to power control, coherent transmission and detection, channel estimation, measurement of adjacent cells and obtaining of the SC [2].

Measuring equipment currently available for this purpose is capable of taking grab samples of the power levels at certain points. For adequate coverage analysis requires the proper spacing between measurements, making it easy to apply statistical techniques such as Krige Method; it is an interpolation technique based in a sample regression, which are irregularly spaced, to predict unknown values from known values. The interpolation method of Krige facilitates the task of analyzing the distribution of power radiated by an antenna, it is sufficient to take a series of grab samples with the proper equipment (which is capable of obtaining the geographic coordinates of the location and analysis parameters) to create an experimental variogram, and based on it make the most appropriate approach to the theoretical variogram to perform interpolation and move from discrete samples to a continuous measurement map [6].

For complete analysis requires the use of different models of Geographic Information Systems (GIS, *Geographic Information System*) which is an organized integration of hardware, software and geographic data, designed to capture, store, manipulate, analyze and display all forms of geographically referenced information, to solve complex problems of planning and management to meet specific information necessary for a general vision of the area of interest [3].

1.1 Geographic Information Systems

A GIS is a geographic system because allows the creation of maps and spatial analysis; is an information system because it focuses on the management, processes data previously stored and allows for efficient, repetitive and standardized spatial consultations, for adding value to the information maintained; and is an informatics system with specialized hardware and software that process the obtained data (spatial databases).

The GIS functions as a database of geographic information that is associated by a common identifier to graphic objects on a digital map. In this case the power level of the CPICH.

By separating information into different layers, are stored separately, allowing to work with them quickly and easily, to generate new information that could not otherwise be obtained [3].

There are three groups of models of GIS

- Vector GIS.

- Raster GIS.

- Object- Oriented GIS.

1.1.1 Vector Model

The focus is on the location accuracy of the elements of the space. To digitally modeling real world entities using three spatial objects: the point, the line and the polygon on a mapping system. For example, satellite photographs.

1.1.2 Raster Model

A study area is divided into small areas or array of square cells identical in size, and the "information" is stored in each bin for each attribute in the database, for example, contour. A greater number of rows and columns in the grid (higher resolution), will involve more effort in the process of capturing information and more computational cost to process it.

1.1.3 Object-Oriented Model

While data modeling vector and Raster structure their information through of layers, object-oriented systems try to organize geographic information from geographical object itself and its relationships with others. Thus, the geographic objects are subject to a number of processes and are grouped into classes, introducing a dynamic character to the information in the system. For this reason, the object-oriented model is more suitable for situations where the nature of the objects that try to model is changing in the time and/or space.

The key advantage that allows this data structure compared to the other is that from a number of parameters in the behavior of geographic objects is possible to simulate the evolution. Because it is so versatile, the area of application of geographic information systems is very broad, can be used in most activities with a spatial component [4].

That is why precisely this model was used in this study for analyzing the behavior of the power levels radiated from the Node B to user stations. With the help of a layer of

the vector model (Satellite photograph of the study area) as shown in **Figure 1**, to obtain a clear idea of the obstacles in the propagation of the signal.

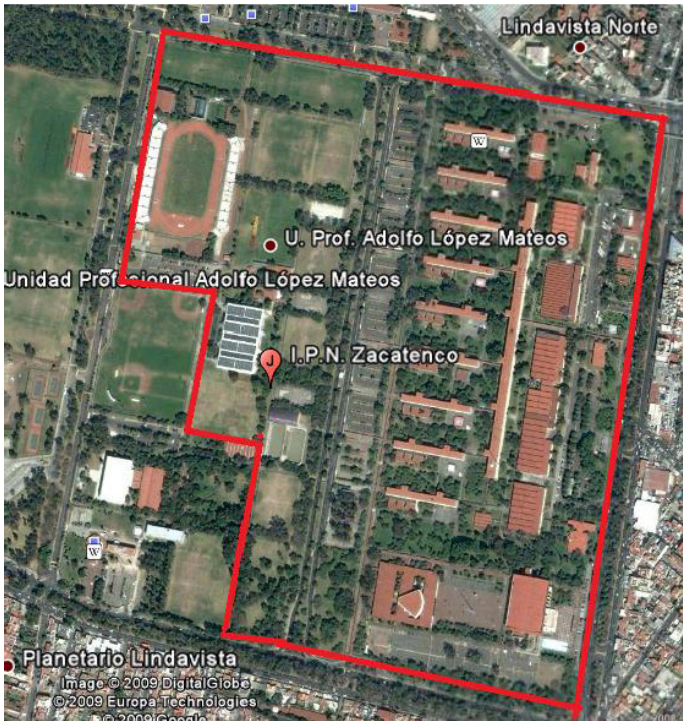


Figure 1 Satellite Photograph of the study area obtained from Google Earth

2. Methodology for the generation of GIS

To have an efficient coverage analysis, is essential that the process of creation of the coverage maps meet a set of basic criteria, ensuring the reliability and usefulness of the information contained in the system. These design criteria are shown in the following sections.

2.1 Data Selection

For this work, the latitude and longitude were selected as a geographical reference of the system, and the CPICH power level as an attribute, because through the measurement of this power level, the user terminal is able to establish a comparison between the Node B near, and decide which of them will provide the best service. This will allow that the user station know which is the dominant pilot that would define the coverage area.

2.2 Measurement process

The measurements were made with a spectrum analyzer, BTS Master MT8222A, made by ANRITSU. El BTS Master can measure the performance of Node B by connecting directly to the Node B equipment or through the air by connecting an antenna. To measure a signal WCDMA in air, you must connect the appropriate antenna to the frequency band to be measured. To know the location information of each measurement requires a GPS Antenna (Global Positioning System). The measuring equipment requires at least pick up the signal from four different satellites to ensure accurate location information of the sample [5]. Were obtained a total of 1519 measurement, the distance between them was about 3 meters, in the area shown in Figure 1, whose surface is about 0.7 Km^2. In each measurement the spectrum analyzer was placed at a height between 1.10 and 1.30 meters, since it is the average height to which the user carries his mobile equipment.

2.3 Storage or pre-processing data

The measuring equipment has an internal memory lets it store each of the measurement, and then are removed either through USB port or by networking computer equipment through Ethernet port. From the files obtained useful information is extracted with an in program language C++ and settles into a text file as shown in **Figure 2**. In which the data are arranged in descending using criteria column longitude.

LONGITUD	LATITUD	CPICH
-99.13472222	19.49777778	-82.553
-99.13472222	19.49777778	-80.167
-99.13472222	19.49777778	-83.676
-99.13472222	19.49805556	-88.068
-99.13472222	19.49833333	-84.118
-99.13472222	19.49833333	-85.757
-99.13472222	19.49833333	-88.932
-99.13472222	19.49861111	-86.631
-99.13472222	19.49861111	-83.616
-99.13472222	19.49888889	-79.743
-99.13472222	19.49916667	-82.326
-99.13472222	19.49944444	-86.039
-99.13472222	19.49972222	-84.787
-99.13472222	19.49972222	-85.6
-99.13472222	19.49972222	-87.987
-99.13472222	19.50027778	-81.967
-99.13472222	19.50027778	-81.457
-99.13472222	19.50083333	-86.443
-99.13472222	19.50083333	-88.63
-99.13472222	19.50111111	-89.686
-99.13472222	19.50111111	-91.257
-99.13472222	19.50166667	-89.88
-99.13472222	19.50166667	-93.867
-99.13472222	19.50166667	-91.075
-99.13472222	19.50166667	-88.946
-99.13472222	19.50194444	-92.528
-99.13444444	19.49583333	-59.089
-99.13444444	19.49583333	-64.401
-99.13444444	19.49583333	-61.553
-99.13444444	19.49583333	-61.553
-99.13444444	19.49583333	-59.87

Figure 2 Data file format

2.4 Data processing

Data processing is performed to obtain useful information from data previously entered into the system.

At this point coverage continuous maps are generated through **"EasyKrig"** which is a software application implemented on MatLab software platform, this makes the prediction of power levels continuously using Krige Method initially developed by Danie G. Krige in an attempt to more accurately predict ore reserves through an algorithm of least squares regression. The interpolation method of Krige facilitates the task of analyzing the distribution of power radiated by an antenna, it is sufficient to take a series of grab samples with the proper equipment (which is capable of obtaining the geographic coordinates of the location and analysis parameters) to create an experimental variogram, and based on it make the most appropriate approach to the theoretical variogram to perform interpolation and move from discrete samples to a continuous measurement map. The semivariogram provides information of the spatial behavior of a variable. Ordinary Krige was used because the mean value is not known, and we know that the value is not constant throughout the study area, but locally can be considered constant; because measurements are made at a distance will be very similar to those made in the vicinity of that point.[6]

To ensure the effectiveness of the prediction is needed validation process; included in the application of **"EasyKrig"** (as shown in the **Figure 3**) in which the approximation error is within the acceptance region determined by the variability of the measurement power.

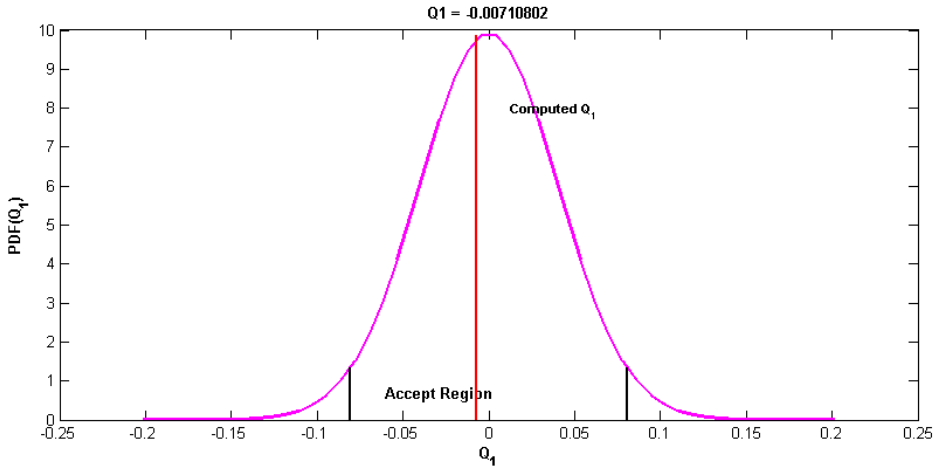


Figure 3 Graphical validation process Krige Method using the software tool "EasyKrig"

The number of measurements needed to ensure a correct prediction of the measured power levels, depend on the range of variability that this power. That is to say, if the power is very variable, must make a greater number of measurements that when the power does not change quickly.

Knowing the function most appropriate to the behavior of the measured power level the information is plotted so that the axes are defined by geographical coordinates, and the power level determines the color which represents the sample, as shown in **Figure 4**. It forming the second layer of the GIS.

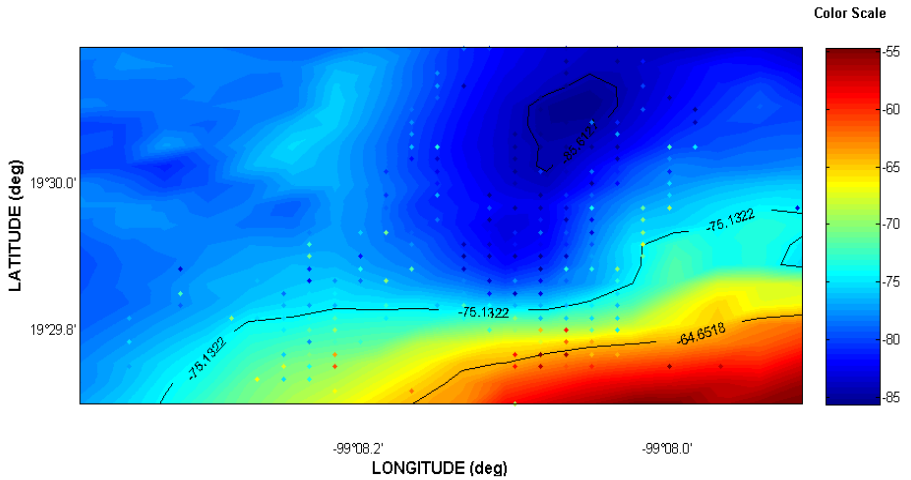


Figure 4 Continuous map of the power levels

2.5 Production data

After data processing we have two layers; satellite photography of the study area and the coverage map. These overlap to produce new data, as obstacles in propagation, the ratio of distance/attenuation of the signal, radiation pattern of transmitting antenna, to name a few. The result of the overlay is shown in **Figure 5**.

Because different Base Stations were radiating towards the area of interest, added more layers to GIS, allowing analysis expands the possibilities, because the process information of each base station in a different layer, thanks to the measuring equipment can identify each SC, it is possible to study interference between adjacent cells, as shown in **Figure 6**.

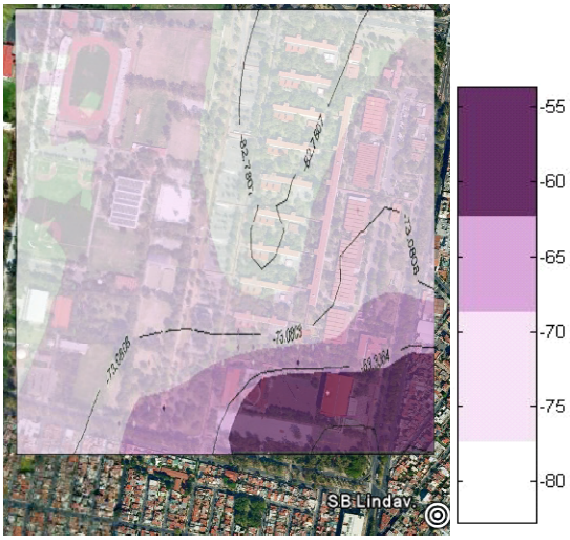


Figure 5 CPICH power distribution in dBm

Figure 6 Comparison of the coverage area of two Base Stations

3. Interpretation of Results

In each layer of the information system were analyzed separately each of the main base stations that provide service in the study area. This allowed a deeper analysis.

For example in Figure 5, we can notice that as the signal collides with building of different heights suffers attenuation proportional to the height of the same. We can explain why there is an increase in power in the upper left of Figure 5; this is due to the effect of multipath propagation, in third generation cellular systems by the type of media access is a favorable effect on propagation environments contaminated.

On the other hand, Figure 6 shows that the Pilot Dominance (strongest signal, indicating the possibility of providing better service) between two base stations with more influence in the radio signal has a conflict, since both radiate a power of similar intensity in the same area, causing the mobile device has a conflict in the choice of the base station will provide service.

Using a SIG in this work, is possible know the configuration of the segmentation of the base station antennas, as shown in **Figure 7**. And indirectly shows the areas where the call will transfer smoothly, because they will not switch to another base station, just the call is transferred to another base station sector.

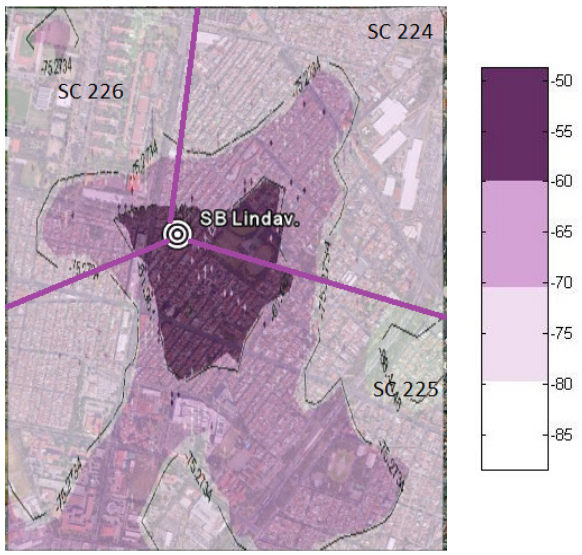


Figure 7 CPICH power distribution for the sectors of the Base Station identified by the SC 224, 225, 226 [dBm]

4. Conclusions

The rapid growth of cellular networks in Mexico and throughout the world with the aim of providing more benefits to users, causing the cell outline is saturated, creating

problems such as interference between base stations. Conflict can be avoided with adequate coverage analysis.

It is at this point that GIS are useful, as a perfect complement for different types of information about a specific geographic area to obtain information not previously known,

- ✓ as the coverage area of each base station system for purposes of interference analysis and transfer areas,
- ✓ obstacles in the signal propagation,
- ✓ multipath propagation effects,
- ✓ Pilot pollution, etc.

Number of parameters that can be studied depends on the capabilities of the measuring equipment, and existing vector model of GIS for the area in question, such as satellite images, contour surveys, hydrographic, etc.

5. References

1. Tachikawa, Keji, "*WCDMA Mobile Communications System",* John Wiley & Sons, 2002.

2. Laiho Jaana and Wacker Achim, Novosad Tomás, "*Radio Network Planning and Optimization for UMTS, second edition*", John Wiley and Sons, 2006.

3. Geographic Information Systems http://www.inbio.ac.cr/es/inbio/inb_sig.htm

4. *Geographic Information System in the management of natural hazards* http://www.oas.org/dsd/publications/unit/oea65s/ch10.htm

5. ANRITSU "*Practical Tips on WCDMA Measurements*", Application Note No. 11410-00378, Rev. B Printed in United States 2008-09.

6. Ponce Rojas Sonia Jazmín, "*Krige Method Application for the Coverage Analysis of a Node-B in a WCDMA Network*", International Journal of Communications, Network and System Sciences (IJCNS), 2011.

7. Ponce Rojas Sonia Jazmín, "*Analysis of the phenomenon of multipath and cell size in networks of third generationWCDMA type*", http://itzamna.bnct.ipn.mx:8080/dspace/bitstream/123456789/6503/1/ANALISISIMPACTO.pdf

8. Nagendra Sah, Neelam Rup Prakash, Deepak Kumar, "*Implication of Propagation Modeling in Coverage Prediction*", I S S N : 0 9 7 6 - 8 4 9 1 (On l i n e), IJCST Vol. 1, Iss ue 1, September 2010.

9. Valkealahti K., Höglund A., Parkkinen J. and Hämäläinen A., "*WCDMA Common Pilot Power Control for Load and Coverage Balancing"*, Proceedings of the 13th IEEE

10. International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2002), vol. 3, pp. 1412-1416.

Complexity improved Sphere Decoder for Highly Correlated and LOS channels

Juan Francisco Castillo León, Rafael Aguilar González, Ulises Pineda Rico, and Enrique Stevens-Navarro

Facultad de Ciencias
Universidad Autónoma de San Luis Potosí
Salvador Nava S/N, Zona Universitaria
S.L.P., Mexico
e-mail:jncastillo45@gmail.com, rafa_ag108@hotmail.com,
u_pineda@galia.fc.uaslp.mx, estevens@galia.fc.uaslp.mx

Abstract. In recent years it has been shown that iterative decoding techniques improve the performance (bit error rate) of various digital communication systems. Techniques for Multiple-input multiple-output (MIMO) are a key technology to promote high-speed wireless communication under the need of a low complexity iterative scheme for detection. Spherical decoding (SD) has been suggested as an efficient algorithm to solve such detection problem. SD is known as an algorithm of polynomial complexity without clearly specifying the assumptions made about its structure. Recently, SD has become a powerful tool to achieve a performance close to maximum likelihood (ML) algorithm (considered ideal) but involving lower complexity. In this paper we analyze the spherical decoder performance compared to other decoders on different channels using Montecarlo simulations.

Keywords: Sphere Decoding, Wireless Communications, Multi-Antenna Systems, Maximum Likelihood, Zero Forcing, Complexity.

1 Introduction

Wireless communications have captured the attention and imagination of the world and have become the segment's largest and fastest growing subject in the area of telecommunications. The main reasons are the desire for mobility and access to the network without a physical connection (wired). Various technologies and systems have been proposed to provide wireless communication services [1].

The first generation mobile systems (1G) is characterized by analog transmission of voice and it was introduced in the early 80's. Systems of second generation (2G) are distinguished by the digital transmission of voice and data at low rates. The third generation (3G), arises from the need to increase data transmission capacity in order to offer services like Internet access, video conferencing, television, and downloading files [2].

The success of wireless communications has been primarily associated with a steady increase in system capacity and quality of service. The bandwidth is limited and expen-

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 308-316
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

sive and in order to continue this trend should be used new technologies to provide greater spectral efficiency and reliability.

Traditionally the antenna systems are formed by one transmitter and one receiver, e.g. Single-Input Single-Output (SISO) systems. In some cases this setting is changed by increasing the number of antennas to provide diversity to the system, transforming it in a Multiple-Input Multiple-Output (MIMO) system. The objective of MIMO systems is to increase the capacity given the rich scattering propagation environment offered to the signal.

The work of Foschini [3] and Telatar [4] show that by increasing the number of antennas on both sides of the channel, substantially increases the number of bits that can be transmitted (capacity), something unthinkable in SISO systems. This increased capacity is associated with a wealth of dispersion in the environment, which allows the transmission of information by independent paths.

Due to its advantages over traditional systems, the MIMO communication systems have emerged as a key technology. MIMO techniques have been proposed as extensions of existing wireless communication standards such as IEEE 802.11, HSDPA and are part of emerging standards such as IEEE 802.16.

There are generally three categories of MIMO techniques. The first aims to improve power efficiency and maximization of spatial diversity. For example, delay diversity, STBC (Space Time Block Codes), STTC (Space Time Trellis Codes). The second approach uses layers to increase the capacity, e.g. V-BLAST (Vertical-Bell Laboratories Layered Space-Time) where signals are transmitted over multiple antennas to increase transmission speed. The third type exploits the channel knowledge at the transmitter. This channel information is used for pre and post filtering in the transmitter and receiver, which can achieve a gain in capacity.

Pre-coded data to be transmitted cannot completely prevent the effects of the channel due power constraints. Also, there are few problems with the calculation of the inverse of the channel, especially when the matrix-channel is near singular or singular. Therefore, it is necessary a stage for detection at the receiver in order to ensure successful information recovery. The detection methods can be optimum (that are often complex) or suboptimal (Heuristic) which have a low computational complexity.

The Maximum likelihood (ML) detector, in general terms, it requires joint detection of an entire block of symbols [5]. Although optimal, the extreme complexity of the decoder is opposed to practical use in multiple antenna systems. Especially, when using modulations of several bits per symbol and / or many transmit antennas are involved. For multiuser detection (MUD), the block of symbols increases and so does the number of operations needed to detect, making it virtually impossible for practical use [6]

Zero Forcing Detector (ZF) uses the reverse of the channel to remove the effects of it, but despite its low complexity is not useful for practical applications since the calculation of the inverse of the channel becomes complex by increasing its size. Additionaly, the channel matrix may be not invertible and its performance is far below the ML detector.

Consequently, there has been a growing interest in the field of decoding for ML detection in digital communications. The sphere decoding (SD), offers to decrease the

computational complexity because it only explores the possible outcomes in a radio "r" thus reducing the number of operations performed to obtain a result [7].

In this paper, we analyze the performance and complexity advantages of a proposed spherical decoder comparing it with others using Monte Carlo simulations.

2 System Model

The model consists of a MIMO system with M_T transmitters and M_R receiving antennas, the received signal vector of dimension M_R is given by:

$$y = Hx + n \tag{1}$$

where H denotes the channel matrix $M_R \times M_T$, $x = [x_1 \quad x_2 \quad ... \quad x_{M_T}]^T$ is the signal transmitted vector of M_T elements and n is a complex Gaussian noise vector that is added with dimensions M_R. Inputs x are chosen independently from a constellation O to the bits per symbol Q, e.g. $|O| = 2^Q$. The set of all possible symbols to transmit is denoted by O^{M_T}. We assume for the simulation that the number of receivers equals the number of transmitters $M_T = M_R$ and also H is modeled as a Rayleigh fading channel, Rician or correlated [8].

3 Overview of Methods

3.1 Maximum Likelihood

Maxium likelihood (ML) is based on the method of least squares and the objective is to find the minimum Euclidean distance of each element from the vector received while at the same time analyzing all existing solutions, see Figure 1.

$$\widehat{x} = \arg \min_{x \in R^{M_R}} \|y - Hx\|^2 \tag{2}$$

Basically, ML consists in solving (2) from a set of possible symbols which depend entirely in the shape of the modulation scheme used. From (2) we have that $y = [y_1 y_2 \cdots, y_{M_R}]$ is the received vector and $H_{M_T \times M_R}$ corresponds to the channel, $x = [x_1 x_2 \cdots x_{M_T}]$ is the potential vector data that has been sent to, $\widehat{x} = [\widehat{x}_1 \widehat{x}_2 \cdots \widehat{x}_{M_R}]$ is the vector of received data according to the algorithm, which theoretically was sent. With this method, the complexity grows exponentially as it makes 2^{M_T} iterations before delivering a result.

3.2 Zero Forcing

For a channel with a response H, ZF decoder inverts the channel response calculating its inverse.

$$inv(H) = H^{-1} \tag{3}$$

$$H * H^{-1} = I \tag{4}$$

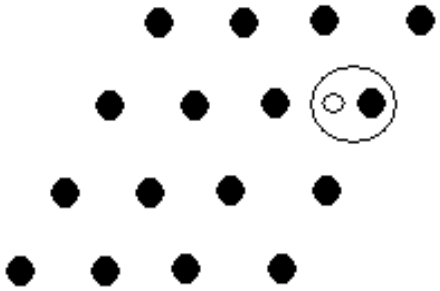


Fig. 1. Interpretation of ML.

Ideally, when performing this process, the channel effects are nullified as seen in (4).

$$y = Hx + n \tag{5}$$

$$\widehat{x} = inv(H) * y \tag{6}$$

$$\widehat{x} = x + n' \tag{7}$$

however, as shown in (7), the noise vector has been amplified by H^{-1}.

3.3 Sphere-Decoder

The basic premise of sphere decoding is quite simple: it comes to finding the minimum Euclidean distance within a sphere centered at y and radius r reducing the search space and therefore the required calculations, as shown in Figure 2.

It is clear that the closest point within the radius of the sphere is also the closest point within the full mesh.

The point Hx is a sphere of radius "r", if and only if:

$$r^2 \geq \|y - Hx\|^2 \tag{8}$$

To divide the above problem into subproblems, it is useful to consider the QR factorization of the matrix H.

$$H = QR \tag{9}$$

$$H = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{21} & Q_{22} & Q_{23} \\ Q_{31} & Q_{32} & Q_{33} \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ 0 & R_{22} & R_{23} \\ 0 & 0 & R_{33} \end{bmatrix} \tag{10}$$

So then the condition can be described as:

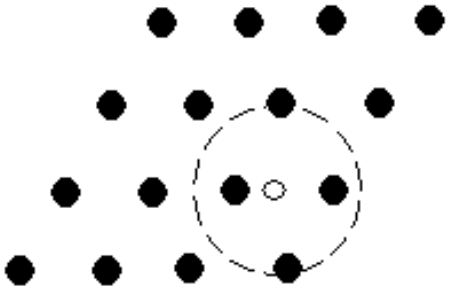


Fig. 2. Interpretation of SD.

$$r^2 \geq \|Q_1^* y - Rx\|^2 + \|Q_2^* y\|^2 \tag{11}$$

or:

$$r^2 - \|Q_2^* y\|^2 \geq \|Q_1^* y - Rx\|^2 \tag{12}$$

defining $y = Q_1^* y$ and $r^{\backprime 2} = r^2 - \| Q_2^* x \|^2$ allows us to rewrite this as

$$r^{\backprime 2} \geq \sum_{i=1}^{m} (y_i - \sum_{j=i}^{m} r_{i,j} x_j)^2 \tag{13}$$

where the first term depends only on x_m, the second term on x_m, x_{m-1}, and so on. Therefore a necesasary condition for Hx to lie inside the sphere is that $r^{\backprime 2} \geq (y_m - R_{m,m} x_m)^2$. This condition is equivalent to x_m belonging to the interval

$$\left\lceil \frac{-r^{\backprime} + y_m}{R_{m,m}} \right\rceil \leq x_m \leq \left\lfloor \frac{r^{\backprime} + y_m}{R_{m,m}} \right\rfloor \tag{14}$$

where $\lceil \bullet \rceil$ denotes rounding to the nearest larger element in the set of numbers that spans the lattice. Similarly, $\lfloor \bullet \rfloor$ denotes rounding to the nearest smaller element in the set of numbers that spans the lattice.

Of course, (14) is by no means sufficient. For every x_m satisfying (14), defining $r_{m-1}^{\backprime 2} = r^{\backprime 2} - (y_m - R_{m,m} x_m)^2$ and $y_{m-1|m} = y_{m-1} - R_{m-1,m} x_m$ a stronger necessary condition can be found by looking at the first two terms in (13), which leads to belonging to the interval

$$\left\lceil \frac{-r_{m-1}^{\backprime} + y_{m-1|m}}{R_{m-1,m-1}} \right\rceil \leq x_{m-1} \leq \left\lfloor \frac{r_{m-1}^{\backprime} + y_{m-1|m}}{R_{m-1,m-1}} \right\rfloor \tag{15}$$

Algorithm

Input: Q, R, $y = Q_1 x$, r.

1. (initialize) $k = m$, $r_m^{\backprime 2} = r^2 - \|Q_2^* x\|^2$, $y_{m|m+1} = y_m$
2. (limits) $UB(k) = \lfloor (r'_k + y_{k|k+1})/R_{k,k} \rfloor$, $x_k = \lceil (-r'_k + y_{k|k+1})/R_{k,k} \rceil - 1$
3. (increase x_k),$x_k = x_k + 1$. If $x_k \leq UB(k)$, go to 5; else , goto 4.
4. (increase k) $k = k + 1$; if $k = m + 1$, terminate the algorithm; else, goto 3.
5. (Decrease k) If $k = 1$, ir a 6; else $k = k - 1$, $y_{k|k+1} = y_k - \sum_{j=k+1}^{m} R_{k,j} x_j$, $d_k^{\backprime 2} = d_{k+1}^{\backprime 2} - (y_{k+1,k+2} - R_{k+1,k+1} x_{k+1})^2$, and goto 2.
6. Solution found. save x and it's distance from y, $d\backprime_m^2 - r\backprime_1^2 + (y_1 - R_{1,1} x_1)^2$ and goto 3.

Where Q and R come from the decomposition QR, y is the received data, $y = Q_1 x$, r is the radius of the sphere, m is the dimension received data vector and $\widehat{x}$ is the estimated figure [7].

4 Results

In figures 3 and 4 is observed as a reference the ZF detector performance over a Rayleigh fading channel, although the detector has a low computational complexity, its performance is inferior to the ML detector and the SD.

With a SNR of 10 dB and a Rician fading channel (which corrupts the data in a more aggressive way than the Rayleigh channel) with $k = 0.1$, we obtain a gain of 2.5 dB compared to ZF detector. However, this gain causes an increase in computational complexity, see Figure 3

Using the SD under similar conditions to those mentioned above, but this time on a correlated fading channel with $alpha = 0.5$, we obtain a performance similar to ZF. When $alpha < 0.5$ (highly correlated channel), we obtain a certain gain causing again a complexity increment, see Figure 4

The ideal solution to the problem is given by the ML scheme, but due to the exhaustive search performed along all the constellation of possible outcomes, it becomes prohibitive to practical use. In other words, complexity rises rapidly with increasing the number of antennas or change in the encoding (increase of bits per symbol). The SD reduces the complexity and try to get similar performance so it can be implemented and its complexity in the worst case is polynomial [5] making it more practical than ML.

The Figure 5 compares the computational complexity of ML detector and SD in terms of FLOPS (Floating Point Operations Per Second), we can see that increasing the number of antennas ($Tx = Rx$) and therefore the possible outcomes, the ML scheme increases exponentially the number of operations needed for the detection stage. On

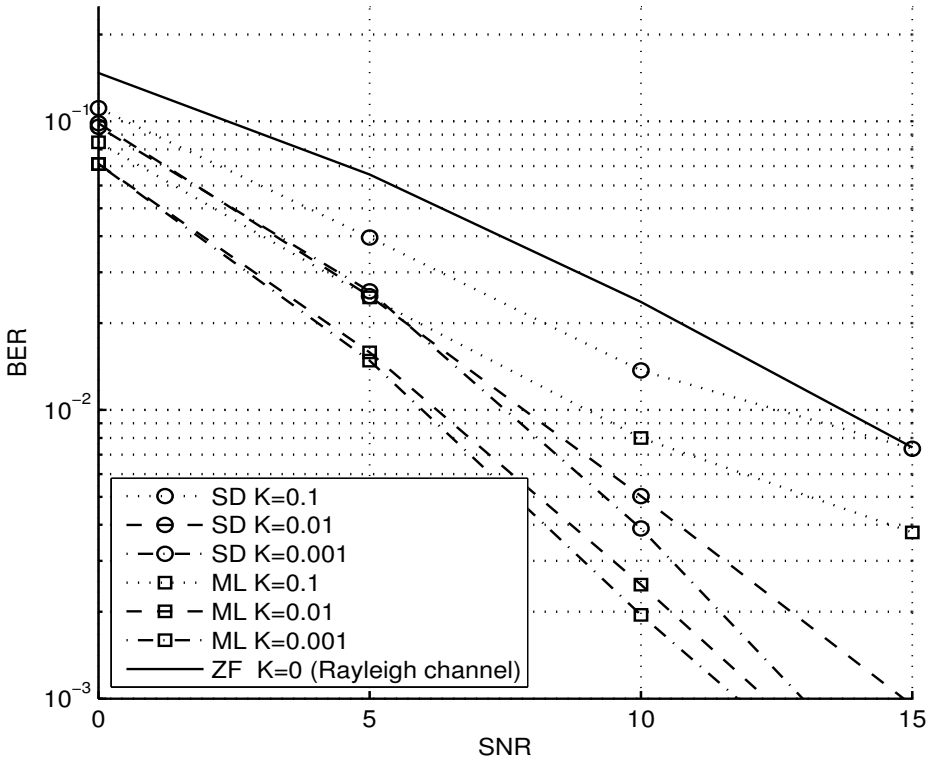


Fig. 3. BER performance graph, ML vs SD through a Rician fading channel with 2 Tx and 2 Rx, with BPSK modulation.

the other hand, SD grows far slower than ML saving a big amount of operations and making it practical to deploy.

5 Conclusions

In this paper we have studied the advantages of the spheric detector which performance is similar to ML without use as many operations making it more easy to deploy for mobility applications. Also, its lower complexity makes the SD to offer a higher throughput that the ZF detector.

Even under conditions of high correlation or fading, the proposed SD detector shows superior results in comparison with linear detectors such as ZF which its main advantage it is the low complexity but lacks of good performance. Additionally, with the rapid development of electronic devices (faster processors), the low levels of complexity and high performance, SD is set as a candidate for implementation in wireless systems of multiple antennas under critical conditions such a highly-correlated or Rician channels.

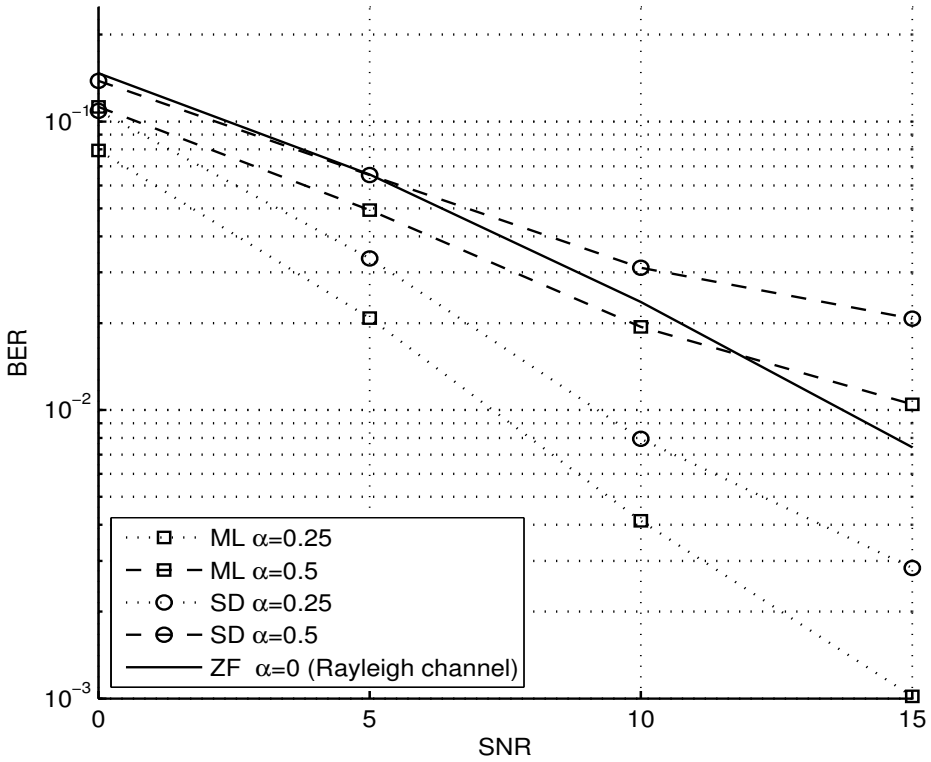


Fig. 4. BER performance graph, ML vs SD through a Correlated fading channel with 2 Tx and 2 Rx, with BPSK modulation.

Acknowledgements

This work was suppoted under grants PROMEP/103.5/10/4520, PROMEP/103.5/09/4245 and FAI: C10-FAI-05-10.38.

References

1. Theodore S. Rappaport, Wireless Communications: Principles & Practice, Prentice Hall Inc, 2002.
2. Phillip Ames, John Gabor, Wireless Communications and Computing Group, Intel Corporation, "The Evolution of Third-Generation Cellular Standards", Intel Technology Journal Q", 2000.
3. G. J. Foschini and J Gans, "On limits of wireless communications in a fading environment when using multiple antennas", Bell Labs Technical Journal, vol. 1, no. 2, Lucent Technologies, pp 41-59, Autumm 1996.
4. Telatar, I. E, "Capacity of multi-antenna Gaussian channels" ATδT Bell Labs internal report, June 1995

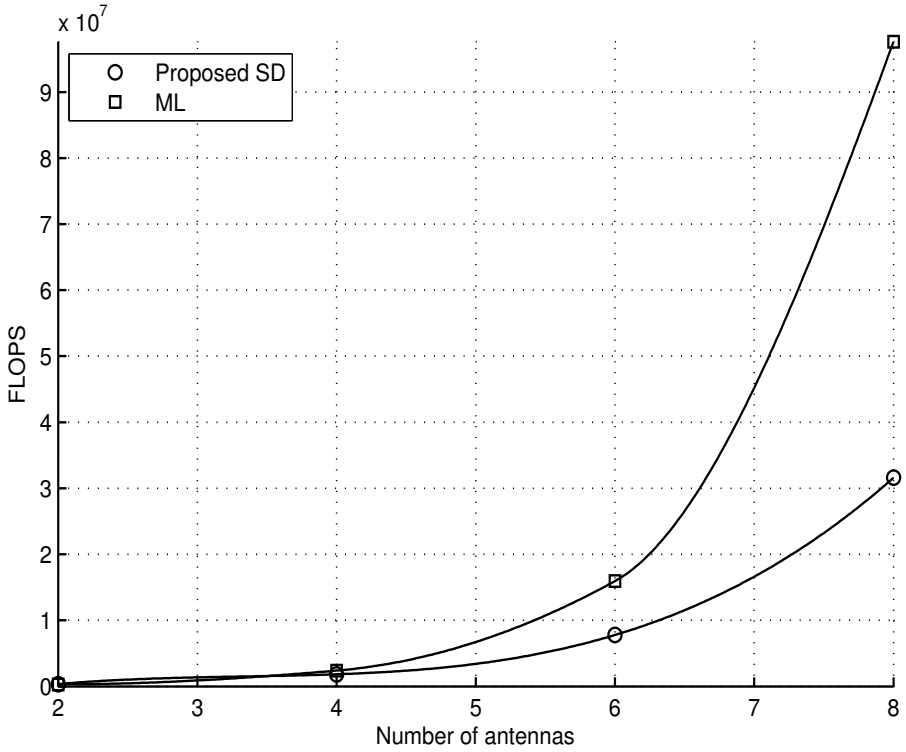


Fig. 5. Graphic of complexity, FLOPS vs. Number of antennas (Rx = Tx) and SNR = 10.

5. Joakim Jalden, Student member IEEE, "On the Complexity of Sphere Decoding in Digital Communications"
6. Oussama Damen, Ammar Chkeif, and Jean-Claude Belfiore, "Lattice Code Decoder for Space-Time Codes", IEEE communications letters, vol. 4, No. 5, May 2000.
7. B. Hassibi and H. Vikalo, "On sphere decoding algorithm I. Expected Complexity.", IEEE Transactions on Signal Processing, 53 (2005), pp. 2806-2818
8. Ulises Pineda Rico, "Link Optimisation for Mimo Communication Systems", LAP Lambert, 2010.

Analysis of Plasma Loop Antenna with Uniform and Nonuniform Distribution

Fatemeh Etesami[1], Farzad Mohajeri[1],

[1] Dept. of Electrical and Communication Eng.
School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran
{Etesami, Mohajeri}@Shirazu.ac.ir

Abstract. This paper presents the results of study of radiation pattern of surface wave plasma loop antenna using of numerical method. By solving electric field integral equation (EFIE) using moment method (MOM), the current distribution and then the radiation pattern of antenna is derived. The side lobes in the radiation pattern of nonuniform plasma antenna are more defined than uniform case. The simulation method of plasma antenna is described in both uniform and nonuniform case. The effect of various spaces between two larger sides of plasma and metal loop antenna on radiation pattern is studied. In VHF band, the radiation pattern of plasma loop antenna is elevated in both cases.

Keywords: Plasma loop antenna, plasma density, moment method, radiation patterns.

1 Introduction

Plasma antennas are usually radio frequency (RF) antennas base on plasma element instead of metal conductor. Recently they have achieved more attentions and have developed, since have attractive advantages over the conventional metal ones [1]. One important advantage is possibility of rapid changing of the effective length of the antenna by input RF power [3]. Plasma antenna can be reconfigurable with respect to shape, frequency and radiation parameters on very short time [5]. Plasma can switch on and off in microseconds and also can be a good electrical conductor when is energized, but is non-conducting when is de-energized. Therefore an unenergized plasma antenna can be difficult to detect by enemy radar. This property is useful for military communication [3]. Other unique property of plasma antenna is having variable impedance and broadband matching [7].

There are several plasma sources for plasma antenna that surface wave discharges comparing to others are the most flexible. In 1982 Moisan et al., have suggested the RF plasma surface technique for plasma discharging by one electrode in one end of tube. This technique not only doesn't have previous problems such as plasma contamination by electron erosion, but also is more acceptable in stability, lost, decreasing radar cross section (RCS) and simplicity of design [4, 5]. Recently base on this idea the plasma in antenna is rapidly created and destroyed by applying proper RF power pulses to discharge tube [2].

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 317-326

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

In physical experiments, the possibility of changing dimensions and parameters of a structure is limited. Numerical simulations can solve this problem to some extent. So, in recent decade, attention to numerical methods and simulators for analysis of plasma antennas has increased.

In this paper, we implement a computer code base on MOM technique to solve electrical field integral equation (EFIE) for current distribution on a plasma loop antenna on infinite ground plane. The radiation patterns in two cases, uniform and nonuniform plasma, are studied. Also the effect of the changing of the space between two bases of antenna on the radiation pattern is shown.

2 Theory

2.1 Principle of plasma antenna

The isotropic cold plasma is a type of dispersive medium. The relative permittivity ε_{rp} of uniform cold plasma is as follow [7]

$$\varepsilon_{rp} = 1 - \frac{\omega_{pe}^2}{\omega(\omega - j\nu_m)} = 1 - \frac{\omega_{pe}^2}{\omega^2 + \nu_m^2} - j\frac{\omega_{pe}^2 \nu_m}{\omega(\omega^2 + \nu_m^2)}. \quad (1)$$

where ω is operating frequency [rad/s], $\omega_{pe} = \sqrt{ne^2/m\varepsilon_0}$ the electron plasma frequency [rad/s], n the electron density [m^{-3}], ν_m collision frequency [Hz], m electron mass [kg], e charge of electron [C], ε_0 the free space electric permittivity [F/m].

For electromagnetic waves propagating in cold plasma, it behaves like a dielectric with permittivity less than unity for frequencies above the plasma frequency. But for frequencies below the plasma frequency, where the real part of plasma permittivity is negative, electromagnetic wave not be allowed to propagate in plasma [5, 6]. Therefore plasma is a high pass filter.

For a time harmonic wave with a time dependence of $e^{j\omega t}$ propagating in cold plasma, the propagation constant can be expressed as

$$\gamma = j\omega\sqrt{\mu\varepsilon}\left(1 + \frac{\sigma_p}{j\omega\varepsilon}\right)^{\frac{1}{2}} = \alpha + j\beta \quad (2)$$

where $\varepsilon = \varepsilon_0 \varepsilon_{rp}$ is the electric permittivity in plasma [F/m], $\mu = \mu_0$ free space magnetic permeability [H/m], $\sigma_p = \dfrac{\varepsilon_0 \omega_{pe}^2}{\nu_m + j\omega}$ the complex conductivity of plasma [S/m], α the attenuation constant [Np/m] and β the phase constant [rad/m].

2.2 Plasma column parameters

In fact the plasma density in a tube is not uniform completely and decreases axially. In a given pressure p and for an input power of P_0, the plasma density along a column decreases in an approximately linear manner from the launcher to end of the column as given by [3]

$$n(z) \approx n_L - C\nu_m(p)(h - z) \tag{3}$$

where ν_m is collision frequency, C a constant with a value $C \approx 5 \times 10^9\, m^{-4} s$, z the position along the plasma column, h the length of plasma column and n_L is a characteristics number density at the plasma frequency corresponding to the radio frequency of the source, ω, that is given as

$$n_L = \frac{\varepsilon_0 m_e}{e^2} \omega^2 (1 + \varepsilon_g) \tag{4}$$

where ε_g is the dielectric constant of the insulator (usually glass) surrounding the plasma.
The length of plasma column, h, is proportional to square root of the applied input power as shown below

$$h = \frac{n_0}{C\nu_m(p)} = \frac{A(p)\sqrt{P_0}}{C\nu_m(p)} \tag{5}$$

where $\nu_m(p)$ and $A(p)$ are constant in a definite pressure and n_0 the density at the base of the column.

2.3 Formulation of numerical method

Generally the electric field integral equation (EFIE) is given by [9]:

$$-\frac{j}{\omega\mu}[\hat{t}(r).E^{i}(r)] = \hat{t}(r).\iiint_{V}[1+\frac{1}{k^{2}}\nabla\nabla.]J(r')G(r,r')dr' \tag{6}$$

where $\hat{t}(r)$ is the tangent unit vector at position r, $E^{i}(r)$ the radiation field in free space, k wave number in free space, μ free space permittivity, V the volume of source, J the volume density of current, ∇ and $\nabla.$ respectively gradient and divergence operator on the observation coordinates and $G(r,r')$ electro dynamic Green's function in three dimensions via

$$G(r,r') = \frac{e^{-jk|r-r'|}}{4\pi|r-r'|} \tag{7}$$

where r' and r are respectively source and observation points.

Because our antenna has junction, we need to consider the EFIE in an appropriate form for this geometry. In this form, the current will be linear and a vector function of position; therefore, the basis and weighting functions in moment method will also be vectors. So let us make the line current [11]

$$\vec{I}\ (r) = I(r)\hat{t}(r) \tag{8}$$

where $I(r)$ is replaced by $J(r)$ in (6) and integration is done on length of antenna.

3 Modeling

To study and analysis of plasma antenna, it is easy to assume that it is located on infinite ground plane. First plasma is considered uniform and then nonuniform. Fig.1 shows the geometry of our plasma loop antenna on the infinite ground plane where a is the radius of plasma column, L the length of total antenna, h the height of antenna and s is the spacing between two larger sides of antenna. In Fig.1, the samples of tangent unit vector $\hat{t}(r)$ are shown.

Since ε_{rp} and σ_{p} that is mentioned in theory part, are related to uniform plasma, we need to consider an appropriate model for our geometry. So for simulation of plasma loop antenna, the length of each larger sides of antenna, h, is subdivided into N segments of length $\Delta h = h/N$ .Then the plasma density at both end of each segment is calculated using equation (3) assuming P_{0} and p are constant. For plasma density in each section, the average amount of these two end density, n_{ave}, is considered. Now we can use the relative permittivity and the conductivity of uniform plasma in

each segment [10]. The density of last segment must be considered n_L, where is the minimum value of n for which the wave propagates.

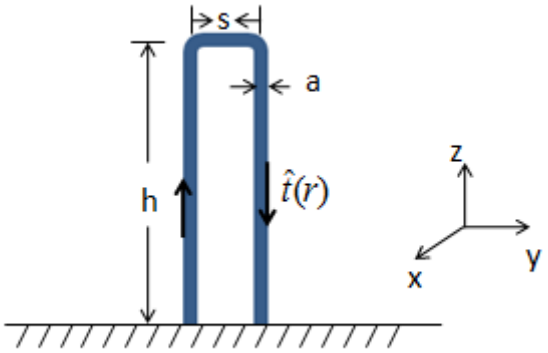


Fig. 1. The geometry of plasma loop antenna on infinite ground plane

We assume separated plasma sources for each larger side of antenna. Since the length of top slice of antenna is very smaller than two larger sides and wave length, the plasma density in it is supposed approximately uniform and equal n_L.

4 Simulation and Results

Here, we assume a = 12.5mm, the height of antenna 450mm and the space between two larger sides 100mm.The operating frequency range is in VHF band. For comparing, the aluminum (with a conductivity of $3.8\times 10^7 S/m$) antenna with the same size is applied. For a usual pressure of $400\,\mu b$, $A(p) = .016\times 10^{18}\,m^{-3}W^{-1/2}$ and ν_m = 500MHz **.** For uniform case the plasma density is assumed $n = 5\times 10^{17} m^{-3}$**.**

Fig.2 shows the normalized radiation power of plasma loop antenna in two cases, uniform and nonuniform density, at f = 300*MHz* for ν_m = 500MHz. It is obvious that the radiation power in side lobes is stronger for nonuniform case.

In Fig.3 the effect of changing the space between two larger sides of plasma and metal antennas, s, at f=300MHz is shown. Increasing this space leads to decrees of mutual effect of two sides and displacement of the main lobe of radiation pattern toward less angles. By comparing Fig.3 (a), Fig.3 (b) and Fig.3(c), one finds that the loop plasma antenna in uniform and nonuniform cases, possess radiation patterns similar to the metal one at each corresponding space. Note that the amounts of spaces are expressed in regard with wave length in Fig.3.

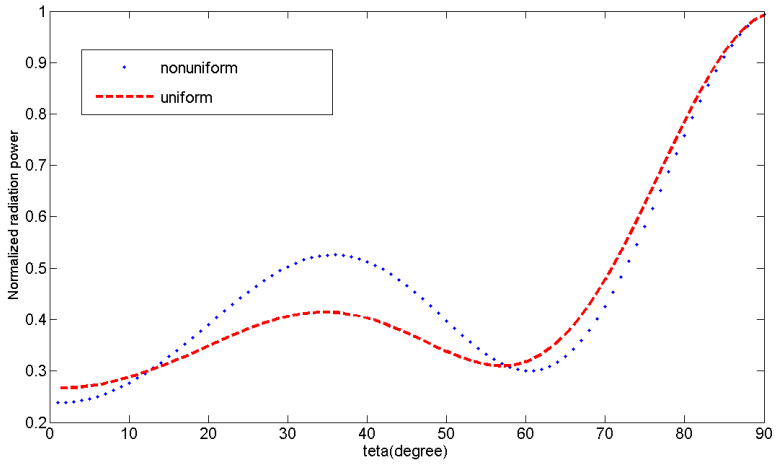
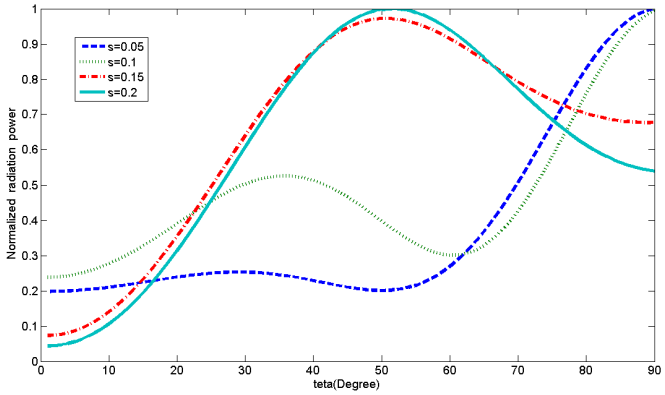


Fig. 2 The normalized radiation power at 300MHz for uniform and nonuniform plasma loop antenna

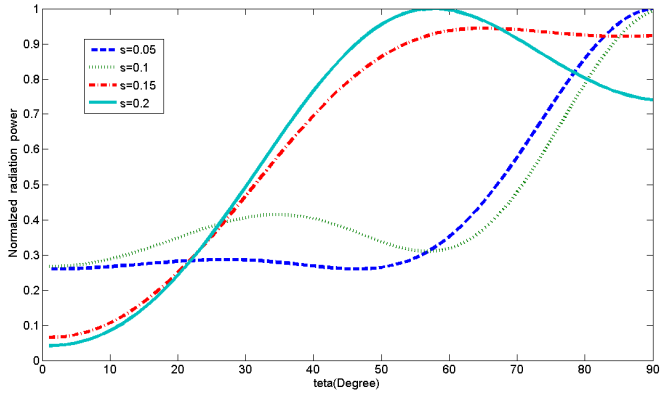
Comparing between the radiation power of the uniform and nonuniform plasma loop antenna and metal at 50*MHz*, 150*MHz*, 200*MHz*, 250*MHz* and 300*MHz*, is revealed in Fig.4. In this part the plasma density in uniform case is assumed $n = 1\times 10^{18} m^{-3}$. Approximately the radiation pattern of plasma antenna in two cases follows the radiation pattern of metal one at most of frequencies. When frequency comes to 50MHz, the plasma antenna is electrically short, so the radiation pattern is intended to metallic one. For sake of better appreciation the radiation patterns of both antennas at variant frequencies are plotted in form of polar in Fig.5.

5 Conclusion

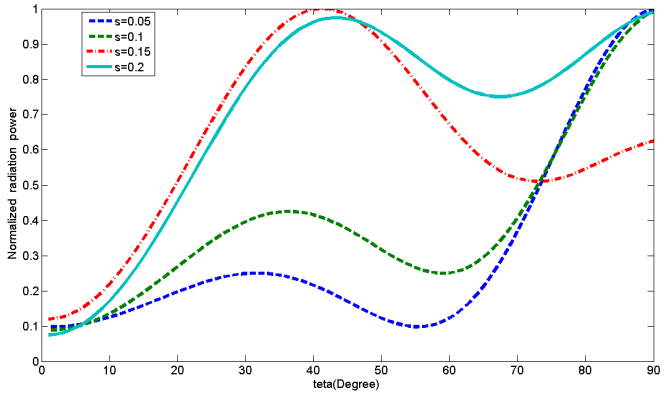
By employing the moment method for solving EFIE, the radiation pattern of uniform and nonuniform plasma loop antenna is studied. The side lobes in radiation pattern for nonuniform case are stronger than uniform one. Increasing the space between two larger sides of plasma and metal loop antenna leads to displacement of the peak of radiation pattern toward lower angles. By changing the operating frequency, the radiation pattern of plasma antenna follows metal one, approximately. Whit increasing the operating frequency the side lobes is appeared in the radiation patterns and difference between the radiation pattern of plasma and metal antenna becomes more evident.



(a)

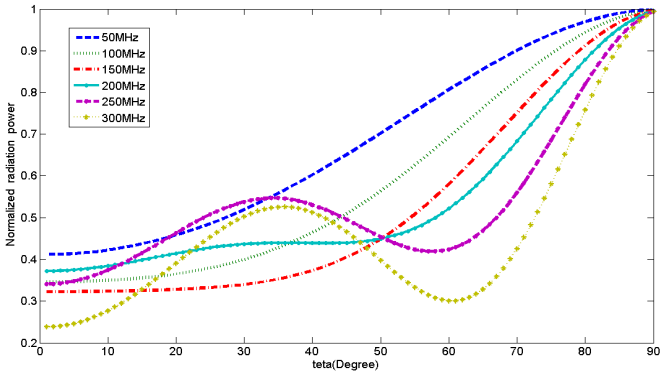


(b)

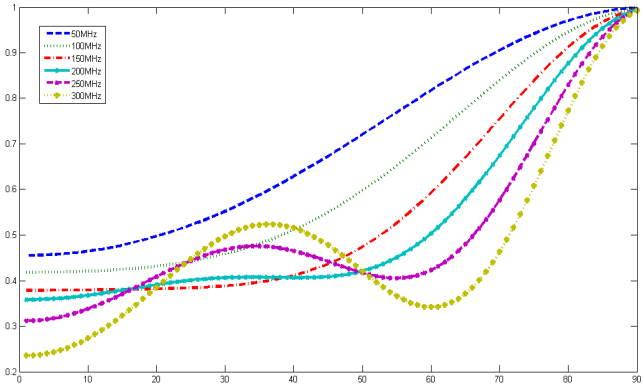


(c)

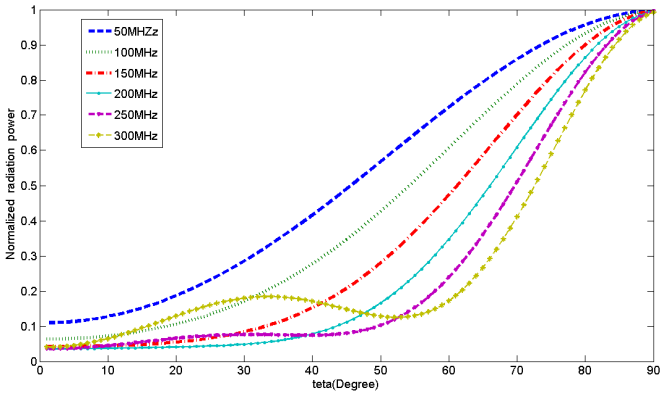
Fig. 3 The normalized radiation power of antenna for various s at f=300MHz.(a) nonuniform plasma antenna. (b) uniform plasma antenna. (c) metal antenna



(a)



(b)



(c)

Fig. 4 The normalized radiation power for s =.1m at several frequencies of (a) nonuniform plasma antenna. (b) uniform plasma antenna. (c) metal antenna

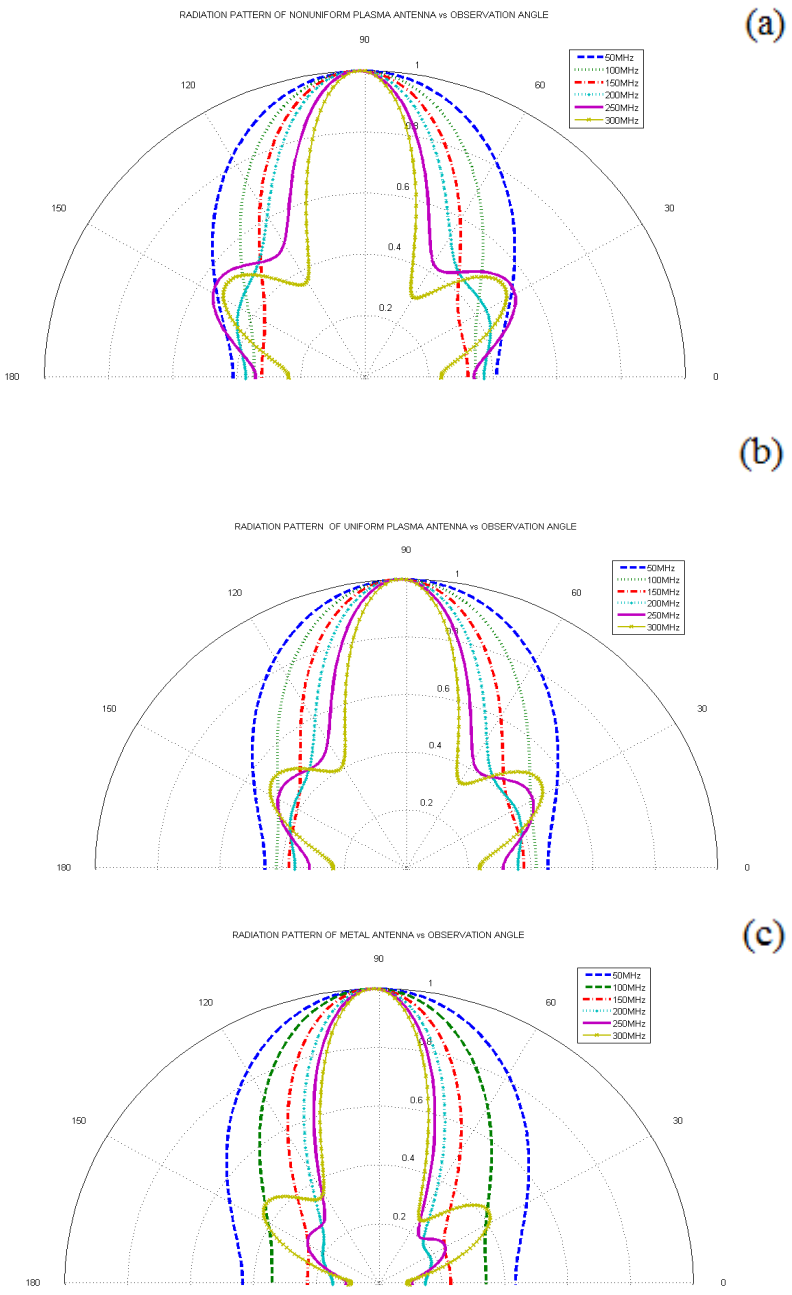


Fig. 4 The radiation pattern in observation angles for s =100mm at several frequencies of (a) nonuniform plasma antenna. (b) uniform plasma antenna. (c) metal antenna

References

1. Alexeff, I., Anderson. T., and et al.: Experimental and theoretical results with plasma antenna. IEEE Trans. Plasma Sci., Vol. 34, No. 2, 166–172, 2006.
2. Cerri, G., Leo, R. D., Primiani, V. M. and Russo P.: Measurement of the Properties of a Plasma Column used as a Radiating Element. IEEE Trans. On Instrumentation and Measurement. Vol. 57, No. 2, 242–247, 2008.
3. Rayner, J. P., Whichello A. P. and Cheetham, A. D. Physical Characteristics of Plasma Antennas. IEEE Trans. Plasma Sci., Vol. 32, No. 1, 269–281, Feb. 2004.
4. Borg, G. G., Harris, J. H., Martin, N. M., Thorncraft, D, Milliken, R., Miljak, D. G., Kwan, B., Ng, T. and Kircher, J.: Plasmas as antennas: Theory, experiment and applications. Phys. Plasma, Vol. 7, 2198–2202, July 2000.
5. Moisan, M., Shivarova, A. and Trivelpiece, A. W.: Review Article: Experimental investigation of the propagation of surface waves along a plasma column. *Plasma Phys.* Vol. 24, No. 11, 1331–1400, Feb. 1982.
6. Lee, Y., and Ganguly, S.: Analysis of plasma-column antenna using FDTD method. Microwave and Optical Technology. Lett., Vol.46, 252–259, Jan. 2005.
7. Lieberman, M. A., and Lichtenberg, A. J.: Principles of Plasma Discharges and Materials Processing. New York: Wiley, 1994, 95–96.
8. Maier, S. A.: Plasmonics Fundamentals and Application. New York: Springer, 2007, 12.
9. Gibson, W. C.: The Method of Moments in Electromagnetics. Chapman & Hall/CRC, 25–26, 73–77, 2008.
10. Sadeghi-Kia, F., Hodjat-Kashani, F., Rashed-Mohassel, J.: Analysis of Tapered Column Plasma Using Full Wave Simulator. Proceeding of the 2009 IEEE International Conference on Antenna & propagation, 301–304, 16-17 Nov. 2009.

Power Distribution Analysis of CPICH Channel in a WCDMA Network

Mariana Eslava Gutiérrez, Sergio Vidal-Beltrán,
Marco Antonio Acevedo- Mosqueda, Montserrat Jimenez-Licea

Instituto Politécnico Nacional, Escuela Superior de Ingeniería Mecánica y Eléctrica
Maestría en Ciencias en Ingeniería de Telecomunicaciones, México.
e-mail: misatokatsura2@hotmail.com

Abstract. This paper shows an analysis of the power distribution of channel CPICH in the WCDMA technology; experimental measurements were used to generate coverage maps with different power levels, allowing us to make the corresponding analysis.

Keywords: CPICH, pilot, WCDMA.

1 Introduction

Today, it is known that there are three generations of mobile communications. The first generation was only an analog network and offered services of voice. Because of the growing demand of services, the systems began to develop, first with the second generation that was a digital network and also of voice, this gave us low data transfer speed. However, it does not meet the communication requirements of today and develop to 3G.

Wideband Code Division Multiple Access (WCDMA) is a technology of third generation mobile communications that increases data transmission rates. WCDMA provides new capabilities of service and lower costs for voice and high-speed data (full-motion video, Internet access and videoconferencing). WCDMA is one technology of Spread Spectrum; this technology spreads the signals over a bandwidth of 5 MHz and it is able to carrying voice and data simultaneously. These features allow a data rate of 384 kbps and a transfer rate in the downlink up of 2Mbps and average processing speed of 22 to 320 Kbps. [1]

The organization of channels used by WCDMA is composed of three layers: physical channel, logical channels and transport channels. Of these channels, the logical describe the type of information to be transmitted, the transport channels described as the logical channels are transferred and the physical channels are means of transmission that provide the radio platform through which information is sent.

The Common Pilot Channel (CPICH) transmits a carrier used to estimate the channel parameters. It is the physical reference for other channels. It is used for power control, transmission and coherent detection, channel estimation, measurement of adjacent cells and obtaining the Scrambling Code (SC). The mobiles use this channel as a reference for deciding when to perform a transfer from one base station to another. The mobile terminals performed a transfer when the signal from another base

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 327-334
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

station is above a certain margin of discretion with respect to the received signal. [1-2].

Each base station of WCDMA (Node B) transmits a pilot code; this code makes possible the coherent demodulation, reducing the level of Ec/Io and improves system performance. The pilot code is a pseudorandom sequence, the same for each Node B, but is identified by its sequence of displacement for each Node B. The pilot channel does not carry data, but is used by the user to start the acquisition system, assist in the process of soft handover, synchronization and channel estimation.

2 Measurement methodology

The measurement of the CPICH power is a way of knowing how is distributed the presence of the WCDMA network, the analysis of these power levels; it is possible to determine if the provider's coverage fulfills the customer needs. For a correct analysis of experimental results was necessary to have georeferenced information. To obtain this information, it was necessary to have specialized equipment and generate position information of GPS (Latitude, Longitude, Altitude and Time).

As mentioned, we performed measurements of CPICH power level in the "Lindavista" area. The main focus, was the analysis of the base station "Lindavista" and learns how it is distributed the coverage of WCDMA network of the main operator of the test region. This base station (Node B) with the sectors belonging to the SC 224, 225 and 226 is our central Base Station and platform of our analysis. To perform power measurements, we used a spectrum analyzer, BTS Master MT8222A, Anritsu. Among its features is that it has a GPS receiver, which helped us to locate the measurement point and generate coverage maps georeferenced. In addition, it seems measures the power CPICH of six SC (Scrambling Codes) present at the measured point [3]. The measurements were made by placing the equipment at a height approximately between 1 and 1.5 meters above ground level, simulating the average height at which mobile phone users make use of mobile terminals.

The information of position coordinates was obtained with the spectrum analyzer, it was necessary to configure the equipment and connect the GPS antenna. This feature requires that the equipment get positional information of four satellites for better accuracy.

A total of 1750 measurements distributed along the test site, and use the spectrum analyzer BTS Master MT8222A. In particular, we focus on the power of the CPICH, the analyzer displays in dBm. The distance between measurements was approximately 4 to 5 meters, covering the area of analysis.

With the 1750 point measurements, it was possible to generate coverage maps. The data were processed using software Easy Kriging [4] to implement the Krige geostatistical method (or kriging). . Krige method is an interpolation technique based on regression of samples used to predict unknown values from known values spaced irregularly. The method takes the point values and generates continuous graphic, performing an interpolation of them

It is a geostatistical method that uses a variogram model for data acquisition. The variogram or semivariogram is a tool to analyze the spatial behavior of a variable on a defined area, resulting in the influence of data at different distances.

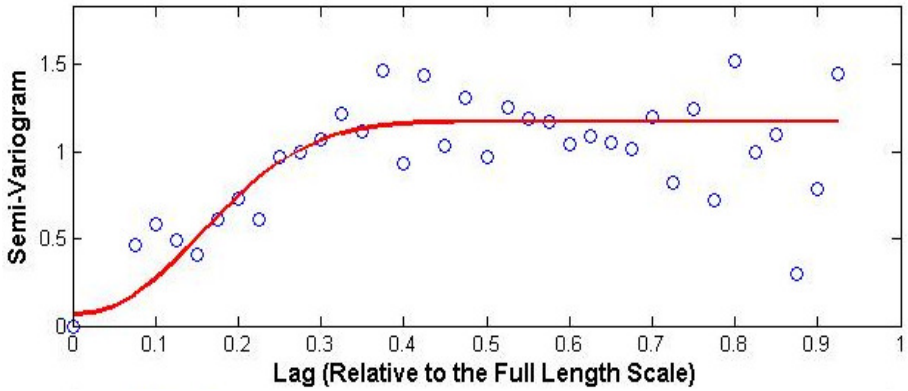


Fig 1. Variogram

Figure 1 shows the variogram generated through of the data, then trying to adjust the variogram, approximating graphics in red on the blue dots; varying parameters model on the right side of the screen. Finally, it generates the coverage map.

2.1 Measurement plan

The first part of the plan of measures was delimiting the area of measurement. This area of work relates to the area covered by the Base Station "Lindavista". **Figure 2** shows the area where measurements were made. In this area we can find different types of scenarios as open area, suburban and urban areas. The model of scenarios open area covers open spaces, without large trees or buildings within the path. In open area the presence of trees is low and building is zero. The suburban area represents some obstacles near of the mobile. This area was chosen because we planed analyze the parameters of the Base Station Lindavista. [3]

2.2 Configuration parameters of WCDMA carrier

The practical parameters of the WCDMA carrier depend of dominant operator in the region of measurement. These operating parameters are:

- Channel: 1087.
- Carrier Frequency: 887.5 MHz.
- Operating Banda: V.

It should be noted that weather conditions affect the values obtained in measurements, so it was necessary that the measurements be made under similar climatic conditions (the sky was not cloudy or partly cloudy, the reason was that the GPS connection is more difficult if the concentration of clouds is large).

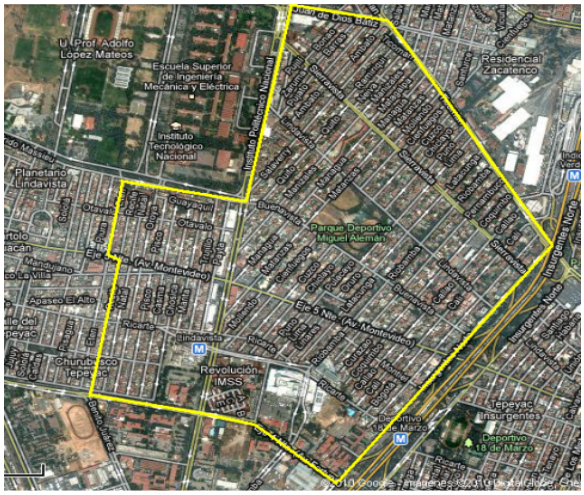


Fig 2. Measurement Area

3 Experimental results

As a first step in obtaining the coverage maps, the experimental results were processed and obtained to a text file to start the interpolation process using the Kriging Easy software. The text file format is shown in Figure 3. The Base Station “Lindavista", presents three SC (Scrambling Codes): 224, 225 and 226. Below is a map (Figure 4) that shows the power levels radiated by each base station sectors of “Lindavista”.

This map allows us to appreciate the coverage areas of each sector. Although the distribution of power is irregular, it does cover with acceptable power levels around the base station. Figure 5 shows the power distribution level of each sector that makes up the base station "Lindavista”. It also shows the color scale used in each of the maps generated.

Whereas the sensitivity of most mobile devices is -80 dBm, can be seen that the levels are acceptable, because in the border areas barely this level is reached.

LONGITUD	LATITUD	CPICH
-99.13472222	19.49777778	-82.553
-99.13472222	19.49777778	-80.167
-99.13472222	19.49777778	-83.676
-99.13472222	19.49805556	-88.068
-99.13472222	19.49833333	-84.118
-99.13472222	19.49833333	-85.757
-99.13472222	19.49833333	-88.932
-99.13472222	19.49861111	-86.631
-99.13472222	19.49861111	-83.616
-99.13472222	19.49888889	-79.743
-99.13472222	19.49916667	-82.326
-99.13472222	19.49944444	-86.039
-99.13472222	19.49972222	-84.787
-99.13472222	19.49972222	-85.6
-99.13472222	19.49972222	-87.987
-99.13472222	19.50027778	-81.967
-99.13472222	19.50027778	-81.457
-99.13472222	19.50083333	-86.443
-99.13472222	19.50083333	-88.63
-99.13472222	19.50111111	-89.686
-99.13472222	19.50111111	-91.257
-99.13472222	19.50166667	-89.88
-99.13472222	19.50166667	-93.867
-99.13472222	19.50166667	-91.075
-99.13472222	19.50166667	-88.946
-99.13472222	19.50194444	-92.528
-99.13444444	19.49583333	-59.089
-99.13444444	19.49583333	-64.401
-99.13444444	19.49583333	-61.553
-99.13444444	19.49583333	-61.553
-99.13444444	19.49583333	-59.87

Fig 3. Text File Format

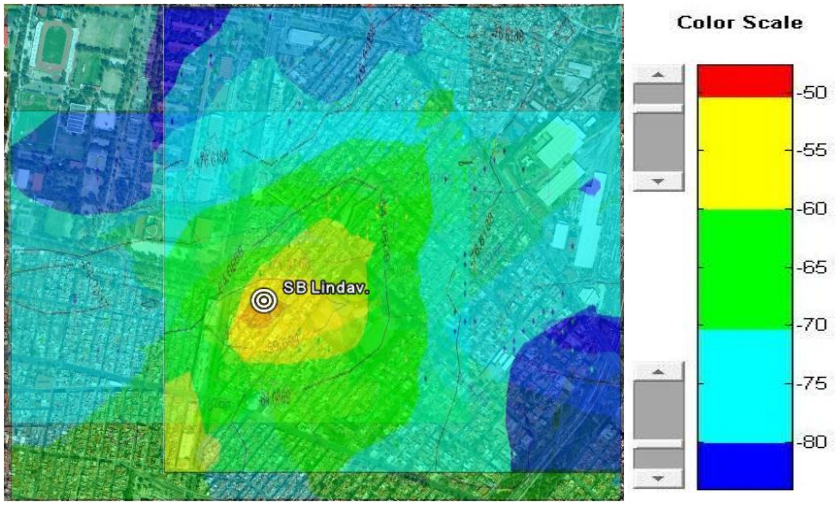


Fig 4. CPICH Power Distribution, “Lindavista” Base Station (dBm)

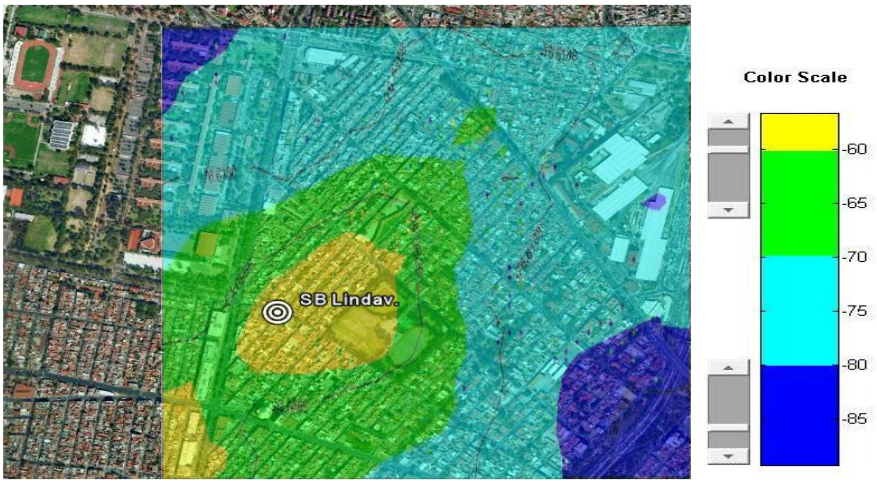


Fig 5. Scrambling Code 224

Fig 6. Scrambling Code 225

4 Conclusions and future work

WCDMA is an idea organized and planned, considering the future, to develop a network that is capable of ensure flexibility in services depending on the needs of users and provide support for service anywhere in the world. WCDMA provides a compromise between coverage and capacity of systems.

The WCDMA network performance can be determined by analyzing several factors, such as: the distribution of CPICH power, multipath, RSCP and the relationship Ec / Io. For purposes of this paper analyzed the distribution of CPICH channel power

The CPICH is a channel that mobile phones used to estimate the fading occurred in the route between the user and mobile.

Fig 7. Scrambling Code 226

The CPICH power level allows us to determine that base station will serve to the mobile terminal, although this doesn't guarantee that the signal that is received is acceptable due to interference. Although the received signal has a power level suitable for that the receiver can recognize the signal, communication can be degraded due to interference, and the transmission rates are reduce.

With the data obtained we can conclude that the base station operates in some respects with minor deficiencies, sometimes the power levels are very high and in some cases very weak. We can also observe a concentration of multiple CPICH pilots in the same area, whose powers are very close together, causing what is known as pilot pollution.

Having a coverage map georeferenced (distribution of radiated power) offers the possibility to know clearly the obstacles that deteriorate the signal quality, areas where the service is received well. The result is to get a general idea about how the Node B is distributing the power of the signal.

Following this line of research, we employ the Krige method to analyze the behavior of the node B under different scenarios to design a propagation model that

considers the particular characteristics of Mexico City for better planning and performance of networks of third generation.

It would be desirable study and analyze the relationship of power and interference, checking the impact that has on the received signals. And also to analyze different scenarios to those presented in this work, for better analysis, considering different materials and environments, and various concentrations of population.

References

1. Tachikawa, Keji, WCDMA Mobile Communications System, John Wiley & Sons, 2002.
2. Laiho Jaana and Wacker Achim, Novosad Tomás, Radio Network Planning and Optimisation for UMTS, Segunda Edición, Editorial John Wiley and Sons, 2006.
3. Ibáñez Gual María Victoria y Simón Vidal Amelia, Modelos Estadisticos Espacio-Temporales en Perimetría, Tesis Doctoral, Departamento de Matemáticas, Escuela Superior de Tecnología y Ciencias Experimentales, Universidad Jaume 1, 2003.
4. Practical Tips on WCDMA Measurements, Application Note No. 11410-00378, Rev. B Printed in United States 2008-09.

Omni-Directional Camera and Fuzzy Logic Path Planner for Autonomous Sailboat Navigation

Miguel Romero, Yan Guo, Sio-Hoi Ieng, Frédéric Plumet,
Ryad Benosman, and Bruno Gas

Institut des Systèmes Intelligents et Robotique, Université Pierre et Marie Curie (CNRS - UMR 7222)
4, place Jussieu 75252 Paris cedex 05, France
firstname.lastname@upmc.fr

Abstract. Autonomous surface vehicles are used to explore, measure and observe many kinds of environments which are often complex and arise many challenges. The ASAROME project [1] (Autonomous SAiling Robot for Oceanographic MEasurements) is focused on an autonomous sailboat to make measurements and observations in marine environments for extended periods. This paper describes a fuzzy inference engine which integrates a routing strategy for obstacle avoidance using an omnidirectional camera for obstacle detection system.

1 Introduction

Operating unmanned autonomous vehicle is a challenging task which is far more difficult than classic mobile autonomous robots because of the outdoor context of application. Most of the autonomous outdoor navigation works have been done for ground and air vehicles but unmanned surface vehicles (USV) have also gathered attention of robotics researchers like [1],[2],[3],[4],[5],[6] for autonomous sailboats. It is usual to equip ground or aerials vehicle with inertial and vision sensors but in the case of surface ones it is more marginal to our knowledge [7]. Active sensors like sonar, radar or even laser are more likely used in surface vehicle for obstacle detection [8].

The ASAROME project fits into this context of autonomous USV, aimed to the design and construction of an autonomous vessel able to carry out long term oceanographic measurement campaign. An accurate and efficient strategy for obstacle detection and navigation is fundamental to preserve the integrity of the ship. Multisensory approaches are usually the solution to the problem as it is shown in several past works [9],[10] [11]. In this project, several on-board sensors will be used for obstacle detection. We will focus on the omnidirectional camera system to build a vision based obstacle detector. Future work will encompass other embedded sensors. Our route determination fuzzy inference system is conceived to optimize the trajectory looking for a favorable wind propulsion through the entire route avoiding unnecessary tack maneuvers and obstacles detected by the panoramic camera.

The paper is organized in several sections. In section 2, we present the calibration method for a catadioptric camera and its use for obstacle detection. In section 3, we

[1] This work is funded under the project ANR ASAROME (Num. ANR-07-ROBO-0009)

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 335-346
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

introduce a fuzzy logic method for local routing and obstacle avoidance. Experimental and simulation results are presented in the last section of this paper.

2 Omnidirectional vision sensor

We are focusing on the vision sensor of the ASAROME embedded system to address the problems of obstacle detection and avoidance. Sensors data are used to feed the navigation command loops but they need firstly to be calibrated correctly, i.e., the mappings between measurements and the real world metric must be estimated.

2.1 Catadioptric sensor calibration

The omnidirectional sensor we use is a non-central catadioptric system, combining a perspective camera with a reflective surface. We have to estimate the relative poses of each component of the sensor.The perspective camera must be calibrated intrinsically using standard techniques described in literature; typically, we used [12] to achieve this task. With the knowledge of the intrinsic the matrix K, we estimate the relative pose of the camera and the mirror with a variation of the method described in [13], based on an homography between two judiciously chosen planes according to [14].

2.2 Boat pose estimation

A global coordinate frame has to be set at first to localize all sensors in relationship with the sailboat, which is defined by the axes of the ship, taken as an ellipse from the images. The boat is first segmented from the background and the covariance matrix of the pixels distribution is computed. Principals axes directions of the ellipse are then given by the matrix eigenvectors (see Fig. 1). Any structure or scene object will now be referenced within this frame for the rest of the paper.

With the sensor calibrated as described and placed on the the top of the mast, it is possible to compute each pixel's projection on the plane defined by the water under the assumption of a calm sea state, if the relative pose of this plane and the vision sensor (thus the boat) can be estimated. This operation is required if we want to extract metric information from the images without the need of stereovision. In the similar way, as we did for the camera pose with respect to the mirror estimation, we can estimate the omnidirectional sensor pose by detecting projections of the sea plane on the catadioptric image. To achieve this task we place into the water several easy-to-detect buoys that are acting as seamarks to underline the sea plane.

If the structure defined by the buoys is known, i.e., their relative distances of the buoys are known, then enough geometric constraints can be provided in addition to the coplanarity one for the plane estimation. A standard parametrization of a plane we can use is:

$$ax + by + cz + d = 0 \tag{1}$$

Given n buoys used to built the seamark, the distance between the i^{th} and the j^{th} buoys, D_{ij} is manually measured. There are $m = \binom{n}{2} = \frac{n!}{2!(n-2)!}$ of such distances and

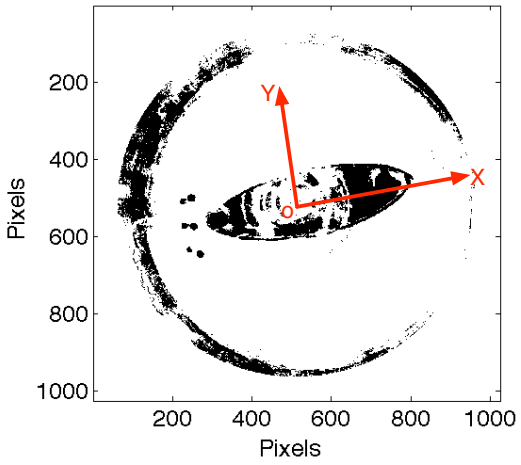


Fig. 1: The ship's axes are used to define a global coordinate frame assuming the camera is placed perpendicular to it.

m is higher than the number of parameters for $n \geq 4$ (in our case $n = 6$). The buoys projected on the camera are also reduced to their centroids $\mathbf{p}_i$ in the image plane and, with the sensor calibration, n 3D lines $(\mathbf{m}_i, \mathbf{u}_i)$ are computed for the n centroids. There is likely only one plane under the distances constraints that intersects a given bundle of rays which in our case is the set of all rays associated to the buoys. We use a Hough-based method for the plane decision. Where a, b, c are precomputed from -1 to 1 with a constant step of 0.002 and d is calculated from 1 to 500 with a constant step of 1. We define π_k, the k^{th} plane that intersects the ray bundle. Each ray intersects the plane and the distances between every two intersections are computed:

$$D_{ij}^k = |\mathbf{P}_i^k - \mathbf{P}_j^k| \tag{2}$$

Where $\mathbf{P}_i^k$ and $\mathbf{P}_j^k$ are the positions of intersections for the point i and j. If π_k is the plane that we are looking for, it will satisfy the condition $D_{ij} \approx D_{ij}^k$. Given the possibility to determine the camera pose with respect to the sea surface, we can map any pixel $\mathbf{p}_i$ to a 3D point $\mathbf{P}_i$ on the plane (Fig. 2). The distance of each pixel to the boat in the image can be computed as: $d = |\mathbf{p}_i|$ in the global coordinate frame introduced earlier. This latter result allows to built a resolution map of the sensor on the sea surface since we can project all the pixels on it.The plane pose estimation method is tested with synthetic data by randomly generating 3 planes defined by their four parameters. For each plane, 6 points are placed randomly on it, assuming that their relative distances with respect to each other point are known. For each of this setting, we execute the plane detection algorithm. This operation is repeated 100 times to provide enough results to produce meaningful statistics. The table 1 summarizes the results obtained. According to the result, the plane estimation algorithm accuracy is reasonable with estimates errors below 7% and a mean value of 3.33%.

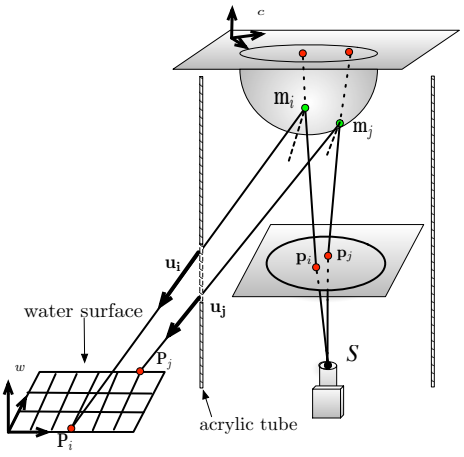


Fig. 2: The mapping of any image pixels for the catadioptric sensor and its intersection with the surface defined by the water.

Table 1: Result of estimated parameters

	Parameters estimation error rates(%)			
Parameters	a	b	c	d
Case 1	4	1	1	3.3
Case 2	1	0	0	0
Case 3	1	2	0	6.7

The resolution map is built with the mast normal to the water plane (calm sea state), however an oscillating sea is more likely expected, implying to recomputed the resolution map for different orientations of the ship with respect to the sea surface. To avoid to run constantly the costly construction, a set of maps is precomputed for several values of orientations one time. Each time the ship detects a significant change of orientation, the correct map is loaded into the computer. The maps are precomputed for angle α spanning from -90 to 90° with a constant step of 1.80°.

2.3 Obstacle detection

A colorimetric criterion is used to segment objects in the image. To be able to detect an obstacle, a color signature of the sea is first computed by selecting samples of regions representing the water. Each pixel of these regions are reprojected to the three planes of the RGB coordinate frame according to their color components. This operation produces a cluster in each plane and is representative of the color distribution of the background (i.e sea).

To segment an object pixel from the background, we project it into the RGB components planes and it is considered being an obstacle if at least one of its RGB component

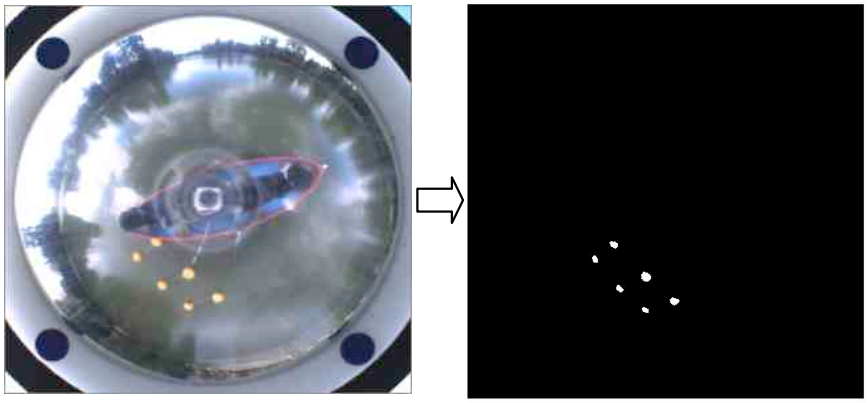


Fig. 3: Seamark detections using colorimetric signature.

does not belong to the clusters defined earlier. The Fig. 3 show an example of segmentation of a seamark formed by six yellow buoys using the combination of image difference and colorimetric signature of the water.

Detected obstacles are usually sets of pixels $\mathbf{P}^i_{obs}(x_i, y_i)$ from which the centroids $\mathbf{P}_c(x_c, y_c)$ are computed. We can also estimate roughly the size of a detected osbstacle as being the maximal distance between two pixels of the set:

$$l_{obs} = max(|\mathbf{P}^i_{obs} - \mathbf{P}^j_{obs}|) \quad \text{for } i \neq j. \tag{3}$$

With the obstacles properly detected and segmented, their positions and their respective size are parsed to the heading calculation.

3 Fuzzy Inference System for Local Routing Strategy and Obstacle Avoidance

In sailing, there are no precisely defined criterions to navigate. This situation gives to the *skeeper* the hard task to take decisions based in his own experience. Then, as Zadeh describe in [15], the problem of sailing is able to be resolved using fuzzy logic.

There are multiple approaches in literature to solve the routing problem for a sailboat using fuzzy logic as [16], [17], [18]. Also there are several papers that describes obstacle avoidance using fuzzy inference engines as [19], [20]. In this section we describe a mixed approach that uses a fuzzy inference engines to solve our sailboat navigation and obstacle avoidance problem.

For a sailboat, the propulsive force comes from the aerodynamics effects of the wind on the sails. This leads to the conventional points of sail diagram which describes a sailing boat's course with respect to the wind direction (Fig. 4). The white sectors

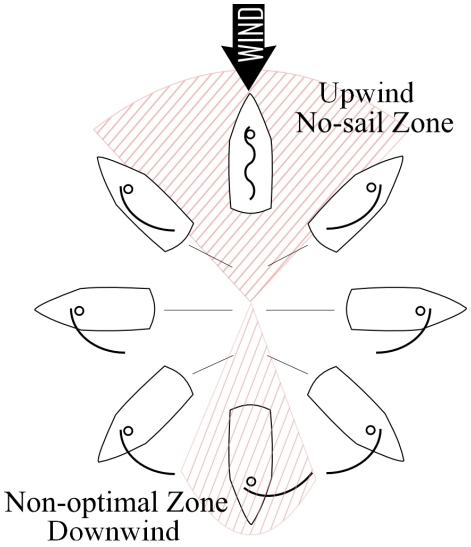


Fig. 4: Standard points of sail diagram

correspond to normal sailing zones and the two shaded zones are not sailable zones or no-go zones (in fact, the downwind zone could be sailed in theory but is not very efficient and rather unstable).

So, unlike conventional motorized robotic vehicles, where a straight line to the goal leads to a shortest path (in time and distance), in this case there is no such easy solution for a sailboat if the target is located directly upwind or downwind. In these cases, the sailboat has to beat (i.e., take a zig-zag course) to reach the goal.

The speed vector of a sailboat depends on many factors like the wind angle, wind speed, sail trimming, currents and waves. This behavior is usually represented for a given boat by a specific polar diagram (Fig. 5). This polar diagram shows the maximum boat speed along a given heading with respect to the wind. Each curve on the polar diagram corresponds to a given wind speed. As we can see on Fig. 5, the wind speed mainly modifies the amplitude without modifying the global shape of the polar diagram.

In the following, we suppose that this boat-specific polar diagram is known for a set of wind speeds.

The proposed method for local routing and obstacle avoidance is based on the calculation of an optimal heading that maximize the crisp output of the Mamdani type fuzzy inference system (FIS) . This FIS have as entries the speed to the objective and the obstacle influence value to tend to minimize the time to reach the goal and maximize the distance between the sailboat and the obstacles. This method will react to the changing of the environmental conditions (wind speed, wind angles,...) by periodically re-computing the optimal heading.

As a measure of the boat efficiency to reach the goal, we take the value of the boat speed vector $\mathbf{V}(h)$ for a given heading angle h, projected in the direction of the waypoint

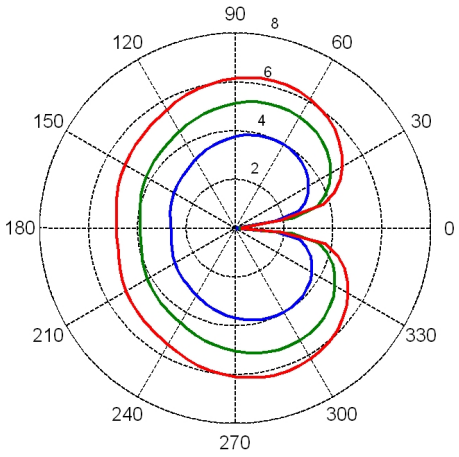


Fig. 5: Boat-specific polar diagram.

WP, that is:

$$V_G = \mathbf{V}(h)^T \cdot \mathbf{T_g} \tag{4}$$

where $\mathbf{T_g}$ is an unit vector pointing to the way point. In order to conveniently use this measure in our routing algorithm, the boat velocity is supposed to be normalized, i.e, the maximum value of the boat speed on a given polar diagram is supposed to be equal to 1, regardless of the wind speed value.

Steering through the eye of the wind, that is, into and across the flow of the wind, is usually a maneuver that must be avoided due to the speed decreasing and, in certain situations, the risks of rollover due to a wind shift. To take these two facts into consideration, we use a penalty factor η_w for the computation of the cost function to minimize the time to reach the goal:

$$C_W = \eta_w (V_G) \tag{5}$$

with $\eta_w = 1$ if the actual heading and the new computed heading are in the same side of the wind and $\eta_w = 0.8$ otherwise. Such a penalty factor leads to prioritize a new heading that keeps the course in the same side of the wind, rather than crossing the eye of the wind.

To take into account the obstacles detected by the perception system, we use a cost function based on the measured distance between the boat and an obstacle's centroid:

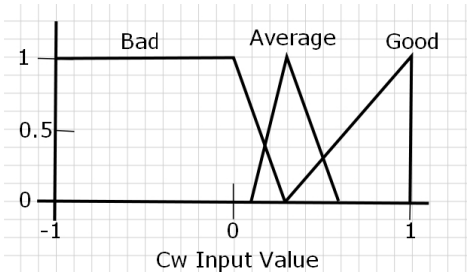
$$C_O = \begin{cases} \eta_o \left(\frac{1}{d_{obs}} - \frac{1}{d_0} \right) & if \;\; d_{obs} \leq d_0 \\ 0 & if \;\; d_{obs} > d_0 \end{cases} \tag{6}$$

with η_o as positive scaling factor, d_{obs} as the Euclidean distance from the center of the sailboat to the center of the obstacle and d_0 as the obstacle influence distance (50 m in our case, which is a trade-off between the range of our sensors and the maneuverability of the sailboat). The final cost function is:

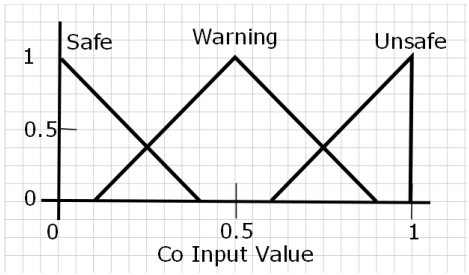
$$C = C_W + \sum_{obs} C_0 \tag{7}$$

Input data for the FIS are the weighted velocity to goal C_w and the obstacle influence value, C_o, given by the equations 6 and 6. The fuzzy sets representing the linguistic variables are as in Fig6, where negative values of C_w represents the heading angles that stray from the goal .

In course control and obstacle avoidance the C_w and the C_o parameters are net to fuzzy controller, then, the acceptance value of the h heading angle is inferred through the fuzzy logic.



(a)



(b)

Fig. 6: Fuzzy system for heading selection

The output variable is the value of pertinence for the tested h heading angle. The fuzzy variable contains tree fuzzy sets that defines this pertinence as bad, average and good, as in Fig7.

The rule base of the heading selection FIS contains 9 rules of the form:

```
If desired direction IS x AND obstacle is y THEN
pertinence is z
```

In order to avoid undesirable trajectory angles(Fig. 4) and to keep the course aiming the goal, we introduce three simple fuzzy rules:

1. *If(C_w is Good) then (Output is Ok);*
2. *If(C_w is Average) then (Output is Average);*
3. *If(C_w is Bad) then (Output is Bad);*

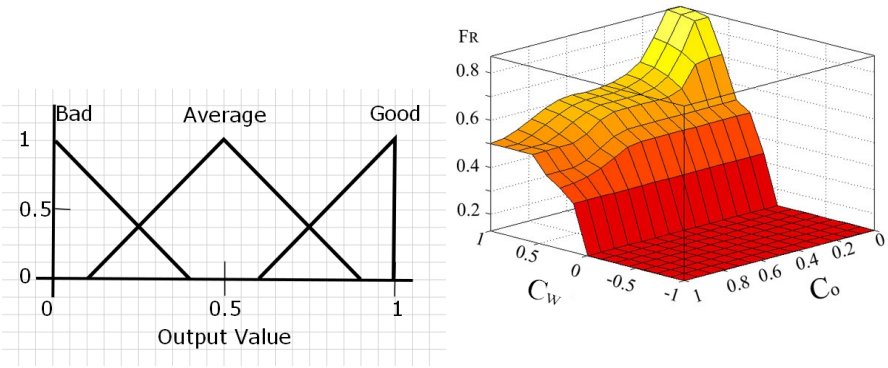
As we can observe in Fig.6a, the heading angle h that gives C_w values ≥ 0.2, which are classified with the linguistic variables *"Average"* or *"Good"*, are the only heading angles that will be considered, since they are the only ones that assure wind propulsion and a path towards the goal.

To keep the sailboat away from the obstacles we must to select a set of fuzzy rules based on the C_o value (Eq.7), then, as is shown in Fig.6b, lower is the value of C_o, safest is the proposed heading angle. From this statement we can infer the next rules:

4. *If(C_w is Good) and (C_o is Safe) then (Output is Ok);*
5. *If(C_w is Good) and (C_o is Warning) then (Output is Average);*
6. *If(C_w is Good) and (C_o is Unsafe) then (Output is Bad);*
7. *If(C_w is Average) and (C_o is Safe) then (Output is Average);*
8. *If(C_w is Average) and (C_o is Warning) then (Output is Average);*
9. *If(C_w is Average) and (C_o is Unsafe) then (Output is Bad).*

This set of nine fuzzy rules assures that the selected heading angle h will minimize Eq.7 taking the sailboat to the goal trough an obstacle free path (Figs. 7, 8).

This FIS(h) is computed periodically (0.5 sec) and the optimal heading angle maximizing this inference system is sent to the low level layer control. Choosing the angle of heading, h, which maximize the output of FIS(h) allows the sailboat to be able to navigate while keeping its course outside of the polar's forbidden area (Fig.4), taking the shortest navigable path and avoiding obstacles. Using this method, there is in fact no guarantee that the system will not be caught in a local minimum. However, since our autonomous sailboat is intended to move in open sea, the probability to have more than one obstacle at a time in the sensors range is actually very low.



(a) Output (b) Fuzzy Surface

Fig. 7: Output fuzzy sets and surface

4 Experimentation

We demonstrate the performance of our algorithm with two different tests: obstacle detection and simulation for the trajectory planning. The first experimentation have been done on a lake, with all the data being recorded during the test. The second one is a simulation, which shows the resulting motion of the sailboat's with and without obstacles.

4.1 Real data analysis for obstacle's detection

We conducted this experimentation on a calm lake with its surface defining a plane. A special seamark with 6 buoys (1m long x 2m wide) is used for this test. Our camera tube is fixed on a 1.9m high tripod.

The seamark position is changed between each record and the distance between the camera and the center of the seamark was manually measured. The distance tested ranges from 2 meter to 14 meter. From the images which are taken by the camera panoramic, we can estimate the length and the width of seamark, also the distance between seamark and our boat.

From table 2, our algorithm seemed to have excellent performance with error rates within 8%. The distances estimated are smaller than the measured distances. The maximum value of estimated error is 0.63m for a distance of 14m, and the minimum is 0.19m for a distance of 8m. The estimated lengths are arround of 1 meter, and the estimated widths are arround of 2 meters. Like the estimated distance error, the maximum size estimated error is 7.40% for a distance of 14m, and the minimum is 2.32% for a distance of 4 m.

Table 2: Real data error estimation

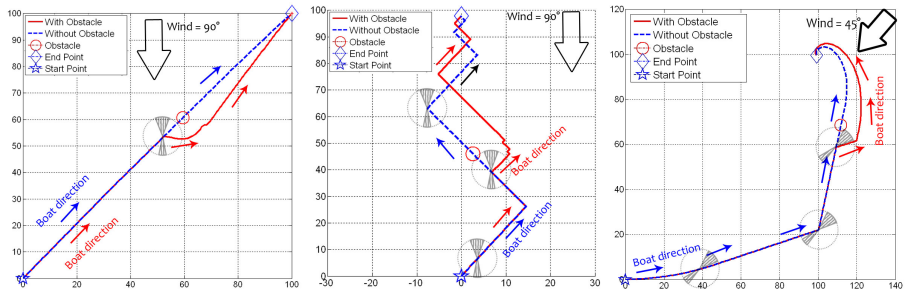
	Distance mesured(m)	Esti. dis.(m)	Esti. angle(°)	Esti. Dis. err. rate(%)	Esti. length(m)	Esti. width(m)	Esti. l+w err. rate(%)
Case 1	2	1.70	259	3.75	0.95	1.94	3.71
Case 2	4	3.72	201	3.5	0.83	2.10	2.32
Case 3	6	5.60	207	5	0.83	2.03	4.67
Case 4	8	7.81	199	2.38	0.91	1.95	5.12
Case 5	10	9.61	186	4.87	0.79	1.99	7.02
Case 6	12	11.44	208	7	0.8	2.09	3.47
Case 7	14	13.37	187	7.87	0.66	2.12	7.40

4.2 Simulation Results

Extensive simulations were conducted of sailboat traveling in different wind directions, with different goal positions and with or without obstacles.

Some representative simulations results are presented in this paper. Fig. 8(a) shows the sailboat's motion for a lateral wind. Figs. 8(b) and 8(c) shows two upwind navigation cases. Each test was conducted two times with the same conditions, except for the obstacle. For all simulation scenarios the wind speed was($15\,Kt$) and remains constant during the simulation time, the wind angle is ($90°$) for the first two cases and ($45°$) for the last one. The initial heading is $0°$, the start position is at $(0,0)$ and the goal positions are $(100,100)$ for the cases depicted on figs. 8(a) and 8(c), and $(0,100)$ for the fig. 8(b) . Those conditions were chosen to show the behavior in a normal navigation situation and also, in cases (b) and (c), to force the sailboat to take an upwind trajectory.

We can see on Fig. 8 that, in all cases, the sailboat reaches the waypoint. In the first case (without obstacle), the sailboat will go straight to goal, in the second try (with obstacle) it turns with out crossing the eye of the wind, to avoid the obstacle. In the second case it have to tack several times to avoid the obstacle and reach the goal. Finally, in the third case, the sailboat tack one time, and then it keeps the wind on its right side.



(a) TWA = 90° (b) TWA = 90° (c) TWA = 45°

Fig. 8: Simulation results with different TWA and different goal position

5 Conclusion

The experimental results demonstrate the ability of panoramic vision system to obtain and process the data to give a precise position and size of the detected obstacles. The conducted simulations validate the routing algorithm because, with the data from omnidirectional camera system, it is possible to modify the motion of the sailboat to avoid the collisions, keep it on the correct wind's angle and reach the fixed waypoint. The current limitation of the presented routing method is the risk to fall in a local minimum. But, as outlined before, in open sea: the probability to have more than one obstacle at a time in the sensors range is actually very low.

The future work in a short term is aimed to perform several tests implying all the sensor for data fusion to validate the whole system in open sea.

References

1. R. Stelzer, K. Jafarmadar, H. Hassler, and R. Charwot, "A reactive approach to obstacle avoidance in autonomous sailing." in *3rd International Robotic Sailing Conference (IRSC)*, 2010.
2. Y. Briere, "Between the drifting buoy and the autonomous sailing boat : the microtransat concept," in *ATMA - ASSS 2008 international autonomous surface ship symposium*, France, 2008.
3. C. Sauze and M. Neal, "A raycast approach to collision avoidance in sailing robots," in *Proceedings of the 3rd International Robotic Sailing Conference*, 2010.
4. G. H. Elkaim, "The atlantis project: A gps-guided wing-sailed autonomous catamaran," *Journal of the Institute of Navigation*, vol. 53, No. 4, pp. 237–247, 2006.
5. J. Sliwka and L. Jaulin, "Autonomous robotic boat of ensieta." International robotic sailing conference, july 2009, pp. 1–7.
6. H. Erckebs, G.-A. Büsser, C. Pradalier, and R. Siegwart, "Avalon, navigation strategy and trajectory following controller for an autonomous sailing vessel," *IEEE Robotics and Automation Magazine*, vol. 17, No 1, pp. 45–54, 2010.
7. A. Subramanian, X. Gong, J. Riggins, D. Stilwell, and C. Wyatt, "Shoreline mapping using an omni-directional camera for autonomous surface vehicle applications," in *OCEANS 2006*, 2006, pp. 1–6.
8. T. Bandyophadyay, L. Sarcione, and F. Hover, "A simple reactive obstacle avoidance algorithm and its application in singapore harbor." in *International Conference on Field and Service Robotics*, 2009.
9. A. Bonci, G. Ippoliti, L. Jetto, T. Leo, and S. Longhi, "Methods and algorithms for sensor data fusion aimed at improving the autonomy of a mobile robot," *in Advances in Control of Articulated and Mobile Robots*, vol. 10, pp. 192–222, 2004.
10. G. Ippoliti, L. Jetto, A. la Manna, and S. Longhi, "Improving the robustness properties of robot localization procedures with respect to environment features uncertainties," *IEEE International Conference on Robotics and Automation*, pp. 1451 – 1458, 2005.
11. S. Giompapa, F. Gini, A. Farina, A. Graziano, R. Croci, and R. Distefano, "Maritime border control multisensor system," in *Aerospace and Electronic Systems Magazine*, 2009, pp. 9–15.
12. J.-Y. Bouguet, *Camera Calibration Toolbox for Matlab*, http://www.vision.caltech.edu/bouguetj/calib_doc/index.html.
13. J. Fabrizio, P. Tarel, and R. Benosman, "Calibration of panoramic catadioptric sensors made easier," in *Proceedings of the IEEE Workshop on Omnidirectionnal Vision*, june 2002.
14. P. Sturm, "Algorithms for plane based pose estimation," in *Proceedings of the IEEE International conference on Computer Vision and Pattern Recognition*, 2000, pp. 706–711.
15. L. Zadeh, "Fuzzy sets," *Department of Electrical Engineering and Electronics Research Laboratory, University of California*, vol. 8, pp. 338 – 353, 1965.
16. R. Stelzer and T. Pröll, "Fuzzy logic control system for autonomous sailboats," *Robotics and Autonomous Systems*, 2007.
17. O. CALVO, "Fuzzy control of a sailboat," *International Journal of Approximative Reasoning*, vol. 16, pp. 359–375, 1997.
18. E. Yeh and J.-C. Bin, "Fuzzy control for self-steering of a sailboat," *Intelligent Control and Instrumentation*, vol. 2, pp. 1339 – 1344, 1992.
19. Y. Jincong, Z. Xiuping, and et.al, "Intelligent robot obstacle avoidance system based on fuzzy control," *The 1st International Conference on Information Science and Engineering*, pp. 3812–3815, 2009.
20. R. Malhotra and A. Sarkar, "Development of a fuzzy logic based mobile robot for dynamic obstacle avoidance and goal acquisition in an unstructured environment," *International Conference on Advanced Intelligent Mechatronics*, pp. 1198–1203, 2005.

A Review on the Detection and Removal of Shadows in Daytime Traffic Images

Sandra Luz Canchola-Magdaleno[1], Carlos Pedraza-Ortega[1], Emilio Vargas-Soto[1], Saúl Tovar-Arriaga[1], Juan Manuel Ramos-Arreguín[1], Marco Aceves-Fernández[1], J. Salvador Hernández-Valerio[1], Hugo Ávila-Mendoza[1]

[1] UAQ, Facultad de Informática, Querétaro, México.

sandra.canchola@uaq.mx, caryoko@yahoo.com, emilio@mecatronica.net, saulotovar@hotmail.com, juan.ramos@uaq.mx, marco.aceves@uaq.mx, valerio@uaq.mx, hugoam13@hotmail.com.

Abstract. Segmentation of moving objects in a video sequence is a challenging issue when the images are acquired in an outdoor scene at daytime. Shadows are often detected along with their respective moving objects that make them more difficult to separate in subsequent processes. But, when shadows are detected and removed from moving blocks, the computer vision algorithms can be applied more accurately. We present and compare the most suitable approaches for detecting shadows in an outdoor traffic scene at daytime.

Keywords: Shadow detection, Traffic images, Segmentation.

1 Introduction

A common problem in traffic applications is that shadows are detected along with their respective moving objects. Shadows can affect the shape and color of the objects and they can even merge areas, therefore the presence of shadows has a negative effect on scene analysis and interpretation systems [1]. In the other hand, with a shadow-free image better outcomes can be achieved during object segmentation, tracking, recognition, understanding scene, etc. Then, it is more desirable to separate objects from its shadings areas during the first processing step. During the past decades many attention was paid to the area of shadow detection and removal applied to specific application such as traffic surveillance [2] [3], face recognition [4] and image segmentation [5].

1.1 Shadows

Shadows are important phenomena when a vision-based computer system works with day-time images. According to the classification reported in [6], shadows are composed of two parts: self-shadows and cast shadows. The first is the section of the

[1] Av. de las Ciencias S/N, Juriquilla, C.P. 76230, Querétaro, México. Tel.: +52-442-192-1200 Ext. 5958; Fax: +52-442-192-1332.

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 347-354
Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

object that is not illuminated by the light source, and the last is the area projected on the scene by the object. And cast shadows can be classified in umbra y penumbra, this is the totally and partially blocked area, respectively. In Fig. 1 a typical car is presented with shadows. In outdoor scenes, the movement of the sun and passing clouds are responsible of changing light conditions. The sun causes a slow, systematic, variation in the intensity and the direction of the received illumination, and, finally, it defines the shape, size and directions of shadows. In Fig. 2 a vehicular scene with shadows is presented during a day long.

Fig. 1 Shadows on a vehicle. The self shadow is denoted with a dashed white line (…). The cast shadow is located under and beside the car. The umbra is denoted with type 2 line (.-.-) .

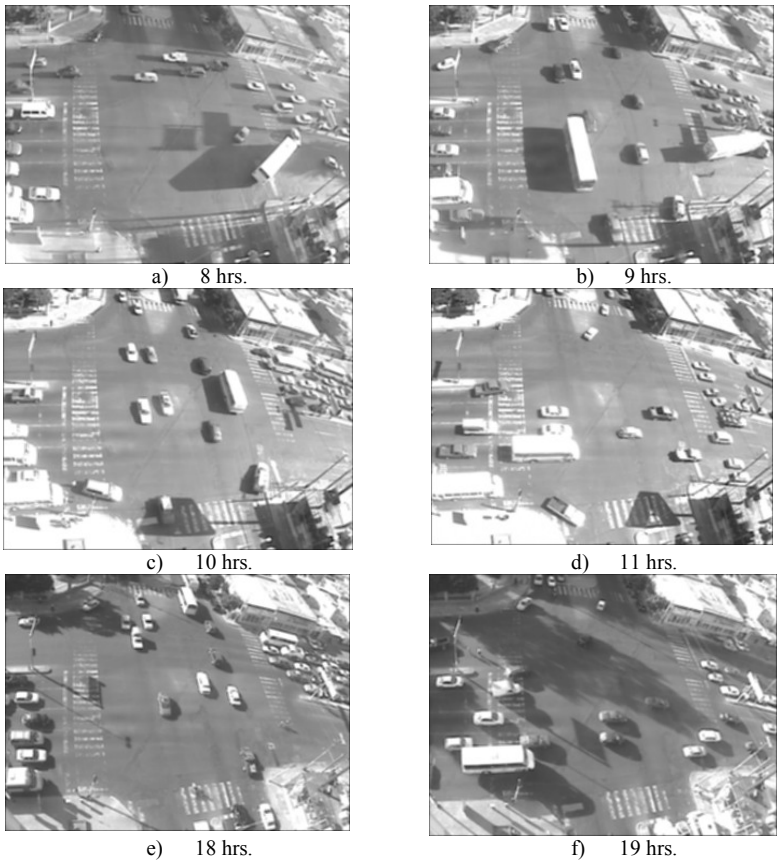


Fig. 2. Vehicular scene with shadows generated by illumination conditions during a day long.

1.2 Traffic applications

In the context of Intelligent Transportation System (ITS), vision-based systems for traffic analysis are developed to help traffic flow management, some examples include speed measurement [7] [8], classifying and counting vehicles [9] [10], etc. The analysis of the events taking place in a crossroads offers the opportunity to avoid harmful situations and the potential to increase security. Then, the challenge is developing an automatic visual system that reasons about the moving vehicles being observed and extracts high-level information [11], useful for traffic monitoring and detection of unusual activity. Unfortunately, computer vision is not massively applied in traffic monitoring applications because existing systems still suffer from poor reliability, high cost and unbalanced accuracy[12]. Their accuracy partially depends on weather conditions: fog, snow, rain; but also by illumination conditions in the scene. Then, to achieve a reliable high level understanding of a traffic scene, the image acquired has to be pre-processed in order to get moving objects segmented as the first stage of processing. In Fig. 3 an example of a traffic image with shadow is shown.

a) b)

Fig. 3. Problems generated in day-time images with shadows. a) Some vehicles are merged; b) Shadows change the color of the vehicles.

2 Shadow detection algorithms

Recognizing shadows in an image is generally a hard task. In vision systems, cast shadow detection is important part of the preprocessing. In this section we present algorithms suitable to apply in traffic scenes.

2.1 Gray-scale video sequences

2.1.1 Multi-gradient shadow identification (MGSI)

In [3] a shadow identification algorithm is presented. First, the ratio between the image and its background is computed, this is,

$$D(i,j) = \frac{B(i,j)}{I(i,j)}k \tag{1}$$

where $B(i,j)$ is the value of the background, $I(i,j)$ is the current image and k is a prefixed factor. Later, a threshold is computed for D.

$$R = Th_m^M(D) \tag{2}$$

where $m < M$ are the threshold values. Then, a multi-gradient analysis is calculated using convolution kernels for detection in three directions, Fig. 4. Finally, the shadow image is the threshold applied to the absolute addition of gradients.

+1	-1
0	0

G_{x_1}

0	0
-1	+1

G_{x_2}

+1	0
-1	0

G_{y_1}

0	-1
0	+1

G_{y_2}

+1	0
0	-1

G_{d_1}

0	+1
-1	0

G_{d_2}

a) b) c)

Fig. 4. Convolution kernels for a) vertical b) horizontal, and c) diagonal edge detection.

2.1.2 Normalized cross-correlation shadow identification (NCCSI)

The algorithm reported in [13] is based in the Normalized cross-correlation (NCC), that is calculated using:

$$NCC(i,j) = \frac{ER(i,j)}{E_B(i,j)E_T(i,j)} \tag{3}$$

where (i,j) is the pixel in this position, $ER(i,j)$ is the value of the pixel by its background value, $E_B(i,j)$ is the mean square root of the background, and $E_T(i,j)$ is the mean square root of the image, the neighborhood is considered for each value. Then, the classifier for shadow pixel is:

$$S(i,j) = \begin{cases} If\ \left((NCC(i,j) \geq \propto) \wedge \left(E_T(i,j) < E_B(i,j)\right)\right), & 1 \\ otherwise, & 0 \end{cases} \tag{4}$$

where $\propto$ is a threshold value. Finally, the shadow refinement is achieved by the analysis of the ration between the image and its background.

2.1.3 Adaptive shadow segmentation (ADSS)

A real-time traffic monitoring system is developed by [12]. Initially, the image differences (D) is obtained between current frame and the last background, then every pixel is classified using give formula,

$$U(i,j) = \begin{cases} If\ D(j,j) > T_b, & BRIGHT \\ If\ D(j,j) < T_d, & DARK \\ otherwise, & BACKGROUND \end{cases} \tag{5}$$

where $T_b$and T_d are the threshold value for pixel classification in bright and dark set, respectively. Then, shadows are identified a dark area connected to bright area that share same direction.

2.1.4 Edge-based moving shadow removal (EBMSR)

In [1] is presented an algorithm for removing moving shadows based on detection of edges when Sobel procedure has been applied. It is assumed that moving objects (C_t) (called foreground) and edges (E_t) were calculated previously. The initial seed region is acquired by:

$$IS_t = C_t - DB_t - IF_t \tag{6}$$

where t means the t-th frame, DB_t the dilated boundary of change mask and IF_t contains the interior regions of foregrounds. Finally, this seed region is analyzed to obtain the minimum rectangle that fits each group of foregrounds.

2.2 Color video sequence

2.2.1 Sakbot Shadow Detection

The Statistical and Knowledge-Based Object Tracker (Sakbot) [14] works in the HSV color space that corresponds closely to the human perception of color. It uses a classifier to define whether a pixel belongs to a shadow, that is,

$$SP_t(i,j) = \begin{cases} if \quad \left(\propto \leq \dfrac{I_t^V(i,j)}{B_t^V(i,j)} \leq \beta\right) \\ \wedge\left((I_t^S(i,j) - B_t^S(i,j)) \leq \tau_S\right) \\ \wedge(|I_t^H(i,j) - B_t^H(i,j)| \leq \tau_H), \quad 1 \\ otherwise, \qquad\qquad 0 \end{cases} \tag{7}$$

where $I_t^n(i,j)$ is the pixel corresponding to the n-th channel, $B_t^n(i,j)$ is the background pixel to the n-th channel given the t-th frame, and τ_S y τ_H are threshold value defined.

3. Performance results

For experimentation, we worked off-line with 282 images taken in a heavily transited crossroads in the city of Querétaro, México. The camera was placed on top of a tower, about 28 m above ground level. The computer had an AMD Turion processor running at 2.1GHz, 4GB of internal RAM, and a Matrox Corona II frame grabber. The computer programs were written in Matlab 2009 to process gray scale images with a resolution of 320 columns times 240 rows. Sakbot is the only method that could not be implemented, because it works with HSV color space images. The results are presented in Table 1. MGSI has the best performance, what makes especially suitable its implementation on-line.

For implementation some values was fixed in every method. In MGSI we set $k = 1$, $m = 0$ and $M = 1.5$. In NCCSI $\propto = 0.9995$. In ADSS we had $T_b = 100$ and $T_d = -50$. Finally, in EDMSR we set $\lambda = 2$ for the structuring element in morphological operations.

4. Comparison

Surveillance traffic applications have strong constraints because they must work on real-time. Five different strategies suitable for this kind of systems have been presented in previous section; in Table 2 they are compared. It is remarkable that most of them exploit two aspects; the first is the relation between the n-th image and its background. And the second is edge information about the scene; based on the fact that the vehicle has significant edges, however cast shadows are edgeless [1]. We need to apply a robust shadow elimination method to improve performance of a vision-based traffic monitoring system. In the future, we will apply this present methods in real video sequences on real-time at daytime, and develop an exhaustive comparison between them. Obviously, a new method of shadow elimination and removal will be proposed.

Table 1. Performance of presented methods: a) time elapse to process 282 images b) mean time to process each image.

Method	a)	b)
MGSI	1 min 25 s.	0.2295 s.
NCCSI	8 min 56 s.	1.9017 s.
ADSS	4 min 17 s.	0.9135 s.
EDMSR	5 min 18 s.	1.1271 s.

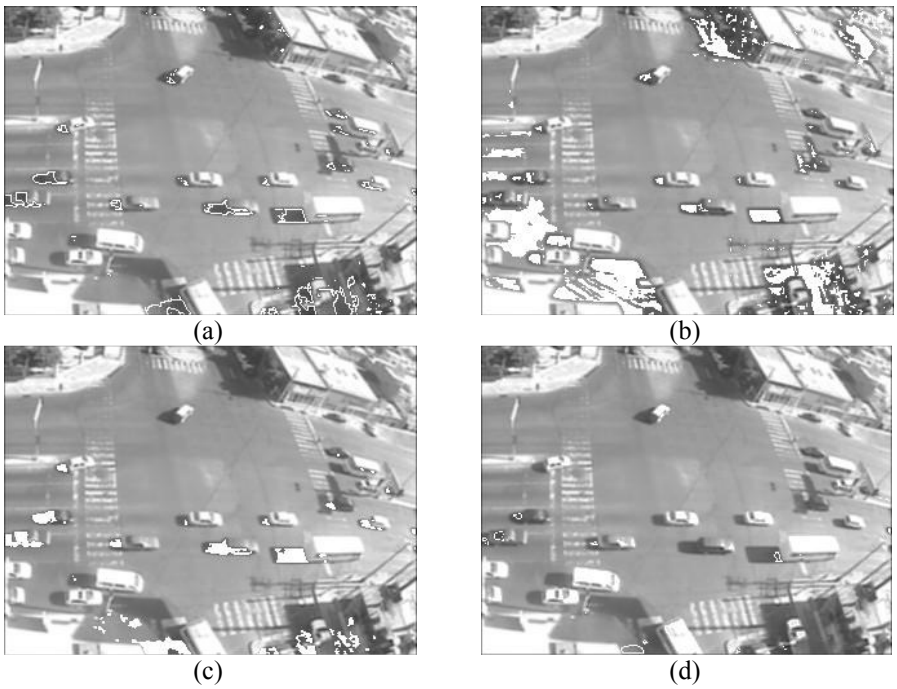


Fig. 5. Examples of images with detection of shadows applying: (a) MGSI (b) NCCSI (c) ADSS y (d) EDMSR. Shadows are denoted with white pixels.

Table 2. Comparison of presented methods: (a) relationship between current image and its background and (b) border information obtained by.

Method	(a)	(b)
MGSI	Ratio	Vertical, horizontal and diagonal edge detection
NCCSI	Ration and cross-correlation	Classification
ADSS	Difference	Proximity
EDMSR	None	Sobel, vertical, horizontal operation
SAKBOT	Ration and difference	Classification

Conclusion

In this paper a comparison between methods suitable to apply in surveillance vehicle systems was presented. In this kind of systems some aspect related to nature scene influence what is captured in images. Shadows specially are present in day-time images, because they are generated by illumination (sun, clouds, rain, etc.). Besides, a surveillance system has to work on-line with images acquired, and it makes more desirable that image processing is a fast process. In the present implementation, we showed that MGSI has the best performance and it is suitable to work on-line as a part of the pre-processing stage. We believe that one new method is feasible with an advised level of performance.

Acknowledgement

We want to acknowledge the financial support of this work through the PROMEP-SEP (UAQ-PTC-162) project.

References

1. *Moving Shadow Detection and Removal for Traffic Sequences.* Xiao, Mei, Han, Chong-Zhao y Zhang, Lei. 2007, International Journal of Automation and Computing, págs. 38-46.
2. *Shadow detection and removal for traffic images.* Wang, J.M., y otros. Taiwan : IEEE, 2004. International Conference on Networking, Sensing and Control. págs. 649 - 654 .
3. *Effective shadow detection in traffic monitoring applications.* Bevilacqua, Alessandro. 2003, Journal of WSCG, Vol. 11, pág. República Checa.
4. *Illumination Normalization for Face Recognition and Uneven Background Correction Using Total Variation Based Image Models.* Chen, T., y otros. 2005. CVPR. págs. 532-539.
5. *A Physical Approach to Color Image Understanding.* Klinker, G.J., Shafer, S.A. y Kanade, T. 1990, International Journal Compuet Vision , págs. 7-38.

6. *Shadow detection algorithms for traffic flow analysis: a comparative study.* Prati, Andrea, y otros. USA : IEEE, 2001. Intelligent Transportation Systems Proceedings. págs. 340 - 345 .
7. *Dynamic Camera Calibration of Roadside Traffic Management Cameras for Vehicle Speed Estimation.* Schoepflin, T. and Dailey, D. 2, 2003, IEEE Transactions on Intelligent Transportation Systems, Vol. 4, pp. 90 – 98.
8. *An Algorithm to Estimate Mean Traffic Speed using Uncalibrated Cameras.* Dailey, D. J., Cathey, F. W. and Pumrin, S. 2, 2000, IEEE Transactions on Intelligent Transportation Systems, Vol. 1, pp. 119-130.
9. *Classifying and Counting Vehicles in Traffic Control Applications.* Archetti, F., y otros. s.l. : Springer, 2006, Lecture Notes in Computer Science, Vol. 3907/2006, págs. 495-499.
10. *Analysis of road image sequences for vehicle counting.* Soh, J., Tae-Chun, B. y Wang, M. Vancouver, BC , Canada. : IEEE, 1995. International Conference on Systems, Man and Cybernetics. págs. 679 - 683.
11. *A Machine-Vision System to Detect Unusual Activities Online at Vehicular Intersections.* Canchola-Magdaleno, S., y otros. 2, 2009, Computación y Sistemas, Vol. 13, págs. 209-220.
12. *An adaptive, real-time, traffic monitoring system.* Rodriguez, T. y N., Garcia. 2010, Machine Vision and Applications, págs. 555-576.
13. *Background Subtraction and Shadow Detection in Grayscale Video Sequences.* Jacques, J.C.S., Jung, C.R. y Musse, S.R. Brazil. : IEEE, 2005. Brazilian Symposium on Computer Graphics and Image Processing.
14. *Statistical and knowledge-based moving object detection in traffic scene.* Cucchiara, R., y otros. s.l. : IEEE, 2000. International Conference on Intelligent Transportation Systems. págs. 27-32.
15. Xu, Li, y otros. *Shadow Detection and Removal in Real Images: A Survey* . Shanghai Jiao University, China. : Computer Vision Laboratory, 2006.
16. *Removing Shadows From Images using Retinex.* Finlayson, G.D., Hordley, S.D. y Drew, M.S. USA : s.n., 2002. Color Imaging Conference: Color Science and Engineering Systems, Technologies, and Applications. págs. 73-79.
17. *Intrinsic Images by Entropy Minimization.* Finlayson, G.D., Drew, Mark S. y Lu, Cheng. 2004, Lecture Notes in Computer Science, págs. 582-595.
18. *Simple Shadow Removal.* Fredembach, Clement y Finlayson, G.D. 2006, International Conference on Pattern Recognition, págs. 832 - 835.

Behavioural Architecture for a Differential-Drive Mobile Robot

Carlos Delgado-Mata, Ramiro Velázquez, and Carlos A. Gutiérrez

IPIT (Instituto Panamericano de Investigación Tecnológica),
Escuela de Ingenieria, Universidad Panamericana campus Bonaterra,
Aguascalientes, México, C.P. 20290 cdelgado@up.edu.mx

Abstract. This paper presents the work in progress of an ethologically inspired Action Selection Mechanism to control a Differential-Drive Mobile Robot. The mathematical model of a two wheel differential-drive model is presented. The model shows how zero turning radius is achieved with only bidirectional movement. Behaviour patterns are used to map the incoming stimuli from ultrasound sensors into responses that affect the voltage's intensity of each wheel's motor. Therefore, it performs translational and rotational movements of the mobile robot described herewith.

1 Introduction

Understanding how wheeled mobile robots (WMR) move in response to input commands is essential for feedback control design and many navigation tasks such as path planning, guidance, and obstacle avoidance.

Campion and Chung classified in [1] the mobility of WMR into five generic structures corresponding to a pair of indices (m, s): mobility degree m and steerability degree s. The first one refers to the number of degrees of freedom the WMR could have instantaneously from its current position without steering any of its wheels while the second refers to the number of steering wheels that can be oriented independently in order to steer the WMR.

1.1 Action Selection Mechanisms

Action Selection has been quintessential in fields like simulation of adaptive behaviour. Applications are constantly developed in Robotics, Intelligent Agents, and lately in Virtual Agents populating Virtual Environments, that is embodied virtual agents. Historically, there have been two approaches for selecting actions: the reactive [2] and the deliberative [3]. The advantages of the former [4] is that they are computationally cheap, and can adapt better to a changing environment. The advantages of the latter is that they can hold in memory a representation of the world and thus they -in theory- could accomplish a more informed and better solution than their counterparts. They do not suffer from the local minima problems inherent in local decision making. Motivated behaviours are governed not only by environmental stimuli but also by the internal state of the animal, being influenced by such things as appetite.

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 355-364

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

The rest of the paper is organized as follows: in Section 2, the popular two wheel differential-drive model is obtained using the general two-active-fixed wheels and one-passive-caster wheel structure. In Section 3 a Behavioural Architecture to drive the two wheel diferential-drive robot is described. Finally, the conclusion summarizes the paper main concepts.

2 Wheeled Mobile Robot

The five mobility WMR classes classified by Campion [1], that correspond to a pair of indices (m, s): mobility degree m and steerability degree s, are:

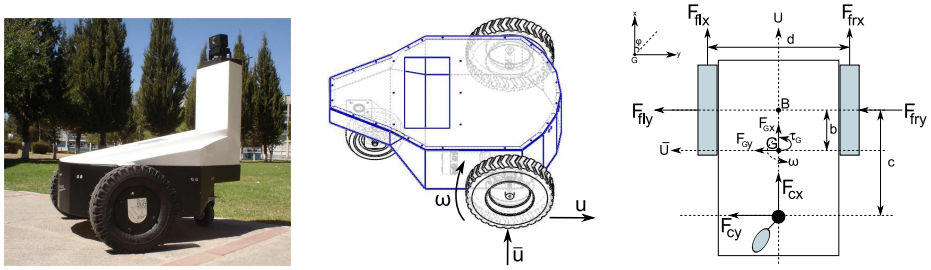
- Type (3,0) robots or omnidirectional robots have no steering wheels (s=0) and are equipped only with Swedish or caster wheels. They have full mobility in the plane (m=3), which means that they are able to move in any direction without any reorientation.
- Type (2,0) robots have no steering wheels (s=0) but either one or several fixed wheels with a common axle. The common axle restricts mobility to a two-dimensional plane (m=2).
- Type (2,1) robots have no fixed wheels and at least one steering wheel. If there is more than one steering wheel, their orientations must be coordinated (s=1). Therefore, mobility is restricted to a two-dimensional plane (m=2).
- Type (1,1) robots have one or several fixed wheels on a common axle and also one or several steering wheels, with two conditions for the steering wheels: their centers must not be located on the common axle of the fixed wheels and their orientations must be coordinated (s=1). Mobility is restricted to a one-dimensional plane determined by the orientation angle of the steering wheel (m=1).
- Type (1,2) robots have no fixed wheels, but at least two steering wheels. If there are more than two steering wheels, then their orientation must be coordinated in two groups (s=2). Mobility is restricted to a one-dimensional plane (m=1) determined by the orientation angles of the two steering wheels.

This paper particularly address type (2,0) robots.

2.1 Differential-Drive WMR

The wheeled mobile robot described herein is a type (2,0) robot. There are many design alternatives; however, the two-wheel differential-drive robot is by far the most popular design.

Let us consider our prototype IVWAN (Fig. 1(a)). Its mechanical structure is based on a differential-drive configuration consisting of two independently controlled front-active wheels and one-rear-caster wheel (Fig. 1(b)). Active wheels are driven by two high-power DC motors which allow IVWAN to achieve a maximum speed of 20 km/hr. IVWAN exhibits both manual and autonomous operation: it can be tele-operated or self-guided by a color camera and an array of ultrasonic sensors that allow the machine to detect and follow visual patterns and negotiate obstacles, respectively [5].



(a) (b) (c)

Fig. 1. Type (2,0) WMR IVWAN (Intelligent Vehicle With Autonomous Navigation): (a) prototype and (b) its differential-drive structure. Two front wheels each driven by its own motor. A third wheel is placed in the rear to passively roll along while preventing the robot from falling over. The wheels exhibit three speeds: u, $\bar{u}$, and ω. (c) Free-body diagram. The first subscript stands for front **f** and caster **c** wheel while the second subscript stands for right **r** and left **l** wheel.

Fig. 1(c) shows a schematic representation of the differential-drive structure. Here, **B** represents the center of the axis connecting both traction wheels; **G** represents the vehicle's center of mass and for simplicity, it is considered as the point to control in position (**x**, **y**) and orientation (φ).

Resultant forces and momentum in the structure can be expressed by eq. (1):

$$\begin{aligned}\sum F_x &= m(\dot{u}-\bar{u}\omega) = F_{frx}+F_{flx}+F_{cx}+F_{Gx}\\ \sum F_y &= m(\dot{\bar{u}}+u\omega) = F_{fry}+F_{fly}+F_{cy}+F_{Gy}\\ \sum M_z &= I\dot{\omega} = \frac{d}{2}(F_{frx}-F_{flx}) - b(F_{fry}+F_{fly}) + \\ &\quad +(c-b)F_{cy}+\tau_G \end{aligned} \tag{1}$$

where **m** is the vehicle's total mass, **I** is the moment of inertia around point **G**, and **u**, $\bar{u}$ and ω are the robot's linear, transverse sliding, and angular speeds, respectively (Fig. 1(b)). Speed $\bar{u}$ can be reasonable neglected assuming that the wheels do not slip during motion. Concerning **u** and ω, they can further be defined by eq. (2):

$$\begin{aligned} u &= \frac{1}{2}[r(\omega_r+\omega_l)+(u_r+u_l)]\\ \omega &= \frac{1}{d}[r(\omega_r-\omega_l)+(u_r-u_l)] \end{aligned} \tag{2}$$

where **r** is the traction wheel radius, **d** is the distance between the traction wheels (see Fig. 1(c)), ω_r, and ω_l are the angular speeds of the right and left wheels respectively, and $\mathbf{u}_r$ and $\mathbf{u}_l$ are the linear speeds of the right and left wheels respectively.

Kinematics of point **G** is related to **u** and ω by eq. (3):

$$\begin{aligned}\dot{x} &= u cos\varphi - b\omega sin\varphi\\ \dot{y} &= u sin\varphi + b\omega cos\varphi\\ \dot{\varphi} &= \omega \end{aligned} \tag{3}$$

As aforementioned, traction wheels are powered by DC motors. These can be modeled by eq. (4):

$$\tau_r = \frac{k_a}{R_a}(E_r - k_b\omega_r)$$
$$\tau_l = \frac{k_a}{R_a}(E_l - k_b\omega_l) \tag{4}$$

where τ_r and τ_r are the torques developed by the motors on the right and left wheels upon input DC voltages $\mathbf{E}_r$ and $\mathbf{E}_l$ respectively, $\mathbf{k}_a$ and $\mathbf{k}_b$ are the motor's torque and electromotive force constants, and $\mathbf{R}_a$ is the motor's electric resistance. Inductive voltages have been neglected.

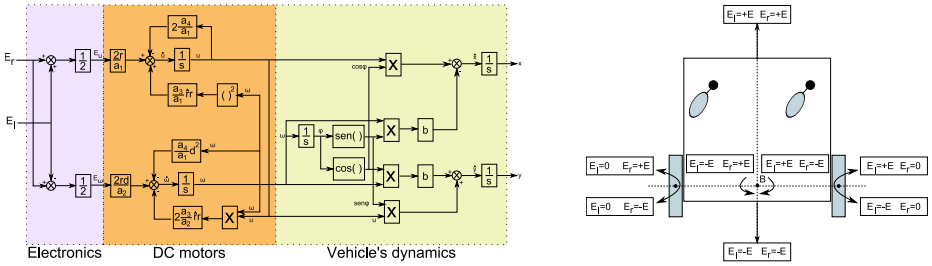


Fig. 2. (a) Block diagram reference for differential-drive robots. (b) Summary of motion upon voltages $\mathbf{E}_r$ and $\mathbf{E}_l$.

Equations describing the wheel-motor system can be simply written as shown in eq. (5):

$$I_e\dot{\omega}_r + D_e\omega_r = \tau_r - F_{frx}\hat{r}$$
$$I_e\dot{\omega}_l + D_e\omega_l = \tau_l - F_{flx}\hat{r} \tag{5}$$

where $\mathbf{I}_e$ and $\mathbf{D}_e$ are the moment of inertia and the coefficient of viscous friction of the wheel-motor system, respectively and $\hat{r}$ is the nominal radius of the traction wheel tires. Using and combining eqs. (1) to (5), the differential-drive model can be summarized by eq. (6):

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\varphi} \\ \dot{u} \\ \dot{\omega} \end{bmatrix} = \begin{bmatrix} ucos\varphi - b\omega sin\varphi \\ usin\varphi + b\omega cos\varphi \\ \omega \\ \frac{a_3}{a_1}\hat{r}r\omega^2 - 2\frac{a_4}{a_1}u \\ -2\frac{a_3}{a_2}\hat{r}ru\omega - \frac{a_4}{a_2}d^2\omega \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \frac{2r}{a_1} & 0 \\ 0 & \frac{2rd}{a_2} \end{bmatrix} \begin{bmatrix} E_u \\ E_\omega \end{bmatrix} \tag{6}$$

with inputs:

$$E_u = \frac{E_r + E_l}{2}$$
$$E_\omega = \frac{E_r - E_l}{2}$$

and constants:

$$a_1 = \frac{R_a}{k_a}(m\hat{r}r + 2I_e) \qquad [V \cdot s^2]$$
$$a_2 = \frac{R_a}{k_a}[I_e d^2 + 2\hat{r}r(I + mb^2)] \qquad [V \cdot m^2 \cdot s^2]$$
$$a_3 = \frac{R_a}{k_a}mb \qquad [V \cdot s^2/m]$$
$$a_4 = \frac{R_a}{k_a}(\frac{k_a k_b}{R_a} + D_e) \qquad [V \cdot s/rad]$$

Note that eq. (6) relates the robot's motion to the motors' input voltages. The block diagram model for differential-drive robots is shown in fig. 2(a). This diagram identifies the electronics, DC motors, and the vehicle's dynamics.

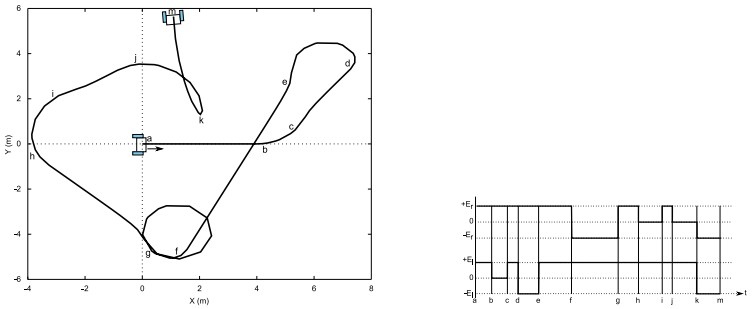
Fig. 2(b) summarizes how differential-drive robots are controlled by the input voltages $\mathbf{E}_r$ and $\mathbf{E}_l$. When both voltages are equal, the two driving wheels turn at the same angular speed and in the same direction, which causes a translation movement. If one voltage is set to zero, one of the wheels turns while the other remains motionless, then the robot describes a circle centered on the motionless wheel. If both voltages are equal in magnitude but opposite sign, the wheels turn at the same speed but in opposite direction which causes a rotation around the center of the axis connecting both wheels (point **B**). Note a zero turning radius in this case.

Numerical values of the parameters involved in eq. (6) can be easily measured from an existent prototype and the specifications of the DC motors can be obtained from the manufacturer. As illustrative example, consider all gain blocks of fig. 2(a) as unity gains. Fig. 3(a) shows a computer simulation of a certain trajectory in the XY plane. Fig. 3(b) shows the driving signals supplied to the DC motors. Note the correspondence with fig. 2(b).

The next section presents the Action Selection Mechanism used to control the differential-drive mobile robot described herewith.

3 Behavioural architecture

This section presents the behavioural architecture used to drive the WMR. The architecture was originally developed for multiple cooperating robots - the Behavioural Synthesis Architecture or BSA [6] - and reapplied it to agents in a virtual environment (VE) in the Virtual Teletubbies project [7]. An object oriented approach (BAMUVA) was developed to simulate conspecific virtual mammals; it is described in [8]. The BSA incorporated three structures at increasing levels of abstraction: behaviour patterns, behaviour packets, and behaviour scripts. An overview of this architecture is presented henceforth.



(a) (b)

Fig. 3. (a) A simulated trajectory of the differential-drive robot and (b) the corresponding driving signals.

3.1 Behaviour patterns

At the most primitive level, a behaviour pattern, (**bp**), was defined as a pair of functional mappings, one from incoming sensory stimulus to outgoing desired motor response, and the other from incoming sensory stimulus to utility. That is, a mapping to define the importance of the motor response for the given level of stimulus. A mobile robot, like the one described herein, possesses a repertoire of behaviour patterns, with each active pattern at any given time proposing its desired motor response (voltage intensity for each high-power DC motor) according to its current sensory input (sonar sensors). These responses are weighted by their utility values and synthesised together to produce an emergent response; the actual behaviour of the mobile robot. Thus, second-to-second variation in emergent behaviour was dealt via weighted synthesis on a continuous basis, unlike the time-sliced Brooksian architecture [4].

The basic component in the architecture is the behaviour pattern, **bp**, where

$$\mathbf{bp} = \begin{bmatrix} r \\ u \end{bmatrix} \tag{7}$$

and

$$r = f_r(s) \tag{8}$$

$$u = f_u(s) \tag{9}$$

r is the desired motion response and is a function, f_r, of a given sensory stimulus, s.

Associated with every response is a measure of its utility or importance, u. This quantity is a function , f_u, of the same sensory stimulus. Hence a **bp** defines not only what the motion response should be for a given sensor input, but it also provides a measure as how the relative importance of this response varies with respect to the same sensor input. The values of r and u constitute a vector known as *utilitor*. Figure 4 shows an example of a simple **bp** that might exist at a given level. Consider the situation where the sensory stimulus relates to a mobile robot's forward facing distance to an obstacle

measuring sensor and the associated motion response relates to the forward velocity for the mobile robot. From figure 4 it can be seen that as the mobile robot gets nearer to the object, its forward translate velocity will be reduced to zero. At the same time the associated utility for the motion response increases. Thus, as the mobile robot gets nearer to an object in its path, the more important it becomes to the mobile robot to slow down. At any point in time, t, multiple conflicting motion responses are typically generated. For example, a mobile robot may be moving towards a goal location when an obstacle unexpectedly appears in its path and at the same time senses that it *needs* to replenish its battery. In such situation what should it do? In BAWMR (Behaviour Architecture for a Wheeled Mobile Robot), conflicting motion responses are resolved by a behaviour synthesis mechanism to produce a resultant motion response. Competing utilitors are resolved by a process of linear superposition which generates a resultant utilitor, UX_t where:

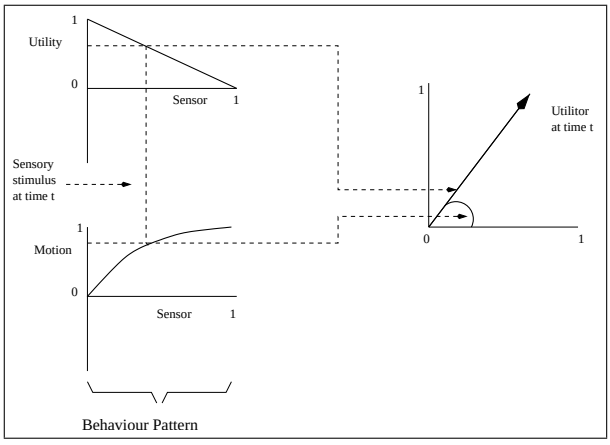


Fig. 4. Behaviour pattern example

$$\mathbf{UX_t} = \sum_{n=1}^{m} u(t,n) \cdot e^{j \cdot r(t,n)} \tag{10}$$

and m equals the total number of related utilitors generated from the different behaviour levels concerned with motion. Given a resultant utilitor, a resultant utility, uX_t, and a resultant motion response, rX_t are simply obtained from

$$\mathbf{yX_t} = \frac{\left|U\ X_t\right|}{m} \tag{11}$$

$$\mathbf{yX_t} = arg(U\ X_t) \tag{12}$$

X identifies the relevant degree of freedom, e.g. forward movement, and the result motion response, rX_t, is then executed by the mobile robot. From equation 4, it can be

seen that generating a resultant utilitor from different behaviours within the architecture constitutes a process of additive synthesis, as shown in figure 5

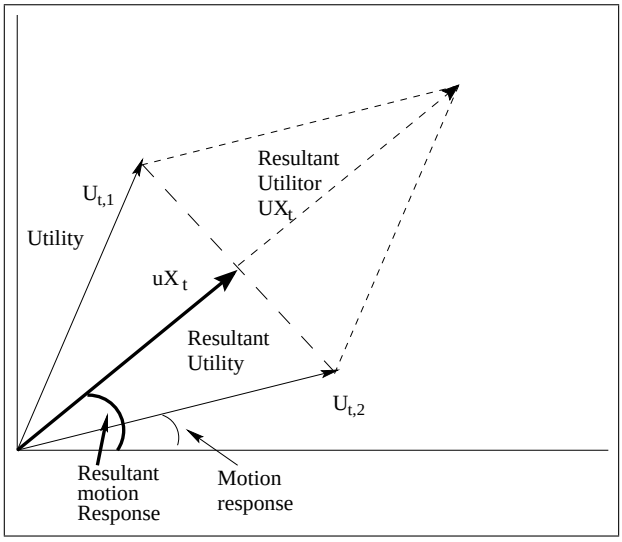


Fig. 5. Generating a resultant utility and motion response from two constituent utilitors

The BAWMR architecture is an Object Oriented extension to the Behaviour Synthesis Architecture, which was developed at the University of Salford [9], to accomplish a task through cooperating robots[6]. This work used ethological knowledge similar to the one described in Animal Behaviour literature [10]. While each robot had a repertoire of simple behaviour patterns, complexity emerged through interactions between behaviour patterns and between robots.

3.2 Behaviour packets

If all the **bp**s in an agent's repertoire were active at the same time then the overall emergent behaviour of the agent might be of little value. For example, patterns designed to produce obstacle avoidance (as described above) are not useful if you want an agent wait for a particular stimuli. The **bp** designer must always bear in mind that the low-level architecture is sensor-driven, and not task or even sub-task dependent. What is needed in this case is an automatic mechanism for deactivating the 'obstacle avoidance' **bp**s when the 'waiting' **bp**s is active. Associated therefore with every **bp** within the mobile robot is an 'active flag', which enables or disables it. Thus obstacle avoidance **bp**s for example can be turned off and on when required. A **bp** is 'deactivated' in the BSA by forcing the respective utility to zero. The action effectively produces a **bp** of zero importance and hence one which does not contribute to the overall emergent behaviour of the agent.

This mechanism is applied by grouping together **bp**s in goal-achieving sets known as behaviour packets. A behaviour packet is a small data structure which includes a sensory pre-condition for activating the **bp**s it references, and a sensory post-condition which controls deactivation of the named **bp**s. Behaviour packets show some similarity with AI production rules [11], though they work at the sub-symbolic level and are driven by incoming sensor data rather than by an inferencing system. They support behavioural sequencing for agents performing at a task (universe) behaviour level. Thus, a sensory pre-condition of 'being near the goal' could be used to move from a behaviour packet in which obstacle avoidance **bp**s were active to one in which they are not.

Therefore, behaviour packets provide a mechanism for contextually sensitive behaviour switching, which is seen as a more flexible mechanism than the finite-state machine definition of inhibition and excitation between behaviours of the subsumption architecture [4].

3.3 Behaviour Script: high-level sequencing and agent drives

A behaviour script is simply a set of behaviour packets assembled for the achievement of a particular task, using the sensory pre-and post-conditions. The original approach was to generate behaviour scripts on the fly using a reflective agent incorporating a symbolic AI planner, and then send the individual scripts to behavioural-based agents. This hybrid approach was taken with the co-operative robots in [12] and is appropriate where the domain is predominantly task-based.

The default script executes a single packet containing **bp**s that effectively lets the low-level module handle wandering in the environment while avoiding obstacles. The default script is changed when another sensory precondition from another set of packets is met.

4 Conclusion

This paper intends to present simple and reliable mathematical model for different design of a type (2,0) robot. In particular, this draft has presented the differential-drive model: the general two-active-fixed wheels and one-passive-caster wheel as well as the belt-drive system. An Action Selection Mechanism to drive the mobile robot was presented. The stimulus received from the sonar sensors is mapped via the BAWMR (Behaviour Architecture for a Wheeled Mobile Robot) into motor responses that affect the voltage's intensity of each of the high-power DC motors. This particular mobile robot is suitable for a non-static environment, which are the most common.

References

1. G. Campion and W. Chung, *Springer Handbook of Robotics*, ch. Wheeled robots, pp. 391–410. Berlin Heidelberg: Springer Verlag, 2008.
2. T. Tyrrell, *Computational Mechanisms for Action Selection*. PhD thesis, University of Edinburgh, Edinburgh, Scotland, 1993.
3. N. Nilsson, *Artificial Intelligence: A New Synthesis*. Morgan Kauffman, 1998.

4. R. A. Brooks, "A robust layered control system for a mobile robot," *IEEE Journal of Robotics and Automation*, vol. 2, no. 1, pp. 14–23, 1986.
5. J. Martinez, G. Moran, B. Romero, A. Camacho, D. Gutheim, J. Varona, and R. Velazquez, "Multifunction all-terrain mobile robot ivwan: design and first prototype," in *Proc. of 2nd Israeli Conference on Robotics*, (Herzlia, Israel), 2008.
6. D. P. Barnes, R. A. Ghanea-Hercock, R. S. Aylett, and A. M. Coddington, "Many hands make light work? an investigation into behaviourally controlled cooperant autonomous mobile robots," in *Proceedings of the first International Conference on Autonomous Agents* (L. Johnson, ed.), (Marina del-Rey, Calif.), pp. 413–20, ACM Press, 1997.
7. R. S. Aylett, A. Horrobin, J. J. O'Hare, A. Osman, and M. Polyak, "Virtual teletubbies: reapplying a robot architecture to virtual agents," in *3rd International Conference on Autonomous Agents*, (Seattle Washington), 1999.
8. C. Delgado-Mata and R. Aylett, "Communicating emotion in virtual environments through artificial scents," in *LNCS Intelligent Virtual Agents* (E. de Antonio and R. Aylett, eds.), (Madrid, Spain), Springer-Verlag, 2001.
9. D. P. Barnes, "A behaviour synthesis architecture for cooperant mobile robot," in *Advanced Robotics and Intelligent Machines* (J. O. Gray and D. G. Caldwell, eds.), pp. 295 – 314, United Kingdom: IEE Control Engineering Series 51, 1996.
10. D. McFarland, *Animal Behaviour*. England: Addison Wesley Longman, 1999.
11. R. Davis and J. King, *An overview of Production Systems in Machine Intelligence*, pp. 300–332. Wiley and Sons, 1977.
12. R. S. Aylett, "Communicating goals to behavioural agents - a hybrid approach," in *Proceedings of the Third World Congress on Expert Systems*, (Seoul), 1996.

A Dexterous Robot Hand with Embedded SMA Actuators

Ramiro Velázquez[1] and Edwige Pissaloux[2]

[1] Mecatrónica y Control de Sistemas (MCS)
Universidad Panamericana
Aguascalientes, Mexico
rvelazquez@ags.up.mx
[2] Institut des Systèmes Intelligents et de Robotique (ISIR)
CNRS UMR 7222 - Université Paris 6
Paris, France
Edwige.Pissaloux@upmc.fr

Abstract. This paper presents the design and implementation of a four-fingered human-size robotic hand intended for dexterous and grasping manipulation applications. The prototype hand has 3 degrees of freedom (DOF) per finger and 2 DOF in the thumb for a total of 11 DOF. Shape memory alloys (SMAs) micro-coils are embedded intrinsically within the hand structure to power the joints while exhibiting some advantageous features such as low-cost, lightweight, compactness and clean silent operation. To increase efficiency and decrease power dissipation, the SMA micro-actuators integrate magnetic bi-stable structures. Mechanical design, actuation approach and first prototype are presented and discussed.

Keywords: Bi-stable actuator, micro-coils, robot hand, shape memory alloys (SMAs)

1 Introduction

The human hand is a highly functional structure which roboticians have attempted to imitate for a long time. With its 21 degrees of freedom (DOF) (Fig. 1), 19 muscles, 17 joints, 19 bones, in addition to its ligaments, nerves and numerous sensors, the human hand is a very complex structure difficult to reproduce mechanically. Moreover, it is dexterous, stable and precise, but also fast moving, strong and flexible.

Over the past 20 years, a number of robotic hands have been developed for dexterous and skillful grasping applications in medical, welfare, space, industrial and virtual environments [1]-[5].

Traditionally, robotic hand designs have tended to be bulky, heavy and noisy due to the use of conventional actuation approaches such as electromechanical and pneumatic technologies. As a result of the problems inherent with these types of actuation, designers have been adapting various new actuation approaches for use in their place. It is believed that the use of smart materials such as shape memory alloys (SMAs), piezoelectric ceramics, electroactive polymers (EAPs) and electrorheological (ER) fluids will provide new design methodologies and paradigms for lightweight robotic hands.

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 365-375

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

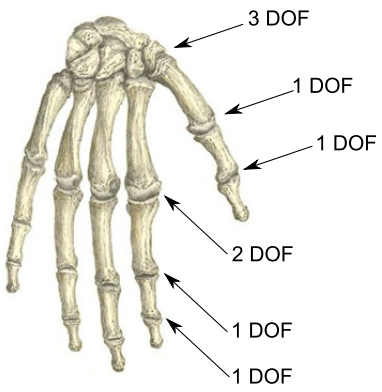


Fig. 1. Joints of the human hand with their corresponding DOF, after [6].

SMA technology seems to be one of the most promising candidates for the development of meso/micro actuators: compact size, high power/weight ratio, extremely high fatigue resistance to cyclic operation and smooth, clean, spark-free and noiseless performance make them an interesting actuation principle [7]. However, two major drawbacks are their relatively slow response speed and their non-linear behavior that make them difficult to control.

In this paper, we present the design and implementation of a robotic hand based on SMAs. The prototype hand is human-size, four-fingered and has 11 DOF. A SMA based micro-actuator intended for powering the fingers' joints is proposed. The actuator consists of an antagonist arranged pair of NiTi micro-coils, which integrates bi-stable structures to increase overall efficiency. This design approach permits to have all the actuators embedded intrinsically within the hand structure while keeping the prototype compact, lightweight, low-power consumption and low-cost.

The rest of the paper is organized as follows: Section 2 overviews the essential features of SMAs. Section 3 presents the design, characterization and implementation of a SMA bi-stable micro-actuator. Section 4 introduces the design of the robotic hand while Section 5 presents the first prototype developed and its grasping capabilities. Finally, Section 6 concludes the paper summarizing the main contributions and future work perspectives.

2 Essential features of shape memory alloys

The shape memory effect (SME) is the ability of a certain group of materials to "memorize" a specific shape when subjected to the appropriate thermal process. Materials exhibiting this effect are metallic alloys such as Ag-Cd, Au-Cd, Cu-Al-Ni, Cu-Sn, In-Ti, Ni-Al and the popular Ni-Ti (Nitinol) [8].

The SME occurs as a result of a temperature dependent transformation between two solid phases: high temperature *austenite* phase and low temperature *martensite* phase. Austenite phase contains the memorized or predefined shape of the material. When

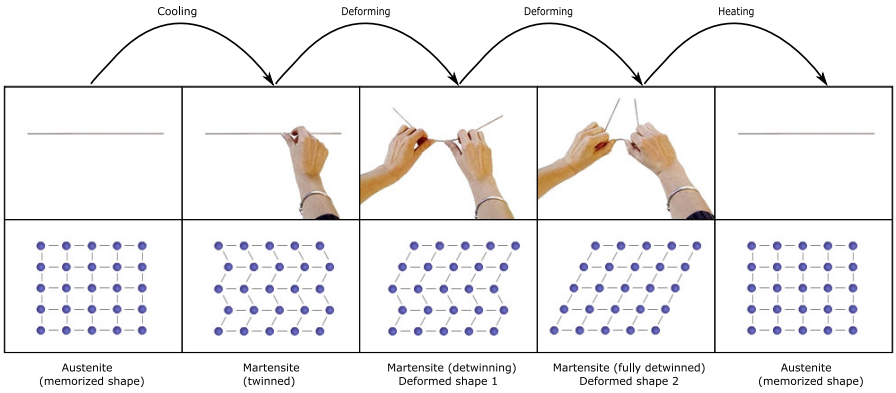


Fig. 2. The SME in SMAs at macroscopic and crystallographic levels.

cooled to martensite, the material presents very low stiffness and yield strength. It is quite malleable and can be easily deformed into a new shape, which it retains. Upon heating, the material returns to its austenitic original and pre-deformed shape (Fig. 2 top).

At microscopic level, the SMA's austenitic crystalline structure is highly symmetric and well ordered. Upon cooling, the crystalline structure collapses leaning in opposite directions along subsequent layers that are self-accommodating (or twinned) so that no macroscopic deformation results. Applying an external stress on the SMA will cause the twinned martensite layers to begin to lean in the same direction. When all the layers are leaning the same way, the SMA is said to be oriented (or detwinned). Upon heating, the crystal layers line up to recover their original symmetry (Fig. 2 bottom).

The SME is repeatable and it can be considered as a transformation phenomenon of thermal energy into mechanical work, usually generating force and displacement. Thus, SMAs can be used as active elements of actuators.

SMAs can be formed into almost any shaped actuator. Most popular shapes are wire, spring, tubing, sheet and ribbon.

Experimental testing of NiTi SMA straight wires has revealed that stroke is limited to approximately 4 to 5 % of their original length [9]. Compared to wires, coil springs do have a significantly higher recoverable strain. Impressive 300 % strokes can easily be obtained using helical springs. However, as they perform in torsion instead of tension, they cannot develop the same force.

This paper suggests that SMA coil springs can be designed to exert significant forces for powering robotic hands having at the same time, stroke and compactness not offered by straight wires.

3 A SMA based bi-stable micro-actuator

3.1 Material

A NiTi micro-coil spring was fabricated with trademark Flexinol wire ($Ni_{52}Ti_{48}$) with the following geometric characteristics: 200 μm of wire diameter, 1.3 mm of mean spring diameter and 12 active coils. Its mass is 30 mg and its laboratory cost is only 1 USD.

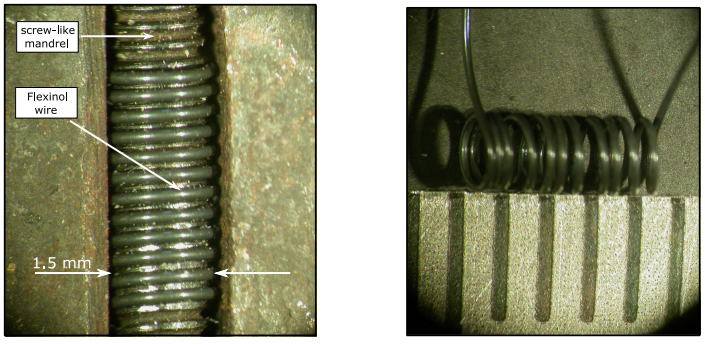
The coil shape was set by winding the wire tightly on a cylindrical mandrel (screw-like) and then heat-treating both wire and mandrel at 600°C for 5 min. Rapid cooling via water quench concluded the process (Fig. 3).

Differential scanning calorimeter (DSC) tests revealed that full austenite phase is achieved at 65°C while full martensite is at 36°C.

3.2 Performance Evaluation

The micro-coil's parameters were adjusted so that it develops an appropriate force according to its deflection and temperature. Fig. 4(a) shows the experimental load-deflection relations under various isothermal conditions for a maximum deflection of 10 mm. Note 2 different behaviors: elastic before load reaches a critical value and plastic beyond this critical value. The curves indicate that the maximum output force of the SMA micro-coil at a deflection of 10 mm is approximately 180 mN at 20°C, 410 mN at 85°C and 720 mN at 105°C.

Using this data, it is possible to formulate a 3D computational model of the SMA micro-coil's behavior in the load-deflection-temperature space (Fig. 4(b)) [10]. Note that, as introduced in fig. 4(a), the tensile force developed by the micro-coil against a load increases with temperature.



(a) (b)

Fig. 3. Fabrication procedure of an SMA micro-coil spring: (a) constrained Flexinol wire on a mandrel and (b) coil shape after heat treatment.

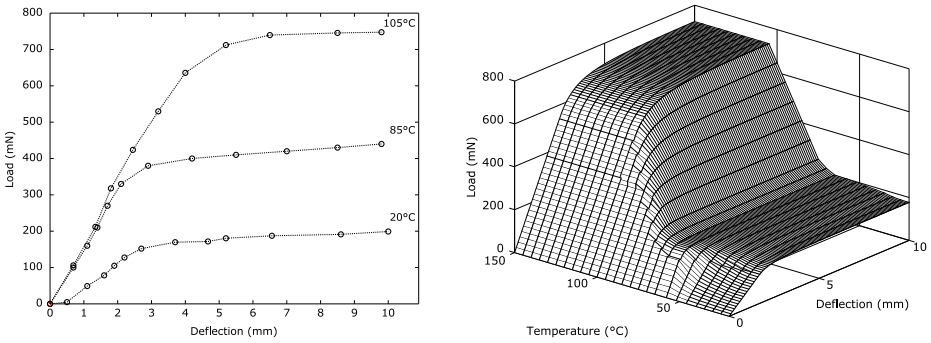


Fig. 4. (a) The micro-coils's experimental load-deflection behavior at constant temperatures (20, 85 and 105°C) and (b) computational model in the load-deflection-temperature space. Experimental observations revealed that beyond 105°C, the load-deflection behavior was practically the same.

3.3 Design of Actuators

As discussed in section 2, SMAs achieve their actuation through the phenomenon of the shape memory effect (SME), which is non-reversible. This implies that, a SMA itself cannot be considered as an actuator since it does not provide reversible motion.

Two general approaches are used to exploit SMA materials in actuator applications: the one-way and the two-way effect.

The one-way effect describes the ability of SMAs to recover a memorized shape when heated up to austenite, but retains this shape when cooled down to martensite. To be used in cyclic actuation, it is necessary to provide a biasing force to induce the initial deformed shape in the martensite phase.

The two-way effect describes a memory process with two stable shapes: one in austenite and the other one in martensite. Therefore, the two-way effect does not require any external mechanisms for cyclic actuation.

While using the two-way effect provides simpler and compacter actuators with many fewer elements involved, it certainly requires extra manufacturing processes, it is difficult to achieve correctly and its strain is only half of that observed in one-way SMAs [11].

Although a one-way effect SMA could be designed such that it exerts a force in three dimensions (when deformed in 3 directions from the memory configuration), the great majority apply a one-directional tensile force and cannot directly apply a compressive force.

In many robotic applications, this is accomplished by using a mass, an elastic element or a second SMA arranged in antagonist mode [12]. The simple mechanisms in fig. 5 show how these configurations can be used to design actuators that create linear motion.

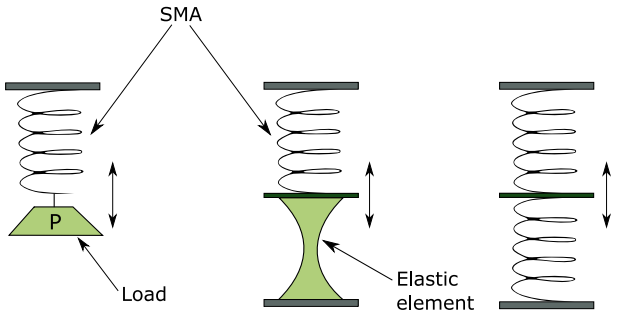


Fig. 5. Main SMA actuator configurations using one-way effect SMAs.

Note that all of these configurations provide motion essentially between two points defined by the active and passive states of the SMA. However, only the passive state can be considered stable since no power is needed to sustain it.

In contrast, the active state not only requires energy to be achieved but also to be retained; the SMA must be kept active, which not only leads to poor efficiency, but also implies serious drawbacks mostly when electrically heated.

In principle, an SMA can be heated arbitrarily quickly by passing a sufficiently large current through it. However, excessive electrical power has the capacity to overheat the SMA, causing thermal stress fatigue and a gradual degradation of its performance.

A strategy for retaining the active position in the martensite state of the SMA is then required to avoid power consumption, useless output work and overheating. One method to achieve two stable positions is the use of clamping/latching mechanisms. Thus, the actuator would need energy only when changing from one state to the other and in neither of the two end positions power would be needed.

If we consider that the hand's joints spend most of their time in a fixed position, it can be concluded that, actuators that spend a great deal of energy maintaining their active state are not acceptable. Bi-stable actuators are more suitable for this application [13].

In physical implementation, it is often advantageous to make use of bi-stable structures: they provide accurate and repeatable motion and, in presence of disturbances or environmental variations, they maintain the desired position [14].

4 Design of a SMA based robotic hand

The conceptual representation of a four-fingered anthropomorphic robotic hand is shown in fig. 6 left.

This design has 3 DOF per finger and 2 DOF in the thumb for a total of 11 DOF. Possible hand motions are shown in fig. 7. Note that the maximum joint deflection for this prototype is 90°. Thus, if all three joints are deflected 90°, the fingertip's total workspace will be 360°.

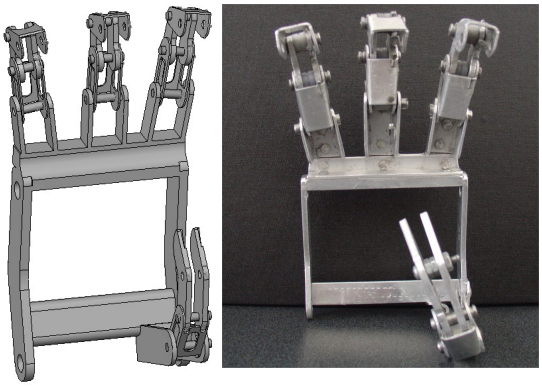


Fig. 6. Conceptual representation of a four-fingered robotic hand and its equivalent mechanical structure.

Fig. 6 right shows the equivalent mechanical structure. This structure is the same size as an adult's hand. Base and fingers were entirely fabricated using lightweight aluminum while joints are made of hardened plastic. Each finger link mass is estimated to be 35 g, so the entire finger is 105 g. The total structure is about 600 g.

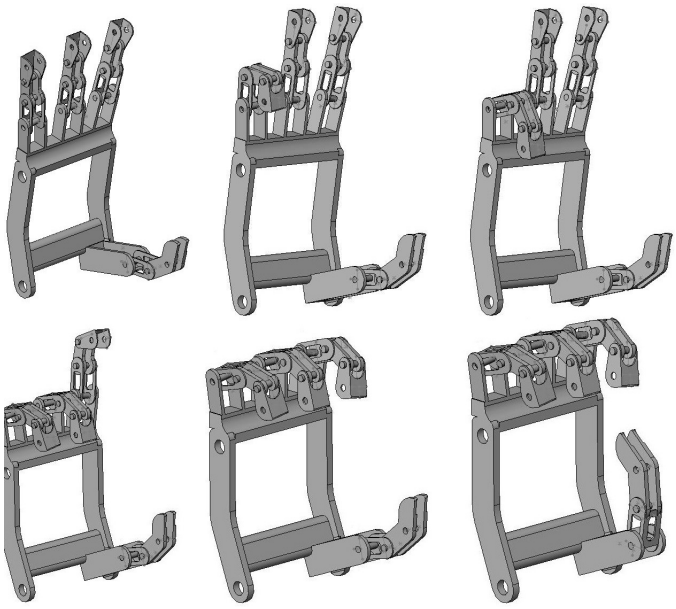


Fig. 7. Possible hand motions: 3 DOF per finger and 2 DOF in the thumb.

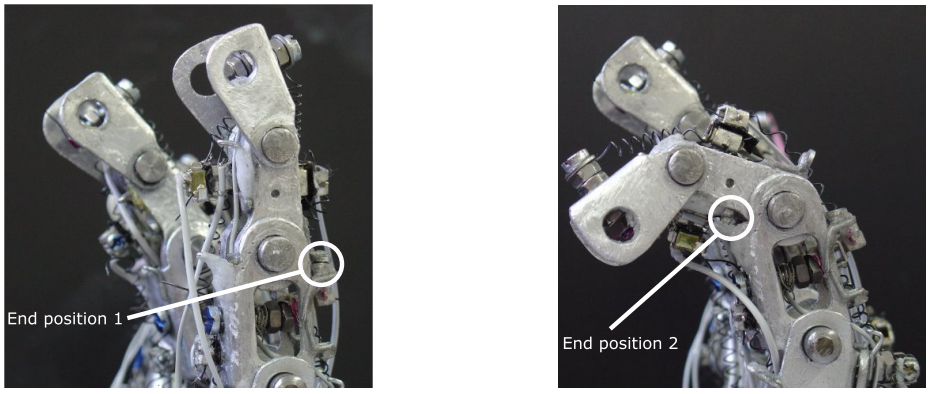
5 Prototype

The mechanical structure in fig. 6 is actuated by a set of SMA micro-coils of characteristics described in section 3. Note from fig. 4(a), that the micro-coil is perfectly capable of developing appropriate forces for powering the 35 g finger links when heated to 85°C. Furthermore, these micro-coils are quite small, so they can be embedded within the hand structure. No external actuation module is needed.

The SMA micro-coils were arranged in antagonist configuration. As seen from fig. 5, the antagonist principle is based on heating one SMA element at a time, so that its austenitic state produces a force and displacement over the second martensitic SMA. To retain the active position, the actuator integrate bi-stable structures.

Many mechanisms exhibit bi-stability: switches, closures, hinges, shampoo bottle caps, bicycle kickstands, tape measures, retractable pens, etc. [14]. A magnet approach seemed appropriate for incorporation into the robotic hand.

The operation principle is shown in fig. 8: (1) at the initial position, both SMA micro-coils are in martensite state. Position is retained by the magnetic attraction of two magnets (Fig. 8(a)). (2) When electrical current flows in one of the SMAs, its austenitic compressed memorized shape unblocks the magnets, extends the martensitic SMA and moves the joint until it reaches the second end position. Two other magnets are attracted ensuring this position and no further energy is required until a change is necessary (Fig. 8(b)).



(a) (b)

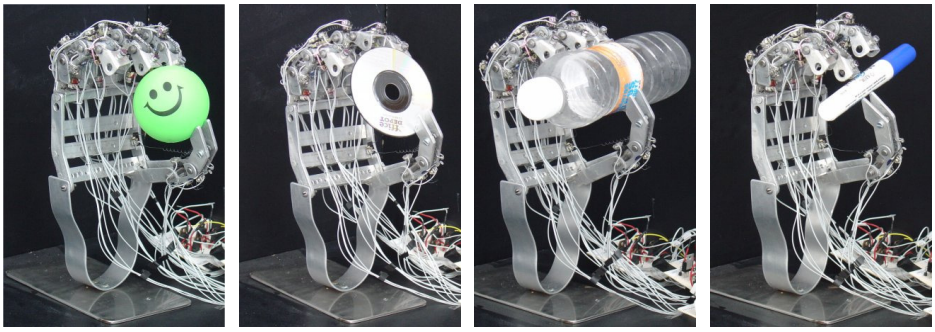
Fig. 8. SMA antagonist actuator with bi-stable structures. When one SMA micro-coil is heated, it shrinks and the link moves to one direction. When the opposite SMA is heated, the link moves to the opposite direction. End positions are ensured by the magnetic attraction of a pair of magnets.

The first prototype hand is shown in fig. 9. It is well known that SMAs cannot be welded, so they were mechanically fastened to the aluminum links. This also eases the replacement task: the micro-coil is simply unscrewed from the link and it can then be lifted out.

Fig. 9. First prototype of robotic hand with SMAs.

A total of 6 SMA micro-coils are embedded in each finger while the thumb contains 4. The entire hand encloses 22 SMA micro-coils or 11 antagonist pairs. A couple of wires are connected to each SMA to electrically control them from its electronic drive. Wire guides are included along the structure to avoid the risk of jamming.

Some grasping examples conducted with this prototype are shown in fig. 10. Note that objects in figs. 10(a)-(c) can be classified as "big" objects or power grasp as termed in [15] while 10(d) is a "small" object or precision grasp. Current work is focused on precision grasp to assist humans with objects difficult to manipulate.



(a) (b) (c) (d)

Fig. 10. Power and precision grasping examples: (a) sphere, (b) disk, (c) "big" cylinder and (d) "small" cylinder.

6 Conclusion

This paper has presented both design and implementation of a four-fingered human-like robotic hand intended for dexterous and grasping manipulation applications. The prototype proposed here uses non-classical types of actuation, such as SMAs, which allow the development of efficient, compact, lightweight and low-cost actuators.

An SMA based electrically driven micro-actuator intended for powering the hand's joints has been presented. This actuator uses an antagonist arranged pair of SMA NiTi micro-coils whose thermomechanical behavior has been characterized. To increase efficiency, ensure repeatability and decrease power dissipation, this micro-actuator integrates magnetic bi-stable structures.

The first prototype developed has 3 DOF per finger and 2 DOF in the thumb for a total of 11 DOF. It integrates 11 SMA antagonist actuators within its structure and is capable of grasping several kinds of objects. Current experiments involve precision grasping.

Future research perspectives include tele-manipulation based on electromyographic (EMG) signals. EMG signals will be measured from the operator's forearm muscles to detect finger motion. The prototype robotic hand is expected to move accordingly.

References

1. I. Yamano and T. Maeno, "Five-fingered robot hand using ultrasonic motors and elastic elements", *Proc. IEEE International Conference on Robotics and Automation*, pp 2684-2689, 2005.
2. K. DeLaurentis, C. Mavroidis and C. Pfeiffer, "Development of a shape memory alloy actuated robotic hand", *Proc. 7th International Conference on New Actuators*, pp 281-285, 2000.
3. J. Butterfass, G. Hirzinger, S. Knoch and H. Liu, "DLR's multisensory articulated hand, part I: hard and software architecture", *Proc. IEEE International Conference on Robotics and Automation*, pp 2081-2086, 1998.
4. J. Zhao, Z. Xie, L. Jiang, H. Cai, H. Liu and G. Hirzinger, "A five-fingered underactuated prosthetic hand control scheme", *Proc. IEEE International Conference on Biomedical Robotics and Biomechatronics*, pp 995-1000, 2006.
5. H. Kawasaki, T. Komatsu and K. Uchiyama, "Dexterous anthropomorphic robot hand with distributed tactile sensor: Gifu hand II", *IEEE/ASME Transactions on Mechatronics*, 7(3), pp 296-303, 2002.
6. Anatomy Atlases (Digital Library). Updated information available at: http://www.anatomyatlases.org/
7. K. Ikuta, "Micro/Miniature shape memory alloy actuator", *Proc. IEEE International Conference on Robotics and Automation*, pp 2156-2161, 1990.
8. C. Mavroidis, C. Pfeiffer and M. Mosley, "Conventional actuators, shape memory alloys, and electrorheological fluids", *Automation, Miniature Robotics and Sensors*, Y. Bar-Cohen Ed. pp 189-214, 2000.
9. N. Chaillet, "Contribution à la modélisation, la commande et la réalisation de microactionneurs et de composants pour la microrobotique". *HDR Thesis*, Franche-Compte University, 1999.
10. R. Velazquez, M. Hafez, E. Pissaloux and J. Szewczyk, "A computational-experimental thermomechanical study of shape memory alloy microcoils and its application to the design of actuators", *Journal of Computational and Theoretical Nanoscience*, 3(4), 538-550, 2006.

11. J. Gill, K. Ho and G. Carman, "Three-dimensional thin-film shape memory alloy microactuator with two-way effect", *Journal of Microelectromechanical Systems*, 11(1), pp 68-77, 2002.
12. T. Waram, "Actuator design using shape memory alloys", T.C. Waram Publishing, Ontario, 1993.
13. R. Velazquez, "Contribution à la conception et à la réalisation d'interfaces tactiles portables pour les déficients visuels", *PhD Thesis*, Paris 6 University, 2006. Available on-line at: http://tel.archives-ouvertes.fr/tel-00085703
14. M. Hafez and M. Benali-Khoudja, "3D tactile rendering based on bi (multi) stable SMA monolithic systems", *Proc. of IEEE International Symposium on Micromechatronics and Human Science*, pp 93-98, 2004.
15. J. Napier, "The prehensile movements of the human hand", *Journal of Bone and Joint Surgery*, 38-B(4), pp 902-913, 1956.

A Study on the Wide-Sense Stationarity and Mean Ergodicity of SOC Rayleigh Fading Channel Simulators

Carlos A. Gutiérrez, Anely Meléndez, Arturo Sandoval, and Hermes Rodríguez

Panamericana University, Campus Aguascalientes,
Josemaría Escrivá de Balaguer No. 101, Aguascalientes 20290, México
{cagutierrez,anely.melendez,jose.sandoval,hermes.rodriguez}@up.edu.mx

Abstract. Stationarity and ergodicity are desirable properties of any stochastic simulation model for small-scale mobile radio channels. These properties enable the channel simulator to accurately emulate the channel's statistical properties in a single simulation run without requiring information on the time origin. In a previous paper, we analyzed the ergodicity with respect to (w.r.t.) the autocorrelation function (ACF) of seven fundamental classes of stochastic sum-of-cisoids (SOC) simulation models for mobile Rayleigh fading channels. In this paper, we continue our investigations on the subject by providing a comprehensive study on the wide-sense stationarity and the ergodicity w.r.t. the mean value of these classes of SOC channel simulators. The obtained results can be used in connection with those presented in our previous paper to design efficient channel simulators for the performance evaluation of modern mobile communication systems.

Keywords: Channel simulators, ergodic processes, mean value, mobile communications, sum-of-cisoids, wide-sense stationary processes.

1 Introduction

The proliferation of low-cost electronic devices with high computational capabilities and the need that exists among telecommunications engineers for affordable and powerful tools for the performance evaluation of modern mobile communication systems have turned the design of computer simulators into a major subject of research. When designing a simulator for the performance assessment of wireless communication systems, it is fundamental to choose a proper model to simulate the channel. This is of primary importance, since the channel exerts a strong influence on the system's performance [16, Ch. 15]. Several different simulation models for multipath radio channels have been proposed in the literature, such as those based on autoregressive processes [1], digital filters [17], linear transformations of complex Gaussian sequences [4], and Karhunen-Lòeve expansions of stochastic processes [19]. However, simulation models based on a finite sum of complex sinusoids (cisoids) have been shown to be an excellent basis for the design of single-input single-output (SISO) [7, 8] and multiple-input multiple-output (MIMO) [15, 20, 21] multipath radio channel simulators. Sum-of-cisoids (SOC) models are well suited for the simulation of fading channels under both isotropic and non-isotropic scattering conditions, as demonstrated in [5]. They have found applications, e.g., in the laboratory analysis of space-time coding schemes [18].

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 376-385

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

Two desirable properties of any stochastic simulation model for small-scale multipath radio channels are stationarity and ergodicity. These properties enable the channel simulator to accurately emulate the channel's statistical properties in a single simulation run (ergodicity) without requiring information on the time origin (stationarity). In the strict sense, a channel simulator is stationary if all marginal and joint probability density functions (PDFs) of the random process characterizing the underlying simulation model are time independent. On the other hand, a channel simulator is ergodic if the time averages of the simulation model are equal to the ensemble averages. These conditions are too stringent and are hardly satisfied in practice. However, the information about the channel's third- or higher-order statistics is rarely required to assessing the performance of wireless communication systems. Hence, for most practical purposes, it suffices if the channel simulator is wide-sense stationary (WSS) and ergodic with respect to (w.r.t.) the mean value and the autocorrelation function (ACF). Indeed, an important part in the statistical characterization of a channel simulator consists in determining whether the simulation model is a WSS, a mean ergodic (ME), or/and an autocorrelation ergodic (AE) random process.

In a previous paper [6], we analyzed the autocorrelation ergodicity of seven fundamental classes of stochastic SOC simulation models for mobile Rayleigh fading channels. In this paper, we continue our investigations on the subject by providing a comprehensive study on the wide-sense stationarity and mean ergodicity of these classes of SOC channel simulators. To the best of the authors' knowledge, the WSS and ME properties of stochastic SOC channel simulators have not been systematically analyzed so far. We notice, nonetheless, that some partial results are available in the literature. In [5], the wide-sense stationarity, mean ergodicity, and autocorrelation ergodicity of a class of SOC models defined by cisoids with constant gains, constant frequencies, and random phases were studied. In [9], the first-order stationarity of the envelope of the seven fundamental classes of stochastic SOC models was investigated. The work in [12–14] is also worth mentioning. There, the authors analyzed the wide-sense stationarity, mean ergodicity, and autocorrelation ergodicity of stochastic sum-of-sinusoids simulation (SOS) for mobile fading channels. Despite the similarities between SOS and SOC models, we point out that the results obtained in [12–14] are not valid for SOC channel simulators. This is because the ACF of an SOC model has more degrees of freedom than the ACF of a conventional SOS model, as explained in [6, Sec. II]. The findings reported in this paper complement those presented in [5,6,9,12–14] and can be used as guidelines to design efficient channel simulators for the performance evaluation of modern mobile communication systems.

The outline to the rest of the paper is as follows. In Section 2, we provide a brief description of an statistical reference model for narrowband mobile Rayleigh fading channels. In Section 3, we review the characteristics of the seven classes of stochastic SOC simulation model for mobile Rayleigh fading channels. In Section 4, we systematically analyze the WSS and ME properties of the classes of stochastic SOC channel simulators. Finally, Section 5 concludes the paper with some remarks and a summary of key results. As a notational convention, we will use bold symbols and letters to denote random variables and stochastic processes, whereas normal symbols and letters will be used for constants and deterministic processes.

2 The Reference Model

A small-scale narrowband mobile Rayleigh fading channel can be represented in the equivalent baseband by a complex Gaussian random process

$$\boldsymbol{\mu}(t) = \boldsymbol{\mu}_I(t) + j\boldsymbol{\mu}_Q(t), \qquad j \triangleq \sqrt{-1} \tag{1}$$

where $\boldsymbol{\mu}_I(t)$ and $\boldsymbol{\mu}_Q(t)$ are stationary real-valued Gaussian processes with mean zero and variance $\sigma_{\boldsymbol{\mu}}^2/2$. Equation (1) may be rewritten in phasor notation as

$$\boldsymbol{\mu}(t) = \boldsymbol{\zeta}(t)\exp\{j\boldsymbol{\phi}(t)\} \tag{2}$$

where $\boldsymbol{\zeta}(t) = \sqrt{\boldsymbol{\mu}_I^2(t) + \boldsymbol{\mu}_Q^2(t)}$ and $\boldsymbol{\phi}(t) = \arctan\big(\boldsymbol{\mu}_Q(t)/\boldsymbol{\mu}_I(t)\big)$. One can easily verify that the first-order PDF of $\boldsymbol{\zeta}(t)$ equals the Rayleigh distribution with parameter $\sigma_{\boldsymbol{\mu}}$ [11, Sec. 5.5], while the first-order statistics of $\boldsymbol{\phi}(t)$ are characterized by a circular uniform PDF [11, Sec. 5.6].

The statistical properties of the Rayleigh fading channel model described by the complex Gaussian process in (1) are completely specified by the time-shift insensitive (TSI) ACF $r_{\boldsymbol{\mu}\boldsymbol{\mu}}(\tau)$ of $\boldsymbol{\mu}(t)$, where $r_{\boldsymbol{x}\boldsymbol{x}}(\tau) \triangleq E\{\boldsymbol{x}^*(t)\boldsymbol{x}(t+\tau)\}$ for an arbitrary random process $\boldsymbol{x}(t)$. The operators $E\{\cdot\}$ and $(\cdot)^*$ stand for the statistical expectation and the complex conjugate, respectively. Assuming a two-dimensional fixed-to-mobile propagation environment, the ACF of $\boldsymbol{\mu}(t)$ can be written as [3]

$$r_{\boldsymbol{\mu}\boldsymbol{\mu}}(\tau) = \sigma_{\boldsymbol{\mu}}^2 \int_{-f_{\max}}^{f_{\max}} p_{\boldsymbol{f}}(f)\exp\big\{j2\pi f\tau\big\}\,df. \tag{3}$$

where $f_{\max}$ is the maximum Doppler shift caused by the movement of the mobile terminal, and $p_{\boldsymbol{f}}(f)$ is the PDF of the random Doppler frequencies of the channel's multipath components.

3 SOC Simulation Models

Most of the statistical properties of $\boldsymbol{\mu}(t)$ relevant for system performance analysis—such as its correlation properties, spectral characteristics, and the first-order distributions of its envelope and phase—can accurately be emulated via a simulation model based on a finite SOC, as demonstrated in [2, 7, 8]. Figure 1 shows the general structure of an SOC Rayleigh fading channel simulator with N homogeneous cisoids[1], the parameters of which—gains, frequencies, and phases—are defined either as random variables or deterministic quantities. An SOC simulation model can mathematically be described by a complex random process $\hat{\boldsymbol{\mu}}(t)$ if any of the cisoids' parameters is random, otherwise, it is to be represented by a complex deterministic process $\hat{\mu}(t)$. A classification of SOC channel simulators based on the type of the cisoids' parameters was introduced in [9]. All in all, eight fundamental classes of SOC simulators were identified in that paper.

[1] By homogeneous cisoids we mean a group of cisoids characterized by the same type of parameters.

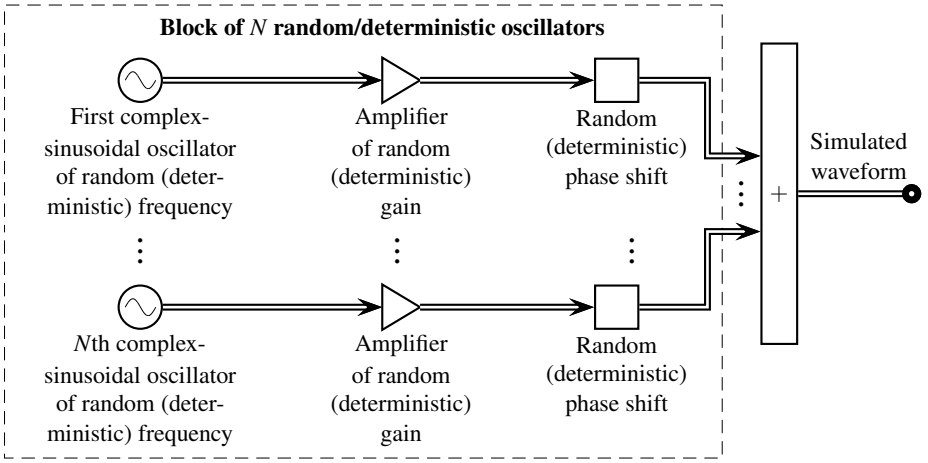


Fig. 1. Block diagram of an SOC simulation model for Rayleigh fading channels [6].

Table 1. Classification of SOC simulation models for Rayleigh fading channels with respect to the type of the cisoids' parameters [9].

Parameters	Gains	Frequencies	Phases
Class I	Deterministic	Deterministic	Deterministic
Class II	Deterministic	Deterministic	Random
Class III	Deterministic	Random	Deterministic
Class IV	Deterministic	Random	Random
Class V	Random	Deterministic	Deterministic
Class VI	Random	Deterministic	Random
Class VII	Random	Random	Deterministic
Class VIII	Random	Random	Random

These classes, which are listed in Table 1, will be taken as a reference in this paper to carry out our investigations on the WSS and ME properties of SOC channel simulators. For the analysis presented herein it is assumed that the following holds:

- All random variables are statistically independent.
- If the cisoids' phases are random variables, then they are uniformly distributed over $[-\pi,\pi)$.
- If the cisoids' Doppler frequencies are random variables, then they have a PDF $p_{\boldsymbol{f}}(f)$ identical to that characterizing the statistics of the reference model's Doppler frequencies.
- If the cisoids' gains are random variables, then they are identically distributed with a mean value m_c and a variance $\sigma_c^2 = \sigma_{\boldsymbol{\mu}}^2/N$.

The concepts of stationarity and ergodicity do not apply on the Class I simulators, since this class of models is completely deterministic. However, the information about the temporal mean value (TMV) of the Class I simulators is fundamental to find out whether or not a given class of stochastic SOC models is defined by a set of ME processes.

4 WSS and Mean Ergodic Properties of SOC Channel Simulators

4.1 Definitions

Before we proceed to analyze the WSS and ME properties of SOC simulation models, it is convenient to formally define the concepts of wide-sense stationarity and mean ergodicity.

Definition 1 (WSS process) *Let $\hat{\boldsymbol{\mu}}(t)$ be a random process. Then, $\hat{\boldsymbol{\mu}}(t)$ is said to be WSS if [10, p. 555]:*

- *The mean value $m_{\hat{\boldsymbol{\mu}}}(t) \triangleq E\{\hat{\boldsymbol{\mu}}(t)\}$ of $\hat{\boldsymbol{\mu}}(t)$ is time independent, i.e., $m_{\hat{\boldsymbol{\mu}}}(t) = m_{\hat{\boldsymbol{\mu}}}$.*
- *The ACF $r_{\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}}(t_1,t_2) \triangleq E\{\hat{\boldsymbol{\mu}}^*(t_1)\hat{\boldsymbol{\mu}}(t_2)\}$ of $\hat{\boldsymbol{\mu}}(t)$ depends only on the time difference $\tau = t_2 - t_1$, meaning that $r_{\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}}(t_1,t_2)$ is TSI, so that $r_{\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}}(t_1,t_2) = r_{\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}}(\tau)$.*

Definition 2 (ME process) *Let $\hat{\boldsymbol{\mu}}(t)$ be a random process whose mean value $m_{\hat{\boldsymbol{\mu}}}(t) \triangleq E\{\hat{\boldsymbol{\mu}}(t)\}$ is constant over time, so that $m_{\hat{\boldsymbol{\mu}}}(t) = m_{\hat{\boldsymbol{\mu}}}$. Then, $\hat{\boldsymbol{\mu}}(t)$ is said to be ME if [10, Sec. 6.6]:*

- *The TMV $m_{\hat{\mu}}^{(k)} \triangleq \langle \hat{\mu}^{(k)}(t) \rangle$ of every sample function $\hat{\mu}^{(k)}(t)$ of $\hat{\boldsymbol{\mu}}(t)$ is equal to $m_{\hat{\boldsymbol{\mu}}}$, i.e., $m_{\hat{\mu}}^{(k)} = m_{\hat{\boldsymbol{\mu}}}$ $\forall k$.*

The notation $\langle x(t) \rangle$ stands for the time average of an arbitrary function of time $x(t)$.

4.2 Classes of SOC Channel Simulators and Their Autocorrelation Properties

For the analysis of the wide-sense stationarity, it is necessary to determine if the ACF of the random process $\hat{\boldsymbol{\mu}}(t)$ characterizing each of the seven classes of stochastic SOC models is a TSI function. This was already done in [6]. The results there obtained are summarized in Table 2, where θ_n denotes the phase of the n-th cisoid. This table also summarizes the conclusions drawn in [6] regarding the autocorrelation ergodicity of the seven classes of stochastic SOC models.

4.3 Classes of SOC Channel Simulators and Their WSS and ME Properties

Class I Channel Simulators The simulation models of Class I are characterized by a deterministic SOC model

$$\hat{\mu}(t) = \sum_{n=1}^{N} c_n \exp\left\{ j(2\pi f_n t + \theta_n) \right\} \tag{4}$$

where the cisoids' gains c_n, Doppler frequencies f_n, and phases θ_n are arbitrary constants. To ensure that the Doppler power spectral density (DPSD) of $\hat{\mu}(t)$ is band-limited, it is assumed that $f_n \in (-f_{\max}, f_{\max})$, $\forall n$. The TMV $m_{\hat{\mu}} \triangleq \langle \hat{\mu}(t) \rangle$ of this class of deterministic SOC models is given as

$$m_{\hat{\mu}} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \hat{\mu}(t)dt. \tag{5}$$

Table 2. Classes of SOC models and their autocorrelation properties [6].

Class	Gains	Frequencies	Phases	AE	TSI ACF
I	Deterministic	Deterministic	Deterministic	—	—
II	Deterministic	Deterministic	Random	Yes	Yes
III	Deterministic	Random	Deterministic	No	No/Yes[a]
IV	Deterministic	Random	Random	No	Yes
V	Random	Deterministic	Deterministic	No	No/Yes[b]
VI	Random	Deterministic	Random	No	Yes
VII	Random	Random	Deterministic	No	No/Yes[b]
VIII	Random	Random	Random	No	Yes

[a] If the boundary condition $\sum_{n=1}^{N}\sum_{m=1,m\neq n}^{N}\exp\{j(\theta_m-\theta_n)\}=0$ is satisfied.

[b] If the mean value of the random gains is equal to zero.

Substituting (4) in (5), and assuming that $f_n \neq 0$, $\forall n$, we have

$$m_{\hat{\mu}} = 0. \tag{6}$$

The k-th sample function of an stochastic SOC model can be represented by a deterministic process $\hat{\mu}^{(k)}(t)$ similar to that defined in (4). Therefore, the TMV $m_{\hat{\mu}^{(k)}} \triangleq \langle \hat{\mu}^{(k)}(t) \rangle$ of $\hat{\mu}^{(k)}(t)$ is given as in (6) for all k regardless of the class of stochastic SOC models under consideration.

Class II Channel Simulators Simulation models of this class are characterized by a stochastic SOC model of the form

$$\hat{\boldsymbol{\mu}}(t) = \sum_{n=1}^{N} c_n \exp\left\{ j(2\pi f_n t + \boldsymbol{\theta}_n) \right\}. \tag{7}$$

It is straightforward to verify that

$$m_{\hat{\boldsymbol{\mu}}}(t) = E\{\hat{\boldsymbol{\mu}}(t)\} = 0. \tag{8}$$

With reference to Definition 2, and taking account of (6) and (8), we can conclude that the SOC models of Class II are ME, since $m_{\hat{\mu}}^{(k)} = m_{\hat{\boldsymbol{\mu}}} = 0$, $\forall k$. Furthermore, given that $m_{\hat{\boldsymbol{\mu}}}(t)$ is constant over time and the ACF of $\hat{\boldsymbol{\mu}}(t)$ is TSI (see Table 2), it follows that the Class II SOC simulators are WSS processes.

Class III Channel Simulators This class of simulators is defined by the set of stochastic processes of the form $\hat{\boldsymbol{\mu}}(t) = \sum_{n=1}^{N} c_n \exp\left\{ j(2\pi \boldsymbol{f}_n t + \theta_n) \right\}$. For this class of SOC simulators, we have

$$m_{\hat{\boldsymbol{\mu}}}(t) = \frac{r_{\boldsymbol{\mu}\boldsymbol{\mu}}(t)}{\sigma^2_{\hat{\boldsymbol{\mu}}}} \sum_{n=1}^{N} c_n \exp\{j\theta_n\} \tag{9}$$

where $r_{\boldsymbol{\mu}\boldsymbol{\mu}}(t)$ is the TSI ACF of the reference channel model [see (3)]. Since the mean value of $\hat{\boldsymbol{\mu}}(t)$ is time dependent, it follows that the Class III SOC simulators are not WSS or ME processes. However, $\hat{\boldsymbol{\mu}}(t)$ proves to be a ME process if any of the following conditions is met:

- The number of cisoids N is even and the phases θ_n are given such that $\theta_n = -\theta_{n+N/2} = \pi/2$ for $n = 1,\ldots,N/2$.
- The number of cisoids N is even, the phases θ_n are equal to each other, and the gains are given such that $c_n = -c_{n+N/2}$ for $n = 1,\ldots,N/2$.

In turn, the wide-sense stationarity property of $\hat{\boldsymbol{\mu}}(t)$ holds if any of the aforementioned conditions is fulfilled and $\sum_{n=1}^{N}\sum_{m=1,m\neq n}^{N}\exp\{j(\theta_m-\theta_n)\}=0$ (cf. Table 2).

Class IV Channel Simulators The Class IV simulators are characterized by a stochastic process $\hat{\boldsymbol{\mu}}(t)=\sum_{n=1}^{N}c_n\exp\left\{j(2\pi\boldsymbol{f}_n t+\boldsymbol{\theta}_n)\right\}$. It is straightforward to show that the mean value of the Class IV simulators is time independent and equal to $m_{\hat{\boldsymbol{\mu}}}=0$. We can therefore conclude that the Class IV of SOC simulators is defined by a set of ME random processes. Moreover, since the ACF of this class of SOC simulation models is TSI (see Table 2), it follows that the SOC models of this class are also WSS random processes.

Class V Channel Simulators This class of simulators is defined by the set of stochastic SOC models $\hat{\boldsymbol{\mu}}(t)=\sum_{n=1}^{N}\boldsymbol{c}_n\exp\left\{j(2\pi f_n t+\theta_n)\right\}$. In this case, the mean value of $\hat{\boldsymbol{\mu}}(t)$ is equal to

$$m_{\hat{\boldsymbol{\mu}}}(t)=m_c\sum_{n=1}^{N}\exp\{j(2\pi f_n t+\theta_n)\}. \tag{10}$$

It is clear from the previous equation that if $m_c\neq 0$, then the mean value of the Class V simulators is time dependent and $\hat{\boldsymbol{\mu}}(t)$ is not a WSS nor a ME process. However, if the mean value of the random gains $\boldsymbol{c}_n$ is equal to zero, then $m_{\hat{\boldsymbol{\mu}}}(t)=0$. Thereby, $\hat{\boldsymbol{\mu}}(t)$ proves to be a WSS and a ME random process, since $m_{\hat{\mu}(k)}=m_{\hat{\boldsymbol{\mu}}}$ for all k, and $r_{\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}}(t_2,t_2)=r_{\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}}(\tau)$ if $m_c=0$ (cf. Table 2).

Class VI Channel Simulators The Class VI simulators are characterized by a stochastic process $\hat{\boldsymbol{\mu}}(t)=\sum_{n=1}^{N}\boldsymbol{c}_n\exp\left\{j(2\pi f_n t+\boldsymbol{\theta}_n)\right\}$. For this class of simulators, we have that $m_{\hat{\boldsymbol{\mu}}}(t)=m_{\hat{\boldsymbol{\mu}}}=0$. The simulation models of Class VI are therefore ME processes. Moreover, they are also WSS, since their ACF is always TSI (cf. Table 2).

Class VII Channel Simulators This class of simulators is defined by the set of stochastic SOC models of the form $\hat{\boldsymbol{\mu}}(t)=\sum_{n=1}^{N}\boldsymbol{c}_n\exp\left\{j(2\pi\boldsymbol{f}_n t+\theta_n)\right\}$. The mean value of this class of simulators is equal to

$$m_{\hat{\boldsymbol{\mu}}}(t)=\frac{m_c\, r_{\boldsymbol{\mu}\boldsymbol{\mu}}(t)}{\sigma^2_{\boldsymbol{\mu}}}\sum_{n=1}^{N}\exp\{j\theta_n\}. \tag{11}$$

From the previous equation, it follows that the mean value of the Class VII simulators is time independent if $m_c = 0$ or if the number of cisoids is even and $\theta_n = -\theta_{n+N/2} = \pi/2$ for $n = 1,\ldots,N/2$. If any of these conditions is fulfilled, then the SOC simulators of Class VII are ME processes. On the other hand, based on the results presented in Table 2, we can conclude that the SOC simulators of this class are both ME and WSS if and only if $m_c = 0$.

Class VIII Channel Simulators Simulation models of the Class VIII are characterized by a random process $\hat{\boldsymbol{\mu}}(t) = \sum_{n=1}^{N} \boldsymbol{c}_n \exp\left\{j(2\pi \boldsymbol{f}_n t + \boldsymbol{\theta}_n)\right\}$. For this class of SOC models, one can easily verify that $m_{\hat{\boldsymbol{\mu}}}(t) = m_{\hat{\boldsymbol{\mu}}} = 0$. In view of this result, we can conclude that the Class VIII SOC simulation models are ME random processes. They are also WSS processes, as it was found in [6] that the ACF of this class of simulators is a TSI function.

5 Conclusions

In this paper, we continued our investigations on the stationarity and ergodicity of SOC simulation models for mobile Rayleigh fading channels. Specifically, we analyzed the WSS and ME properties of seven fundamental classes of stochastic SOC channel simulators. Based on the results presented in this paper, we can conclude that the SOC simulators of classes II, IV, VI, and VIII are always WSS and ME random processes. On the other hand, SOC simulators of classes III, V, and VII are WSS and ME provided that some specific conditions are fulfilled. The findings reported herein complement those presented in a previous paper, where we analyzed the AE properties of the seven classes of stochastic SOC models. Table 3 summarizes the results obtained in both papers. As a final remark, we observe that only the simulation models of Class II possesses the desired WSS, ME, and AE properties. Hence, this type of models provide an excellent basis for the design of efficient channel simulators for the performance analysis of mobile communication systems.

Acknowledgments

This work was financed in part by the Panamerican Center of Research and Innovation (Centro Panamericano de Investigación e Innovación (CEPii)).

References

1. Baddour, K.E., Beaulieu, N.C.: Autoregressive modeling for fading channel simulation. IEEE Trans. Wireless Commun. 4(4), 1650–1662 (Jul 2005)
2. Cheng, X., Wang, C.X., Laurenson, D.I., Salous, S., Vasilakos, A.V.: New deterministic and stochastic simulation models for non-isotropic scattering mobile-to-mobile rayleigh fading channels. Wirel. Commun. Mob. Comput. (Oct 2009), DOI: 10.1002/wcm.864
3. Clarke, R.H.: A statistical theory of mobile radio reception. Bell Syst. Tech. J. 47, 957–1000 (Jul 1968)

Table 3. Classes of SOC Rayleigh fading channel simulation models and their stationary and ergodic properties.

Class	Gains	Freq.	Phases	TIMV*	TSI ACF	WSS	ME	AE
I	Det.	Det.	Det.	—	—	—	—	—
II	Det.	Det.	Rand.	Yes	Yes	Yes	Yes	Yes
III	Det.	Rand.	Det.	No/Yes[c or d]	No/Yes[a]	No/Yes[a and (c or d)]	No/Yes[c or d]	No
IV	Det.	Rand.	Rand.	Yes	Yes	Yes	Yes	No
V	Rand.	Det.	Det.	No/Yes[b]	No/Yes[b]	No/Yes[b]	No/Yes[b]	No
VI	Rand.	Det.	Rand.	Yes	Yes	Yes	Yes	No
VII	Rand.	Rand.	Det.	No/Yes[b or c]	No/Yes[b]	No/Yes[b]	No/Yes[b or c]	No
VIII	Rand.	Rand.	Rand.	Yes	Yes	Yes	Yes	No

* The acronym TIMV stands for time independent mean value.

[a] If the boundary condition $\sum_{n=1}^{N}\sum_{m=1,m\neq n}^{N}\exp\{j(\theta_m-\theta_n)\}=0$ is satisfied.

[b] If the mean value of the random gains is equal to zero.

[c] If the number of cisoids N is even and $\theta_n=-\theta_{n+N/2}=\pi/2$ for $n=1,\ldots,N/2$.

[d] If the number of cisoids N is even, the phases θ_n are equal to each other, and the gains are given such that $c_n=-c_{n+N/2}$ for $n=1,\ldots,N/2$.

4. Ertel, R.B., Reed, J.H.: Generation of two equal power correlated Rayleigh fading envelopes. IEEE Commun. Lett. 2(10), 276–278 (Oct 1998)
5. Gutiérrez, C.A.: Channel Simulation Models for Mobile Broadband Communication Systems, Doctoral Dissertations at the University of Agder 16. University of Agder, Kristiansand, Norway (2009)
6. Gutiérrez, C.A., Meléndez, A., Sandoval, A., Rodriguez, H.: On the autocorrelation ergodic properties of sum-of-cisoids Rayleigh fading channel simulators. In: Proc. 2011 European Wireless Conference (EW'2011). pp. 1–6. Vienna, Austria (2011)
7. Gutiérrez, C.A., Pätzold, M.: Sum-of-sinusoids-based simulation of flat-fading wireless propagation channels under non-isotropic scattering conditions. In: Proc. 50th IEEE Global Communications Conference (Globecom 2007). pp. 3842–3846. Washington, DC (Nov 2007)
8. Gutiérrez, C.A., Pätzold, M.: The design of sum-of-cisoids Rayleigh fading channel simulators assuming non-isotropic scattering conditions. IEEE Trans. Wireless Commun. 9(4), 1308–1314 (Apr 2010)
9. Hogstad, B.O., Pätzold, M.: On the stationarity of sum-of-cisoids-based mobile fading channel simulators. In: Proc. 67th IEEE Veh. Technol. Conf. (VTC2008-spring). pp. 400–404. Singapore (May 2008)
10. Leon-Garcia, A.: Probability and Random Processes for Electrical Engineering. Addison-Wesley, New York, second edn. (1994)
11. Parsons, J.D.: The Mobile Radio Propagation Channel. John Wiley and Sons, Chichester, England, second edn. (2000)
12. Pätzold, M.: On the stationarity and ergodicity of fading channel simulators based on Rice's sum-of-sinusoids. Int. Journal of Wireless Information Networks 11(2), 63–69 (Apr 2004)
13. Pätzold, M., Hogstad, B.O.: Classes of sum-of-sinusoids Rayleigh fading channel simulators and their stationary and ergodic properties — Part II. WSEAS Transactions on Mathematics 4(4), 441–449 (Oct 2005)

14. Pätzold, M., Hogstad, B.O.: Classes of sum-of-sinusoids Rayleigh fading channel simulators and their stationary and ergodic properties — Part I. WSEAS Transactions on Mathematics 5(2), 222–230 (Feb 2006)
15. Pätzold, M., Hogstad, B.O., Youssef, N.: Modeling, analysis, and simulation of MIMO mobile-to-mobile fading channels. IEEE Trans. Wireless Commun. 7(2), 510–520 (Feb 2008)
16. Sklar, B.: Digital Communications: Fundamentals and Applications. Prentice Hall, New Jersey, second edn. (2001)
17. Smith, J.I.: A computer generated multipath fading simulation for mobile radio. IEEE Trans. Veh. Technol. 24(3), 39–40 (Aug 1975)
18. Sterian, C.E.D., Ma, Y., Pätzold, M., Banica, I., He, H.: New super-orthogonal space-time trellis codes using differential M-PSK for noncoherent mobile communication systems with two transmit antennas. Annals of Telecommunications (Jul 2010), DOI: 0.1007/s12243-010-0191-1
19. Yip, K.W., Ng, T.S.: Karhunen-Loève expansion of the WSSUS channel output and its application to efficient simulation. IEEE J. Sel. Areas Commun. 15(4), 640–646 (May 1997)
20. Zajić, A.G., Stüber, G.L.: Space-time correlated mobile-to-mobile channels: Modelling and simulation. IEEE Trans. Veh. Technol. 57(2), 715–726 (Mar 2008)
21. Zhang, H., Yuan, D., Pätzold, M., Wu, Y., Nguyen, V.D.: A novel wideband space-time channel simulator based on the geometrical one-ring model with applications in MIMO-OFDM systems. Wirel. Commun. Mob. Comput. 10(6), 758–771 (Jun 2010)