# Special Issue:
# Advances in Pattern Recognition

# Research in Computing Science

## Series Editorial Board




***Research in Computing Science*** es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 71,** mayo 2014. Tiraje: 500 ejemplares. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No. : 04-2005-121611550100-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Impreso en la Ciudad de México, en los Talleres Gráficos del IPN – Dirección de Publicaciones, Tres Guerras 27, Centro Histórico, México, D.F. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

**Editor responsable:** *Grigori Sidorov, RFC SIGR651028L69*

**Research in Computing Science** is published by the Center for Computing Research of IPN. **Volume 71**, May 2014. Printing 500. The authors are responsible for the contents of their articles.



Printed in Mexico City, in the IPN Graphic Workshop – Publication Office.

Volume 71

# Special Issue:
# Advances in Pattern Recognition

**José Arturo Olvera-López**
**José Francisco Martínez-Trinidad**
**Jesús Ariel Carrasco-Ochoa**
**Ching Y. Suen (Eds.)**


Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2014

**ISSN: 1870-4069**



Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx



Indexed in LATINDEX and Periodica / Indexada en LATINDEX y Periódica

Printing: 500 / Tiraje: 500

Printed in Mexico / Impreso en México

# Preface

The 2014 Mexican Conference on Pattern Recognition (MCPR 2014, June 25-28) was the sixth event in the series. The conference was jointly organized between the Computer Science Department of the National Institute for Astrophysics Optics and Electronics (INAOE) and the Computer Science Department of the Autonomous University of Puebla (BUAP), under the auspices of the Mexican Association for Computer Vision, Neurocomputing and Robotics (MACVNR), which is affiliated to the International Association for Pattern Recognition (IAPR). MCPR series of conferences aim to provide a forum for the exchange of scientific results, practice, and new knowledge, as well as, promoting co-operation among research groups in Pattern Recognition and related areas in Mexico and around the world.

This year MCPR included a Postgraduate Students' Meeting (MCPR2014-PSM) allowed discussing their research work in order to receive feedback from experienced researchers, advices for future directions, learn from each other, as well as promoting their participation in conference events.

This volume contains original contributions carefully selected which are derived from both Master and PhD students' researches related to the area of Pattern Recognition. We cordially thank all authors who submitted their contributions to build this volume as well as the members of the Reviewing committee for their effort for evaluating and selecting the papers among the submissions that were received.

We hope this volume from the MCPR2014 Postgraduate Students' Meeting will prove useful to the reader, and hope that the meeting itself will provide a fruitful forum to enrich the collaboration between students and the broader Pattern Recognition community.

We also want to thank to the generous grants received from SEP and CONACYT.

The submission, reviewing, and selection process was supported for free by the EasyChair system, www.EasyChair.org.

José Arturo Olvera-López
José Francisco Martínez-Trinidad
Jesús Ariel Carrasco-Ochoa
Ching Y. Suen

May 2014

# Table of Contents

Page

## Advances in Pattern Recognition

# Representative Pattern Mining in Graph Collections

Niusvel Acosta-Mendoza[1,2], Jesús Ariel Carrasco-Ochoa[2], José Fco. Martínez-Trinidad[2], Andrés Gago-Alonso[1], and José E. Medina-Pagola[1]

[1] Advanced Technologies Application Center
$7^a$ ♯ 21406 e/ 214 and 216, Siboney, Playa, C.P. 12200, Havana, Cuba.
{nacosta,agago,jmedina}@cenatav.co.cu
[2] National Institute for Astrophysics, Optics and Electronic.
Luis Enrique Erro No 1, Sta. Ma. Tonantzintla, 72840, Puebla, México.
{nacosta,ariel,fmartine}@ccc.inaoep.mx

**Abstract.** Nowadays, there has been a meaningful increase in the use of frequent approximate subgraph (*FAS*) mining for different applications, for example, graph classification. However, the great amount of mined patterns is one of the fundamental drawbacks of FAS mining. This drawback has a negative effect in the computational performance of classifiers, especially in large graph databases where the number of frequent patterns could be very high. In this paper, we propose a research proposal driven to obtain FAS mining algorithms capable to compute a representative subset of patterns. The representative pattern set should be identified into the mining process improving the efficiency in time, in comparison with the time required if this identification is performed in a post-processing stage over all patterns computed by a general FAS mining algorithm.

**Key words:**Approximate graph mining, representative patterns, graph-based classification.

## 1 Introduction

In data mining, frequent pattern identification has become a meaningful topic with a wide set of applications in several domains of the science [1]. This topic includes different techniques for pattern extraction, where frequent subgraph mining techniques have been highlighted. Using graphs as basic structure allows identifying patterns with spatial and semantic relationships.
Several algorithms have been developed for finding all frequent subgraphs in a graph database [2–6]. Most of these algorithms use exact matching methods for computing the frequent subgraphs, but there are several practical problems where the need to allow some variations in the data arises. This fact is because there are concrete problems where exact matching could not be applied with positive outcome [7, 8]. This means that it is important tolerating certain level

of distortion, semantic variations, vertices or edges mismatched during the frequent pattern search. Thus, it is required to evaluate the similarity between graphs considering approximate matching. In this way, several algorithms have been developed for frequent approximate subgraph (*FAS*) mining, which use different approximate graph matching techniques allowing the detection of frequent subgraphs with some distortions in the data [7–11]. These FAS mining algorithms have been successfully used for supervised classification, where FASs are used as features for representing objects. This approach has been used in several domains of the science as: analysis of biochemical structures [9, 11], analysis of genetic networks [10], analysis of circuits, links and social networks [8], and image classification [7]. However, in most of these applications, usually a large number of frequent subgraphs is computed [12], therefore, discovering interesting patterns in this set of patterns is still a challenge. Several techniques have been proposed for identifying interesting subgraphs, reducing the dimensionality of the identified pattern set, such as: identifying only maximal, cliques, and closed subgraphs, among others. Using only maximal frequent subgraphs instead of using all the patterns is one of the techniques used to avoid redundancy among the computed patterns and consequently for reducing the dimensionality of this set of patterns. A maximal frequent subgraph is a pattern that is not a subgraph of any other frequent subgraph [13]. It is important to highlight that from the frequent maximal subgraphs it is possible to reconstruct the whole set of frequent subgraphs because all of them are summarized into the maximal patterns. However, from the maximal patterns, the information about the support of non-maximal patterns cannot be retrieved. To face this problem, in several applications closed frequent subgraph are used. A closed frequent subgraph is a pattern that does not have any supergraph with the same frequency [14, 15]. Thus, from the closed frequent subgraphs it is possible to reconstruct the whole set of frequent subgraphs including the information about their support. In real applications such as biochemical compounds, clique frequent subgraphs [9, 16] have been used for reducing the amount of mined patterns. A frequent clique subgraph is a pattern where every two vertices are connected by an edge. Using this kind of patterns, specially when the graph collection contains many clique graphs, the amount of patterns is too high.

This paper is structured as follows: in Section 2, some related works about algorithms for computing representative patterns are described. In Section 3, the research problem is presented. In Section 4, the research proposal is discussed. This proposal includes: the research question, the aims, and the expected contributions of this research. Later, in Section 5 we present some preliminary results. Finally, our conclusions are included in Section 6.

## 2 Related work

Several researchers have turned their attention to the problem of mining maximal, closed or clique patterns in graph collections [9, 14–17]; however, only a few of these works are based on approximate graph matching:

- *APGM* [9] computes the frequent approximate subgraphs that are cliques in a graph collection. This algorithm uses a depth-first search (DFS) approach for building each clique candidate pattern by extending the edges and using the Canonical Adjacency Matrix code (CAM code) of each candidate, sub-isomorphism tests are applied. Also, in this process, a substitution matrix, containing probabilities of interchange between vertex labels, is used. Although the authors suggested that this idea can be extended to edge labels, this algorithm only deals with variations between vertex labels.
- Z. Zou *et al.* [16] propose an algorithm that computes top-k maximal clique subgraphs in an uncertain graph. In this approach, a combination of both, maximal and clique, are used for taking advantage from both approaches. This algorithm uses an exact approach for computing sub-isomorphism between graphs, but during the candidate generation process, each candidate is identified as a clique evaluating the probability that the candidate has of being a clique across all processed graphs. Each graph, processed by this algorithm, is uncertain because it is built taking into account the existence probability over the original graph.

In this paper, we are focused on the approximate approach for graph mining, which allows some semantic variations in vertex and edge labels keeping the graph topology.

## 3 Research problem

Frequent approximate subgraph mining have become a very commonly used technique in data knowledge extraction, which has been successfully applied in several domains of the science. This technique has become an important topic in those mining tasks where the mined patterns are detected taking into account distortions in the data. Using these approximate techniques, better results than the exact techniques are reported in some tasks of graph classification, however, it has a main problem that a high number of patterns are identified during the mining process. This high amount of patterns increases the computational resources needed for storing them, affecting the efficiency and efficacy of the methods where they will be used.

## 4 Proposal

In this section, a research proposal to give a solution to the previously commented problem is presented.

### 4.1 Research question

Is it possible to propose new algorithms for computing representative frequent approximate subgraphs, that allow keeping or improving the classification efficacy reported in the state-of-the-art when this type of subgraphs are used as attributes in supervised problems?

### 4.2 Aims

The general aim of this research is:

To propose new algorithms for mining representative FAS that allow us keeping or improving the classification efficacy, in supervised problems, reported in the state-of-the-art when this type of subgraphs are used as attributes.

The specific aims are:

1. Propose a new algorithm for computing maximal FASs in graph collections.
2. Propose a new algorithm for computing closed FASs in graph collections.
3. Propose a new algorithm for computing clique FASs in graph collections.
4. Extend a based-graph classification framework for evaluating the efficacy and efficiency of the representative subgraphs computed by our algorithms.

### 4.3 Expected contributions

The expected contributions of this proposal are:

1. A review of algorithms for frequent subgraph mining in graph collections.
2. An algorithm for computing the maximal FASs in a graph collection.
3. An algorithm for computing the closed FASs in a graph collection.
4. An algorithm for computing the clique FASs in a graph collection.
5. A graph classification framework based on FAS mining using the proposed algorithms.

## 5 Preliminary results

As preliminary results of the proposed research, we propose an algorithm for computing maximal frequent approximate subgraphs (*M-FASs*) based on an algorithm for FASM proposed by Acosta-Mendoza *et al.* [7] (VEAM), where substitution matrices are used to specify which vertices, edges or labels can replace some other ones; allowing variation into the vertex and edge labels, but keeping the graph topology.

Our proposal, which is a modification of the VEAM algorithm, called *M-VEAM*, extracts only the maximal FASs from a graph collection. M-VEAM (see Algorithm 1) starts finding the frequent approximate single-edge set $C$, using a breadth-first search (*BFS*). Later, for each pattern in $C$, a function "Search" (see Algorithm 2) that recursively computes all extensions of a given pattern using depth-first search (*DFS*), is invoked and if the extended pattern is maximal (i.e. none of its extensions is frequent) then it is stored into the output set $F$. The function "appLset" (see Algorithm 3) searches the possible approximate label set for the new edge $e$ which is an extension of a pattern $T$ and the possible label set of the new vertex that $e$ connects with an existing vertex in $T$ (if is necessary). Finally, when all FASs in $C$ have been extended, the set $F$ of all M-FASs in the

**Algorithm 1**: *M-VEAM*

**Input**: $D$ : A graph collection
$MV$ : Substitution matrix indexed by $L_V$
$ME$ : Substitution matrix indexed by $L_E$
$\tau$ : Similarity threshold
$\delta$ : Support threshold.
**Output**: $F$ : Maximal frequent approximate subgraph set.

```
1 F ← ∅;
2 C ←the frequent approximate single-edge set in D;
3 foreach T ∈ C do
4     Search(T, D, MV, ME, τ, δ, F);
5     if T is maximal then
6         Insert T in F;
```

**Algorithm 2**: *Search*

**Input**: $T = (V_t, E_t, I_t, J_t)$ : A frequent approximate subgraph
$D$ : Graph collection
$MV$ : Substitution matrix indexed by $L_V$
$ME$ : Substitution matrix indexed by $L_E$
$\tau$ : Similarity threshold
$\delta$ : Support threshold
$F$ : Frequent approximate subgraph set.
**Output**: $F$ : Maximal frequent approximate subgraph set.

```
1  foreach o_j ∈ O(T, G_i), where G_i ∈ D do
2      foreach e = ExtSet(o_j) do
3          CL ← appLSet(T, MV, ME, G_i, o_j, e, τ);
4          foreach (elabel, vlabel) ∈ CL do
5              The candidate X is built using the tuple (elabel, vlabel);
6              C ← C ∪ {(X, codeCAM(X), score)};
7  foreach T_1 ∈ C do
8      if sup_G(T_1, D) ≥ δ and T_1 ∉ F then
9          Search(T_1, D, MV, ME, τ, δ, F);
10         if T_1 is maximal then
11             Insert T_1 in F;
```

given collection is returned. More details about VEAM algorithm can be found in [7, 18].

In order to show the usefulness of using M-FASs for image (graph) classification, a comparison between the use, as attributes, of all patterns computed by VEAM [7] against the M-FASs computed by M-VEAM for image classification, is shown. Using the M-FASs computed by M-VEAM we build attribute vectors to represent the images of the collection. An image is represented as an attribute

**Algorithm 3**: *appLSet*

**Input**: $T$ : A candidate graph
$MV$ : Substitution matrix indexed by $L_V$
$ME$ : Substitution matrix indexed by $L_E$
$G = (V, E, I, J)$ : A graph of the collection
$G'$ : Embedding of $T$ in $G$
$e = \{u, v\}$ : An extension of $G'$
$\tau$ : similarity threshold.
**Output**: $CL$ : A set of candidate 2-tuples $(elabel, vlabel)$.

```
1 foreach j ∈ U_E^τ(J(e)) do
2     scoreE ← S_max(T, G') * ME_{j,J(e)} / ME_{j,j};
3     if e is a forward extension of G' then
4         foreach i ∈ U_V^τ(I(v)) do
5             if i is less than or equal to the largest of the vertex labels of T then
6                 score ← scoreE * MV_{i,I(v)} / MV_{i,i};
7                 if score ≥ τ then CL ← CL ∪ {(j, i)};
8     else if scoreE ≥ τ then CL ← CL ∪ {(j, ∅)};
```

vector $V = (v_1, \ldots, v_n)$ where the number of columns $n$ is the amount of maximal patterns computed by M-VEAM. The value of each attribute $v_i$ $(1 \leq i \leq n)$ is the maximum similarity between the pattern $i$ and the image. Thus, a matrix where the row number is the number of graphs (images) in the collection is built, and the element of each row is the attribute vector which represents the corresponding image.

Two image databases are used in this experiment: *GREC* [19] that contains images of electronic and architectonic plane symbols grouped into 22 classes. This database was split into 572 (52%) images for training and 528 for testing; and *CoenenDB* that contains synthetic images, taken from the Random image generator of Coenen [3], that represents two landscape views; CoenenDB was split into 1200 (60%) images for training and 800 for testing. In both databases, each image is represented as a graph: in GREC, several critical points were selected and used as vertices to build a graph and the edges contain vertex spacial information. For CoenenDB, a tree for each image using a quad-tree method [20] was created and the information of the leaves of these trees was used to build a graph.

In Table 1, the number of patterns used as attributes for classification are compared. These patterns are obtained using $\tau = 40\%$ in the CoenenDB database and $\tau = 8\%$ in the GREC database. These values were computed as the mean of the similarities among the graphs of the collection. This table is split into two sub-tables, one for CoenenDB and for GREC collections, respectively. The first column of each sub-table shows the support value used, and the other two

[3] www.csc.liv.ac.uk/∼frans/KDD/Software/ImageGenerator/imageGenerator.html

consecutive columns show the number of patterns (computed by M-VEAM and VEAM algorithms respectively) used as attributes for classification, and the third column shows the reduction percentage achieved using only maximal patterns.

**Table 1.** Number of patterns used as attributes in the classification process.

| CoenenDB | | | | GREC | | | |
|---|---|---|---|---|---|---|---|
| **support ($\delta$)** | **M-VEAM** | **VEAM** | **Reduction** | **support ($\delta$)** | **M-VEAM** | **VEAM** | **Reduction** |
| 20% | 437 | 745 | 41.34% | 2% | 1190 | 1422 | 16.32% |
| 25% | 186 | 330 | 43.64% | 3% | 607 | 715 | 15.10% |
| 30% | 86 | 143 | 39.86% | 4% | 366 | 437 | 16.25% |

As we can see in Table 1, using the subgraphs computed by M-VEAM produces a reduction in the amount of subgraphs used as attributes for classification, compared against the patterns computed by VEAM. In this table, we can see a reduction ranging from 15% to 43%.

The next experiment evaluates the classification results reached using the maximal FASs computed by M-VEAM compared against the results obtained using all FASs as attributes for classification. We summarize the classification results of our experiments in Table 2, which is subdivided in two sub-tables: one shows the accuracy results and the other shows the F-measure results, in the same order. The first and second columns of these sub-tables show the collection name and the support threshold values used in this experiment, respectively. The other four consecutive columns show the classification results (accuracy or F-measure), for the classifier specified in the top of these columns, using only the M-FASs computed by M-VEAM and all FAS computed by VEAM, respectively. Notice that the best results appear boldfaced.

As we can see in Table 2, the results achieved with our proposal are competitive regarding to the results obtained using all patterns computed by VEAM. In the CoenenDB database, the best classification result was obtained by VEAM using the J48graft classifier with an accuracy of 97.25, and using the Regression classifier, M-VEAM obtained the same value. According to the F-measure, M-VEAM obtained the best result using the Regression classifier with an F-measure of 97.26. In this database, using the patterns computed by M-VEAM as attributes, a reduction of 43% was achieved. In the GREC database, the best classification result was obtained using the patterns computed by VEAM jointly with the SVM classifier, obtaining an accuracy of 94.51, while using the patterns computed by M-VEAM we got an accuracy of 93.61 also with the SVM classifier. In this database, the patterns computed by M-VEAM allow a dimensionality reduction of 16% regarding the number of patterns computed by VEAM.

In addition, in Table 3, we present a statistical comparison for all pairwise comparisons between our proposal using M-FASs as attributes and the option of using all patterns computed by VEAM. For this comparison, we use a significant

**Table 2.** Classification results (%) using several classifiers.

**(a) Accuracy results achieved using several $\delta$ values**

| Collection | $\delta$ | J48graft | | Decision Table | | Regression | | SVM | |
|---|---|---|---|---|---|---|---|---|---|
| | | M-VEAM | VEAM | M-VEAM | VEAM | M-VEAM | VEAM | M-VEAM | VEAM |
| *CoenenDB* | 20% | 96.38 | **97.25** | **95.38** | 94.38 | **97.25** | 96.25 | **95.50** | 95.38 |
| | 25% | 95.50 | **96.75** | **94.00** | 80.13 | 96.25 | **96.38** | 93.63 | **94.38** |
| | 30% | 95.50 | **96.50** | **95.63** | 95.25 | 96.38 | **96.50** | **95.50** | 95.13 |
| Average | | 95.79 | **96.83** | **95.00** | 89.92 | **96.63** | 96.38 | 94.88 | **94.96** |
| *GREC* | 2% | **53.98** | 45.45 | **57.77** | 33.90 | **75.57** | 73.48 | 93.61 | **94.13** |
| | 3% | 77.52 | **82.20** | 63.64 | **65.72** | 77.65 | **83.14** | 93.37 | **94.51** |
| | 4% | 77.14 | **81.63** | 59.33 | **68.37** | 81.11 | **82.95** | 92.86 | **94.13** |
| Average | | 69.55 | **69.76** | **60.25** | 56.00 | 78.11 | **79.86** | 93.28 | **94.26** |

**(b) F-measure results achieved using several $\delta$ values**

| Collection | $\delta$ | J48graft | | Decision Table | | Regression | | SVM | |
|---|---|---|---|---|---|---|---|---|---|
| | | M-VEAM | VEAM | M-VEAM | VEAM | M-VEAM | VEAM | M-VEAM | VEAM |
| *CoenenDB* | 20% | 96.34 | **97.23** | **95.43** | 94.49 | **97.26** | 96.21 | **95.51** | 95.39 |
| | 25% | 95.47 | **96.73** | **93.94** | 82.51 | 96.22 | **96.33** | 93.57 | **94.35** |
| | 30% | 95.43 | **96.46** | **95.76** | 95.33 | 96.35 | **96.50** | **95.47** | 95.06 |
| Average | | 95.75 | **96.81** | **95.04** | 90.78 | **96.61** | 96.35 | 94.85 | **94.93** |
| *GREC* | 2% | **52.00** | 38.00 | **16.90** | 11.76 | 55.32 | **79.17** | 91.91 | **93.33** |
| | 3% | 81.95 | **86.96** | 27.54 | **28.13** | 74.70 | **78.43** | **93.63** | 89.36 |
| | 4% | 61.32 | **78.43** | 25.69 | **34.29** | 74.68 | **76.00** | 86.11 | **86.96** |
| Average | | 65.09 | **67.80** | 23.38 | **24.73** | 68.23 | **77.87** | **90.55** | 89.88 |

statistical test known as Bergmann test [21]. The value for $\alpha$ used on this test was 0.05.

**Table 3.** Statistical significance results achieved for different classifiers in two image (graph) collections.

| | **CoenenDB** | **GREC** |
|---|---|---|
| **Test/Classifier** | M-VEAM vs. VEAM | M-VEAM vs. VEAM |
| J48graft | VEAM | – |
| Decision-Table | M-VEAM | – |
| Regression | – | – |
| SVM | – | – |

In table 3, the first column of these sub-tables shows the classifiers used in each comparison and columns 2 and 3 show the results for the CoenenDB and GREC image databases, respectively. These columns show the approach that is significant better than the other according to the Bergman test; the symbol "–" indicates that there is not a statistical significant difference between the results of both approaches.

As we can see from Table 3, the use of M-FASs as attributes is a good option since the dimensionality is reduced and we obtain similar classification results than using all the FASs computed by VEAM.

## 6 Conclusions

Frequent approximate subgraph mining is a widely used technique in Data Mining applications where there is some distortion into the data. However, usually a large number of frequent patterns is computed. Using only representative patterns as attributes instead of using all the patterns is a technique that can be used to reduce the dimensionality of the object descriptions (representation space). Therefore, the aim of our research work is to develop new algorithms for mining representative FAS that allows us improving the classification efficiency and efficacy when this type of subgraphs are used as attributes.

In this paper, we present, as preliminary results of this research work, a modification of a FAS mining algorithm of the state-of-the-art, for computing only maximal FASs in graph collections. The experiments show that using only maximal patterns as attributes, instead of all patterns computed by VEAM, allows obtaining similar classification results, while reducing the dimensionality and removing redundant patterns from the set of patterns used as attributes for image classification.

As future work, we are going to keep developing this research proposal for achieving the specific objectives and general goals.

**Acknowledgment.** This work was partly supported by the National Council of Science and Technology of Mexico (CONACyT) through the project grants $CB2008$-106443 and $CB2008$-106366; and the scholarship grant 287045.

## References

1. Jiang, C., Coenen, F., Zito, M.: A survey of frequent subgraph mining algorithms. Knowledge Engineering Review (2012)
2. Inokuchi, A., Washio, T., Motoda, H.: An Apriori-based Algorithm for Mining Frequent Substructures from Graph Data. In: Proceedings of the 2000 European Symposium on the Principle of Data Mining and Knowledge Discovery (PKDD'00), France, Lyon (2000) 13–23
3. Huan, J., Wang, W., Prins, J.: Efficient mining of frequent subgraphs in the presence of isomorphism. In: The 3rd IEEE International Conference on Data Mining, FL, Melbourne (2003) 549–552
4. Yan, X., Huan, J.: gSpan: Graph-Based Substructure Pattern Mining. In: International Conference on Data Mining, Japan, Maebashi (2002)
5. Borgelt, C.: Mining molecular fragments: Finding relevant substructures of molecules. In: Proc. IEEE International Conference on Data Mining (ICDM), Maebashi City, Japan, IEEE Press (2002) 51–58
6. Nijssen, S., Kok, J.: A Quickstart in Frequent Structure Mining can make a Difference. In: The 10th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2004) 647–652
7. Acosta-Mendoza, N., Gago-Alonso, A., Medina-Pagola, J.: Frequent approximate subgraphs as features for graph-based image classification. Knowledge-Based Systems **27** (2012) 381–392

8. Holder, L.B., Cook, D.J., Bunke, H.: Fuzzy substructure discovery. In: ML92: Proceedings of the ninth international workshop on Machine learning, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1992) 218–223
9. Jia, Y., Zhang, J., Huan, J.: An efficient graph-mining method for complicated and noisy data with real-world applications. Knowledge Information Systems **28**(2) (2011) 423–447
10. Song, Y., Chen, S.S.: Item sets based graph mining algorithm and application in genetic regulatory networks. Data Mining, IEEE International Conference on Volume, Issue (2006) 337–340
11. J.L., Zou, Z., Gao, H.: Mining frequent subgraphs over uncertain graph databases under probabilistic semantics. VLDB J. **21**(6) (2012) 753–777
12. Acosta-Mendoza, N., Gago-Alonso, A., Carrasco-Ochoa, J., Martínez-Trinidad, J., Medina-Pagola, J.: Feature Space Reduction for Graph-Based Image Classification. In: Proceedings of the 18th Iberoamerican Congress on Pattern Recognition (CIARP'13). Volume Part I, LNCS 8258., Havana, Cuba, Springer-Verlag Berlin Heidelberg (november 2013) 246–253
13. Kimelfeld, B., Kolaitis, P.: The complexity of mining maximal frequent subgraphs. In: Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2013, ACM (2013) 13–24
14. Yan, X., Han, J.: ClosedGraph: Mining Closed Frequent Graph Patterns. In: Proc. of the 9th ACM SIGKDD of International Conference on Knowledge Discovery and Data Mining (KDD), Washington, DC (2003) 286–295
15. Takigawa, I., Mamitsuka, H.: Efficiently Mining $\delta$-tolerance Closed Frequent Subgraphs. Machine Learning **82**(2) (2011) 95–121
16. Zou, Z., Li, J., Gao, H., Zhang, S.: Finding top-k maximal cliques in an uncertain graph. In: IEEE 26th International Conference on Data Engineering (ICDE 2010). (2010) 649–652
17. Ozaki, T., Etoh, M.: Closed and maximal subgraph mining in internally and extermally weighted graph databases. In: Proceedings of the IEE Workshops of International Conference on Adbanced Information Networking and Applications., IEEE Computer Society. (2011) 626–631
18. Acosta-Mendoza, N., Gago-Alonso, A., Medina-Pagola, J.: On speeding up frequent approximate subgraph mining. In: Proceedings of the 17th Iberoamerican Congress on Pattern Recognition (CIARP'12). Volume LNCS 7441., Buenos Aires, Argentina, Springer-Verlag Berlin Heidelberg (2012) 316–323
19. Riesen, K., Bunke, H.: IAM Graph Database Repository for Graph Based Pattern Recognition and Machine Learning, Orlando, USA (2008) 208–297
20. Finkel, R., Bentley, J.: Quad Trees: A Data Structure for Retrieval on Composite Keys. Acta Informatica **4** (1974) 1–9
21. Bergmann, G., Hommel, G.: Improvements of general multiple test procedures for redundant systems of hypotheses. In: P.Bauer, G. Hommel, and E. Sonnemann, editors, Multiple Hypotheses Testing, Springer, Berlin (1988) 100–115

# Hardware Acceleration of Frequent Itemsets Mining on Data Streams

Lázaro Bustio-Martínez[1,2], René Cumplido-Parra[2], Raudel Hernández-León[1], and Claudia Feregrino-Uribe[2]

[1] Advanced Technologies Application Center.
$7^a$ ♯ 21812 e/ 218 and 222, Rpto. Siboney, Playa, C.P. 12200, Havana, Cuba.
{lbustio,rhernandez}@cenatav.co.cu
[2] National Institute for Astrophysics, Optics and Electronic.
Luis Enrique Erro No 1, Sta. Ma. Tonantzintla, 72840, Puebla, México.
{rcumplido,cferegrino}@ccc.inaoep.mx

**Abstract.** In recent times, processing of data streams is gaining the attention of the scientific community due to its practical applications. Data stream is an unbounded and infinite flow of data arriving at high rates and, therefore, the classical data mining approaches can not be used straightforward in this scenario. Because of this, finding alternatives to achieve better results in the discovering of frequent itemsets on data streams is an active research topic. One of such alternatives is to develop single-pass parallels methods that can be implemented in hardware to take advantage of the inner parallelism of such devices. In this paper, a new method that can mine high incoming rates data streams is presented. As preliminary results, the proposed methods can mine in exhaustive fashion the incoming data streams when its number of single items is low. When the number of single items is high, the proposed method obtains an approximate solution with no false positives itemset produced.

**Key words:** Frequent itemset mining, data stream mining, systolic tree, custom hardware architectures.

## 1 Introduction

In recent years, there has been an explosion on the amount of data generated by all sort of human activities. In order for this data to be useful, it must be processed to obtain hidden knowledge. To perform this task, several approaches have been proposed and implemented mainly in software-based systems that offer limited performance when processing large amounts of data.

Data Mining aims to pride the tools and techniques needed to face such immense data volumes. In Data Mining is extremely useful to record all the occurrences of certain patterns and that is what frequent itemsets mining performs. Frequent itemsets are those sets of data items that can be found always together more than a given number of occurrences in data. In other words, the goal of frequent itemsets mining is to determine which elements in a database (or any other data source) commonly appear together.

One scenario that is gaining a lot of attention of researchers is the data streams mining. Analyzing data streams is an emerging need, and it can be found in video and audio streams, network traffic, commercial transactions, etc, but those applications need to be as fast as they can so hardware-based approaches have been proposed. Frequent itemsets mining in hardware for data streams addresses new challenges and only in [4] is conducted a research to frequent itemsets mining on data streams.

This paper is structured as follow: in the next section, the theoretical basis that support this research is presented. A review of state-of-the-art is addressed in section 3 while section 4 presents the methodological foundations of this research. The preliminary results are shown in section 5 while this paper is concluded in section 6. Also in section 6 the future works is drafted.

## 2 Theoretical basis

Let $I = \{i_1, i_2, .., i_n\}$ be a set of items:

**Definition 1 (Itemset).** *A itemset $X$ is a set of items over $I$ such $X = \{i_i, ..., i_k\} \subseteq I$.*

**Definition 2 (Transaction).** *A transaction $T$ over $I$ is a couple $T = (tid, I)$ where tid is the transaction identifier, and $I$ is a $X \subseteq I$ itemset.*

**Definition 3 (Support).** *The support of an itemset $X$ is the number of transactions that contains $X$.*

An itemset is called $frequent$ if its support is no less than a given absolute minimal support threshold $\phi_{abs}$, with $0 < \phi_{abs} \leq |D|$, while $D$ is the mining database.

**Definition 4 (Data streams).** *A data stream is a continuous, unbounded and not necessarily ordered, real-time sequence of data items.*

Three characteristics appear in data streams: (1) Items in stream arrive continuously at a high rate (*Continuity*); (2) Items can be accessed and processed just once by the processing units in data streams (*Expiration*) and (3) The only assumption that we can make about bounds of streams is that the total number of data is unbounded and potentially infinite (*Infinity*).

**Definition 5 (Window).** *A window in a data stream is an excerpt of items that pertain to the stream.*

Windows can be created using one of this three approaches: (1)*Landmark window model*, (2)*Damped window model* and (3)*Sliding window model*. The Landmark Window Model employs some points (called landmark) to start recording where a transaction begins and ends. The support count of an itemset in this model is the number of transactions containing it between the landmark and the

current time. To distinguish between the oldest and new transactions a variation of this model was proposed and named Damped Window Model. Damped Window Model assigns different weights to transactions where the recent ones have weight near to 1, and older ones have weight near to 0. As time passes, the weight of each transaction will be degraded. The Sliding Window Model uses only the latest $W$ transactions in the mining process. As the new transactions arrives, the old ones in the sliding windows are excluded. The use of this model impose a restriction: as some transactions will be excluded of the mining process, methods for finding expired transactions and for discounting the support count of the itemsets involved are required.

### 2.1 Reconfigurable Computing

Reconfigurable Hardware Computing is referred to the use of hardware devices in which the functionality of the logic gates is customizable at runtime, and FPGAs are the main exponent of this approach. The architecture of a FPGAs is based on a large number of logic blocks which perform basic logic functions. Because of this, an FPGA can implement from a simple logical gate, to a complex mathematical function. FPGAs can be reprogrammed; that is; the circuit can be "erased" and then, a new architecture that implements a brand new algorithm can be implemented. This capability of the FPGAs allows the creation of fully customized architectures, reducing cost and technological risks that are present in traditional circuits design.

Although there are other hardware development platforms for data streams mining (such as Graphics Processing Units, named GPUs), FPGAs are better suited. GPUs are graphic accelerators which are interfaced by the PCI port; while FPGAs can be interfaced by the PCI port, the USB port or the Ethernet connector (Ethernet interface is ideal for network stream analysis). Due to the high incoming rates of items in data streams, the processing such items must be done as fast as it can. FPGAs are great for real-time systems, where even 1ms of delay might be too long, and this capability are extremely valuable for mining data streams. GPUs are ideal for hybrid applications where some instructions must be accelerated while FPGAs can accelerate the whole process. Because of this, FPGA is better suited to be chosen as a development platform for accelerate frequent itemsets mining on data streams.

## 3 Algorithms review in hardware

Hardware implementations of algorithms take advantage of inner parallelism of the hardware device used. In consequence, such devices gain every day more attention to be employed as development platforms. After a proper review of the state-of-the-art, it can be organized as it is shown in table 1.

Analyzing the revised literature it can be noticed that frequent itemsets mining on data streams using reconfigurable hardware is an interesting research area so, it is worth to propose new parallels algorithms to face such task. In

**Table 1.** Algorithms and architectures for frequent itemsets mining in data streams using FPGAs. DB stands for *Database*; Apr for *Apriori* and FPG for *FP-Growth*.

| **Title** | **Based** | **Source** |
|---|---|---|
| An Architecture for Efficient Hardware Data Mining Using Reconfigurable Computing Systems. [2] | Apr | DB |
| Hardware Enhanced Mining for Association Rules. [5] | Apr | Stream |
| Hardware-Enhanced Association Rules Mining With Hashing and Pipelining. [12] | Apr | DB |
| Novel Strategies for Hardware Acceleration of Frequent Itemset Mining With the Apriori Algorithm. [11] | Apr | DB |
| Mining Association Rules with Systolic Trees.[9] | FPG | DB |
| A Reconfigurable Platform for Frequent Pattern Mining.[8] | FPG | DB |
| A Highly Parallel Algorithm for Frequent Itemset Mining. [6] | FPG | DB |
| Design and Analysis of a Reconfigurable Platform for Frequent Pattern Mining. [10] | FPG | DB |
| An FPGA-Based Acceleration for Frequent Itemset Mining. [13] | Eclat | DB |
| FPGA Acceleration for Intersection Computation in Frequent Itemset Mining. [7] | Eclat | DB |

this task, there are three main approaches: algorithms that use Apriori as the starting point, algorithms that use FP-Growth and those that use Eclat.

The algorithms that mimic the Apriori-based schemes in hardware require loading the candidate itemsets and the database into the hardware. This strategy is limited by the capacity of the chosen platform: if the number of items to manage is larger than the hardware capacity the items must be loaded separately in many consecutive times degrading performance. In consequence, the support counting must be executed several times, and this is a very time consuming approach. In addition, several candidates itemsets and a large database may cause a bottleneck in the system. This issues are forbidden in data streams mining.

As well as Apriori-based algorithm, the FP-Growth-based algorithms need to download the mining database to FPGA. They also need two passes over the database except Mesa et al. [6] but this one still need to download the database to the hardware device. This is impractical in data stream mining scenario due to the Expiration restriction. Like others reviewed algorithms, authors focused their attention in better data structures rather than substantial theoretical contributions. As rule, FP-Growth based algorithm can handle a limited number of itemsets, less than 11 in the better cases which is inadequate for real-life applications. Nevertheless, those algorithms based on FP-Growth use the FP-Tree data structure which is very well suited for data stream mining applications.

Eclat-based algorithm uses the vertical database representation in order to save memory and processing time. It use the intersection of items to compute the support, and it is more efficient than hash-trees. All the Eclat-based imple-

mentations propose an hybrid approach, where the most consuming functions were download to hardware while software controls the execution flow and data structures. Due to the bandwidth limitations of used hardware devices, very large transactions must be segmented. In the reviewed papers, no segmentation strategies were reported. Although the vertical database representation allows to save memory and processing time, it is not compatible with the Expiration restriction. Also, the pruning strategy in Eclat is inefficient and introduces delays that affect the performance of the algorithms. This two issues make Eclat impractical to be used as a starting point for data stream mining algorithms.

## 4 Methodological foundations

### 4.1 Research problems

Modern applications generate huge data volumes in data streams way. Due to the increase of this kind of applications it is necessary obtain useful knowledge from those data streams. As it was previously defined, a data streams are a continuous, ordered and potentially infinite sequence of items in real time where data arrives without interruptions at a high speed. Also, data can be accessed only once, and the only assumption that we can make about bounds of streams is that the total number of data is unbounded. It is unrealistic to store all items of data streams to process them offline. These characteristics impose extra difficulties to algorithms and systems that process such data sources.

Due to the high incoming rate, the impossibility to store the data and the huge volumes of items in streams, software that analyzes such data streams can not process exhaustively all items. The supporting hardware and software are not capable to deal with such intense processing. Instead, commercial applications that mine data stream use an "approximate" processing approach. That is, they do not analyze all items that are present in a flow; instead, they use some heuristic or probabilistic approach to determine which item is the most likely to contain the desired information. There are applications that need intense processing requirements, e.g. intrusion detection systems or network analysis systems. In this kind of applications, the immediate data analysis and near-real-time response are extremely valuable. To fulfill these requirements is needed to propose new parallel algorithms running on high-performance computing devices such as FPGAs. FPGAs can perform tasks in a high parallel fashion, and this is very useful in data streams processing applications.

Frequent itemsets mining is one technique that is commonly used in data knowledge extraction and have been used with success in databases scenario. To mine frequent itemsets in data streams efficiently, an alternative would be to develop new parallel approaches that use custom hardware architectures. In the reviewed literature, there is only one architecture to mine frequent itemsets on data streams [5].

### 4.2 Aims and expected contributions

The general aim of this research work is: *To develop parallel methods for frequent itemsets mining in data streams that outperform the state-of-the-art algorithms for data streams analysis and that are suitable for being implemented in hardware-accelerated platforms. The proposed methods must outperform in one order of magnitude (at least) the state-of-art algorithms implemented in software.*

To fulfill the general aim, some specific aims were proposed: (1) To propose a flexible method for separating the incoming data stream into windows that it can be used by the support counting algorithm; (2) To adopt data structures that can be used in frequent itemsets mining on data streams; (3) To develop new algorithms for frequent itemsets mining that use the separation method selected and the data structures adopted; (4) To obtain parallel hardware implementation of the algorithms mentioned above that can perform frequent itemsets mining at least 1 order of magnitude faster (without compromising effectively) than state-of-the-art software implementations.

As results of this research, the following contributions are expected: (1) A new method for frequent itemsets mining on data streams; (2) A design of parallel one-pass algorithms to mine frequent itemsets on data streams and (3) A custom hardware architecture that implements the proposed algorithms. This custom architecture will take advantage of inner parallelism provided by the hardware device used in its implementation.

## 5 Preliminary Results

After the literature was reviewed, the conclusion observed is that the selected window model should not be an issue: our method, and therefore, the hardware designs derived, must work fine regardless of the window model selected. So, the window model will be an input parameter.

The basic idea of the presented method is to develop a tree structure of processing units where the itemsets in data streams flow from the root node to leaf nodes. The tree structure presented is named *systolic tree* and each node has one child and one sibling. For leaf nodes, the child and sibling nodes are null nodes. In this structure, the child node contains, as a prefix, the itemset handled by its parent. Fig. 1 represents the systolic tree.

Using the Apriori property, which states that any subset of frequent itemset must be frequent [1] if a node is regarded as frequent then its parent is frequent too with equal or greater frequency counting. This property is specially useful for frequent itemset selection strategy.

The size (in number of nodes) of the systolic tree is determined by the capacity of the development platform. Assuming that the development platform contains enough computational resources, the size (in number of nodes) of the systolic tree will be $k = 2^n - 1$. The nodes in the systolic tree have its own processing logic, which it is presented in Algorithm 1. The systolic tree data structure presented in this paper implements a distributed control scheme: the

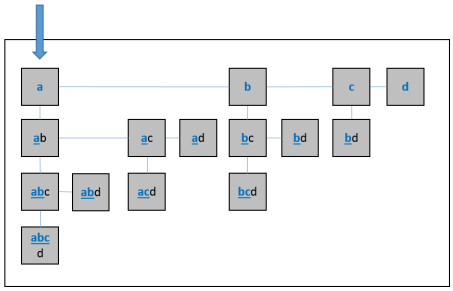



**Fig. 1.** Systolic tree data structure to mine frequent itemsets on data streams.

processing and control logic are distributed in each node of the systolic tree. This allows saving computational resources due to the logic reduction.

When the data stream arrives, each transaction is flowed into the systolic tree to determine the frequency of each item. Algorithm 1 depict the frequency counting scheme proposed. This algorithm will be executed in parallel in each node of the systolic tree. After the frequency of each itemset is calculated, those itemsets that can be regarded as frequent are determined using a backtracking strategy and the Apriori property.

The proposed method is designed to be implemented in a custom hardware architecture. To validate the concept introduced in this research, it was programmed sequentially in software using C# language over the .Net Framework platform.

As it was explained earlier, the systolic tree can process a limited number of items determined by capacity of the hardware device used. If the chosen development platform can hold a systolic tree with 1024 nodes, the maximum number of different items in the incoming transactions that it can be process will be 10. If the number of different items in the incoming transaction is greater, some itemsets will be not processed and therefore, the mining process will be approximate with no false positives produced. However, if the development platform can hold all the possibles itemsets, the mining process will be exact.

Some experiments were conducted and the pursued objectives were to verify the correct performance of the proposed algorithms and to measure how the systolic tree grows according to the length of incoming transactions. To accomplish these goals, MSNBC dataset from UCI repository [3] was used. This dataset is click-stream data that contains 989,818 sequences. The number of distinct item in this dataset is 17. The average number of itemsets per sequence is 13.33. The average number of distinct item per sequence is 5.33. In order to validate that the frequency counting computed by algorithm 1 is correct, it was assumed that the systolic tree can handle all possibles itemsets for used dataset. It is important to notice that for various selected support values, the conducted experiments demonstrate that the frequent itemsets detected by the proposed method and its frequency countingwas the same that the one obtained by the baseline FP-Growth.

**Algorithm 1**: Frequency counting.

**Input**: Transaction's window
**Output**: Systolic tree with the counting frequency of each itemset.

```
1  n_i ⟵ systolyc_tree.RootNode;
2  foreach itemset S_i in window_buffer do
3      Flush S_i into n_i;
4      if n_i.IsOccupied == false then
5          n_i.IsOccupied = true;
6          n_i.Label.Add(S_i[0]) n_i.Counter++;
7          S̃_i = S_i.Exclude(n_i.Item) if S̃_i.IsEmpty == false then
8              StartParallalelBlock:
9              n_i ⟵ n_i.ChildNode;
10             Flush S̃_i to n_i and go to step 4;
11             n_i ⟵ n_i.SiblingNode; Flush S̃_i to n_i and go to step 4;
12             EndParallelBlock;
13     else
14         if S_i.Contain(n_i.Label) == true then
15             n_i.Counter++;
16             S̃_i = S_i.Exclude(n_i.Item);
17             if S̃_i.IsEmpty == false then
18                 StartParallalelBlock:
19                 n_i ⟵ n_i.ChildNode; Flush S̃_i to n_i and go to step 4;
20                 n_i ⟵ n_i.SiblingNode; Flush S̃_i to n_i and go to step 4;
21                 EndParallelBlock;
22         else
23             n_i ⟵ n_i.SiblingNode; Flush S_i to n_i and go to step 4;
24 return systolic_tree;
```

Fig. 2 shows graphically the results obtained. Experiments show that the systolic tree grows exponentially (and therefore memory consumptions grow exponentially too) concerning to the length of the incoming transactions. This effect can be attenuated using larger hardware devices or using external memories, but it is still an issue to be taken in account. The processing time grows linear concerning of the length of the incoming transactions. These values are calculated for the sequential software implementation. The proposed method is designed to be implemented in parallel so the processing will be executed simultaneously, and after some initial time, the results will arrive continuously.

The software implementation of the proposed method pursuits the main objective of determining whether it is a valid solution for frequent itemset mining on data streams while in future works, hardware implementations will be developed. Experiments demonstrate that for different support values, the frequent itemsets and its frequency counting are the same that obtained by the baseline

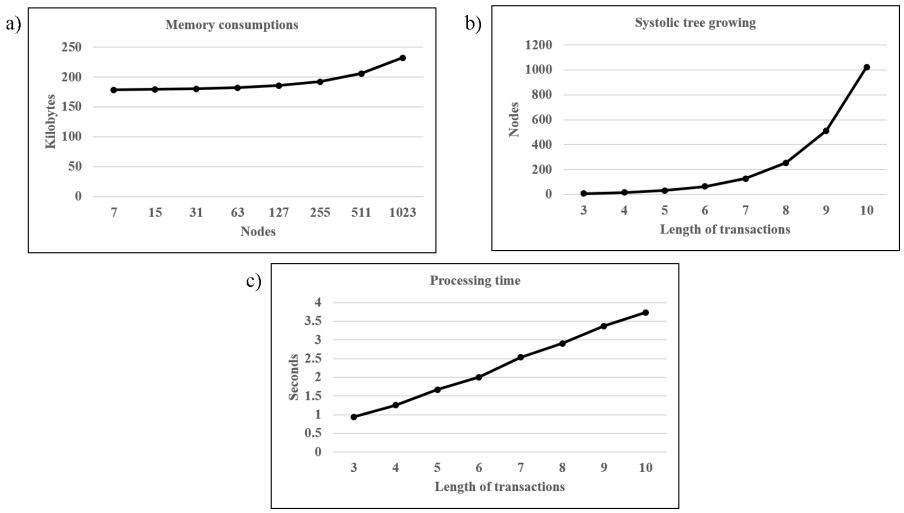



**Fig. 2.** Behavior of the proposed method while the length of the incoming transactions grows. a) Memory consumption graph. b) Systolic tree size graph and c) Processing time graph.

software. The length of the incoming transactions, and therefore the systolic tree size, can affect these results. In this case, not all of the frequent itemsets will be returned, but those itemsets that are regarded as frequents by the proposed method will be regarded as frequent with the same frequency counting by the baseline FP-Growth. In other words, if the available computing resources of the development platform selected can handle any length of the incoming transactions, the mining process will be exact. Otherwise, the mining process will be approximate with no false positives. The software implementation validates the correct functioning of the proposed method and allows to understand its functioning before implement it in hardware.

## 6 Conclusions and future work

Frequent itemset mining is a widely used Data Mining technique with outstanding results in database scenario. Data stream mining is a recent research field where frequent itemsets are introducing. Due to the continuity, expiration and infinity characteristic of data streams it is necessary to explore alternatives that allow to increase the efficiency of the mining process in such datasets. One alternative could be the design of parallel algorithms to be implemented in custom hardware architectures.

This paper introduce a new parallel method for frequent itemset mining in data streams which is designed to be implemented in a custom hardware architecture. The proposed method implements a distributed control logic among all processing nodes, and each node execute the same algorithm. Some experiments

were conducted, and it can be concluded that the proposed method correctly performs this task. When it is executed in a device with no resources restrictions then the exact mining process is performed. By the contrary, when restrictions are imposed, then the approximate mining process with no false positives is performed. From the experiments conducted it is derived that some adjustments must be done to the proposed method in order to save computational resources of the selected hardware device.

In future works, the implementation in hardware is mandatory. Also, a preprocessing strategy in order to determine 1-frequent itemsets which will be flowed into hardware architecture is the next step: this allow to optimize the nodes consumption in systolic tree. A segmentation database strategy that allow to handle larger datasets is an issue to deal with, and it is currently studying.

## References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc. (1994)
2. Baker, Z., Prasanna, V.: An architecture for efficient hardware data mining using reconfigurable computing systems. In: Proc. of the 14th Annual IEEE Symposium on Field-Programmable Custom Computing Machines (2006)
3. Frank, A., Asuncion, A.: UCI machine learning repository (2010), `http://archive.ics.uci.edu/ml`
4. Lin, C.H., Chiu, D.Y., Wu, Y.H., Chen, A.: Mining frequent itemsets from data streams with a time-sensitive sliding window. In: SDM (2005)
5. Liu, W.C., Liu, K.H., Chen, M.S.: Hardware enhanced mining for association rules. In: Proc. of the 10th Pacific-Asia Conf. on Advances in Knowl. Disc. and Data Mining (2006)
6. Mesa, A., Feregrino-Uribe, C., Cumplido, R., Hernandez-Palancar, J.: A highly parallel algorithm for frequent itemset mining. In: Advances in Pattern Recognition, LNCS, vol. 6256 (2010)
7. Shi, S., Qi, Y., Wang, Q.: Fpga acceleration for intersection computation in frequent itemset mining. In: Cyber-Enabled Distributed Computing and Knowl. Disc. (CyberC), 2013 Int. Conf. on (2013)
8. Sun, S., Steffen, M., Zambreno, J.: A reconfigurable platform for frequent pattern mining. In: Proc. of the 2008 Int. Conf. on Reconfigurable Computing and FPGAs (2008)
9. Sun, S., Zambreno, J.: Mining association rules with systolic trees. In: FPL (2008)
10. Sun, S., Zambreno, J.: Design and analysis of a reconfigurable platform for frequent pattern mining. IEEE Trans. on Parallel and Distributed Systems 22 (2011)
11. Thöni, D.W., Strey, A.: Novel strategies for hardware acceleration of frequent itemset mining with the apriori algorithm. In: 19th Int. Conf. on Field Programmable Logic and Applications (2009)
12. Wen, Y.H., Huang, J.W., Chen, M.S.: Hardware-enhanced association rule mining with hashing and pipelining. IEEE Trans. on Knowl. and Data Eng. 20(6) (Jun 2008)
13. Zhang, Y., Zhang, F., Jin, Z., Bakos, J.D.: An fpga-based accelerator for frequent itemset mining. ACM Trans. Reconfigurable Technol. Syst. 6(1) (May 2013)

# Research on Power Line Inspection by Visual Based Navigation

Alexander Cerón[1,2], Iván Mondragón[3], and Flavio Prieto[1]

[1] Universidad Nacional de Colombia - Sede Bogotá,
Carrera 30 No 45 - 03, Bogotá, Colombia.
{aceronco,faprietoo}@unal.edu.co,
[2] Universidad Militar Nueva Granada,
Carrera 11 No. 101-80 Bogotá, Colombia.
[3] Pontificia Universidad Javeriana, Department of Industrial engineering,
Carrera 7 No 40-69, Bogotá, Colombia.
imondragon@javeriana.edu.co

**Abstract.** In this paper, the advances obtained through a research process based on the development of a visual based navigation model for an UAV (Unmanned aerial vehicle) with application in power line tracking for inspection are presented. The main contribution expected is the method for line tracking through UAV navigation. In this process, a new line detection method based on a computer graphics process was created. The method is validated with synthetic and real images. Additionally, a virtual environment for real time simulation and line detection was created. The results obtained in the line detection process are promising.

**Key words:** Inspection, power line detection, navigation, UAV.

## 1 Introduction

Countries like Spain, United Kingdom, China and Australia spent efforts and resources in the development and implementation of technologies for the power line inspection. As well as these, and considering that, Colombia dedicate an important part of its economic development to the natural resources utilization, specially electric energy, it is important to make efforts in developing methods for improving the different processes of electrical infrastructure inspection and maintenance.

According to the power line failure, there are different methods for power line inspection; the main methods are: manual, manned flights and, recently, UAVs; see Figure 1.

Power line detection is an important task in the inspection of electrical infrastructure prior to maintenance, for this reason, there is an interest in developing methods that reduce costs, risks and the logistic problems of manual inspection including manned flights [1], [2], [3] by using UAVs [4]. The UAVs can be used for capturing images from different views that have to be processed in order to detect power lines [5]. It is good to mention that there are companies that

(a) Manual[4]

(b) UAVs[5]

**Fig. 1.** Types of power line inspection.

offer different services and products for electrical inspection, including UAVs; see Figure 1.

There are different methods for line detection that have been used in power line detection for UAV images. These are based on edge and ridge detectors, magnitude and gradient orientation, voting schemes, line support regions, growing regions and chain codes [6–11]. The classical method for line detection is the Hough transform [7], which can detect lines in well contrasted and segmented images. This method was used in combination with a PCNN (Pulse-coupled Neural Network) for removing background and clustering for power line detection [12]. In the work of Zhang [13], a process for power line detection and tracking based on Hough transform with Kalman filter is presented. In this case, they use the Otsu threshold method obtaining better results than PCNN filters.

It is good to mention that some works of line detection have a post processing stage for connecting line segments or cluster interest lines [13], [14]. Recently, a method for line detection based in region growing, ridge filters and chain codes was developed in [11].

According to all of the above, this paper is organized as follows: Section 2 presents the research problem to solve, Section 3 explains the research methodology, Section 4 exposes the main contribution and finally, Section 5 presents the results achieved and their validity.

[4] Source Argano group `http://www.arganogroup.com`.
[5] Source Aibotix GmbH `http://www.aibotix.es`.

## 2 Research problem to solve

Taking into account the importance of line tracking in an automated inspection process, the development of a process of servo-visual control for line tracking is proposed. This process includes the development of a line detection method, since it is required to accomplish the line tracking in order to achieve automatic inspection systems [5]. Additionally, we have not seen many works that approach the line tracking implementation. We also consider that the geometric scene understanding can be very useful for this problem, since it can allow the UAV to make decisions for navigating in complex environments.

The line detection techniques mostly used are: Hough Transform, Line Segment Detector (LSD) [9], Edge Drawing Lines (EDLines) [10] and approaches based on steerable filters and growing regions [14].

The pole detection is an important aspect in the process of inspection, since its detection can help to select the regions where the power lines are located. The graph cut image segmentation combined with priori knowledge and Radon Transform have been used for this task [15].

On the other hand, in the area of visual servoing with UAVs, the works of [16], [17], [18], [19], use feature descriptors such as SIFT and SURF in order to find key points for navigation and reconstruction.

In vision systems, the problem of visual odometry is very important. One approach for this is to incrementally retrieve the path pose to pose and to optimize over the last $n$ poses. In addition to this, by means of visual simultaneous localization and mapping (VSLAM), it is possible to achieve an estimation of the robot trajectory [20]. This implies the maintenance of an environment map and the detection of a robot when it returns to a previous visited area.

Based on the above related work, the main goal of this research is to develop an UAV vision based navigation for power line detection and tracking and a 3D reconstruction of the surroundings; specifically:

- *To develop an algorithm for power line detection.* A comparative analysis of the existent method such as Hough transform, LSD and EDLines is proposed. A linking of line segments stage is considered.
- *To develop a visual servoing system for line tracking from an UAV.* In this process the use of visual servoing is considered. This allows the UAV to locate itself in order to navigate along the power line. Additionally, the use of geometric scene understanding (GSU) is proposed in order to generate control actions during the navigation.
- *To develop a VSLAM for navigation in power line surrounding.* We consider that the three-dimensional reconstruction of the environment of the power line can be useful for the inspection and maintenance. For this, we plan to use features such as SIFT and SURF. For the correspondence stage, RANSAC and ICP algorithms will be used in order to find the correspondences and compute the geometric transformation between images and GPU processing techniques for rendering.

## 3 Research Methodology

The methodological process is composed of three stages: in the first one, a relevance analysis of line detection methods is considered and the creation of a new line detection method is proposed; the second one, is a validation of real time detection in a flight simulator, and, in the third one, the evaluation in a real platform is contemplated.

Due to the fact that many UAV navigation tests will be performed, it is necessary to develop virtual environments for simulating these tests under different conditions. This kind of simulations comprise dynamical models of the UAV, as shown in [21]. These have also been used for SLAM[6], as shown in [22, 20]. This provides different possibilities for the start of operations avoiding damages to the device, as well as the incorporation of other flight platforms and different scenarios.

There are some platforms for UAV simulation such as USARsim[7]; another option is to use ROS (Robot Operating System)[8] as shown in [21]. For the development of this project, we decide to use a gazebo simulator called Tumsimulator, that was developed at the Technical University of Munich[9].

The activities to perform as a part of the research methodology are divided into three specific goals explained in the following subsections.

### 3.1 Power Line detection

- To acquire a dataset of power lines; see Figure 2.
- To implement state of the art line detection algorithms in CPU.
  - Hough.
  - LSD.
  - EDLines.
- To create a new line detection method.
- To validate this method with real images of power lines.
- To analyse the paralelizable stages of this algorithm.
- To implement the new line detection method in CUDA [10].
- To create a virtual environment: It is a scenery with different configurations of power lines. A set of towers and poles will be created by using a 3D modeling software; see Figure 3.
- To validate the line detection method with synthetic images taken from the virtual environment.

[6] Simultaneous localization and mapping
[7] `usarsim.sourceforge.net`
[8] `http://www.ros.org`
[9] `http://wiki.ros.org/tum_simulator`
[10] CUDA is a parallel computing platform and programming model invented by NVIDIA

**Fig. 2.** Some images of the acquired dataset

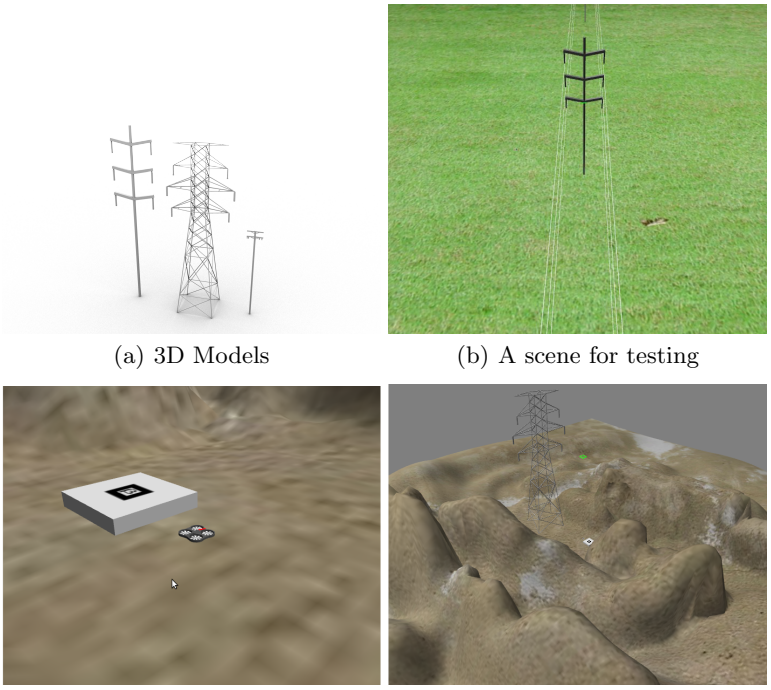
### 3.2 Power Line tracking

- To create a set of videos of power lines.
- To adjust the parameters in order to detect lines in real time.
- To implement the line detection method in a real time flight simulator for UAVs that continuously renders and processes images of scenes.
- To create a technique that allows a virtual UAV of the simulator to track the power line by using the detected lines. This can be a visual servoing algorithm.
- To validate the technique in the virtual environment.
- To validate the technique in a real platform.

### 3.3 VSLAM for navigation in power line surroundings

- To implement a set of 2D feature descriptors.
- To implement a set of algorithms for feature correspondence.
- To evaluate the performance of the selected feature descriptors with the different correspondence algorithms for images of power lines.
- To compute the homography matrix and perform 3D reconstruction of different scenes.
- To implement a VSLAM process in a real platform. In first instance we plan to use an AR-DRone 2.0 platform[11], after that a bigger UAV such as Pelican[12], will be used.
- To evaluate the performance of the system under different environmental conditions.

[11] http://ardrone2.parrot.com/
[12] http://www.asctec.de/uav-applications/research/products/asctec-pelican/



(a) 3D Models (b) A scene for testing

(c) UAV simulated (d) Virtual environment

**Fig. 3.** Development of the virtual environments for simulation.

## 4 Main contribution

The main contribution of this work, is the development of a visual based navigation model for UAVs with application in power line tracking for inspection and 3D reconstruction of the surroundings.

For this reason, it is necessary to develop a vision system composed of computer vision techniques for detecting and tracking objects of interest, such as lines and towers, in order to validate the model.

One derivative contribution of this work is the creation of a new line detection method.

The algorithm is composed of a valid point detector. This is a method that locates points that belong to a line in a segmented image; see Figure 4. This algorithm is based in computer graphic primitives and location of symmetric points in lines. It requires the use of a circle drawing.

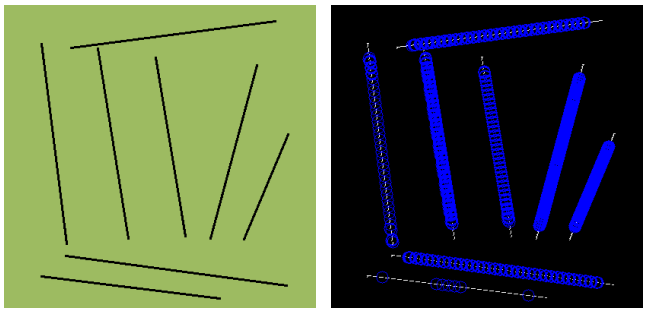
The process comprises the following stages:

- To segment the image with an edge detector, the Canny or a Steerable filter may be used.
- For all pixels at the image that are different than background:
  1. To search valid points by using a circle drawing algorithm.

2. If is a valid point obtain a value of $Dx$ and $Dy$ (see Figure 4.
3. Move towards the symmetry direction by using the values of $(Dx, Dy)$.
4. While (Valid Point)
   - Move the position $(x, y)$ to the values of $(x + skip \cdot Dx, y + skip \cdot Dy)$.
   - To search valid points by using a circle drawing algorithm.
5. Save first and final points.
6. Trace a line between the first and final point, in order to erase the pixels associated to the line in the segmented image.

The main parameters of this method are the following:

– Search radius: this is the circle radius which depends on the length of the line and the size of the image. It has to be larger than 3 pixels in order to capture more details.
– Percent of points detected: when a valid point is selected, it is necessary to take into count how many points are found in a straight line between its two end points. The points are obtained by using primitives for line tracing such as DDA or Bresehnham [23].
– Skip value: This is an increment of skip from a valid point to another. It has a way to control the velocity of the process and its accuracy. This value can be close to the circle diameter.



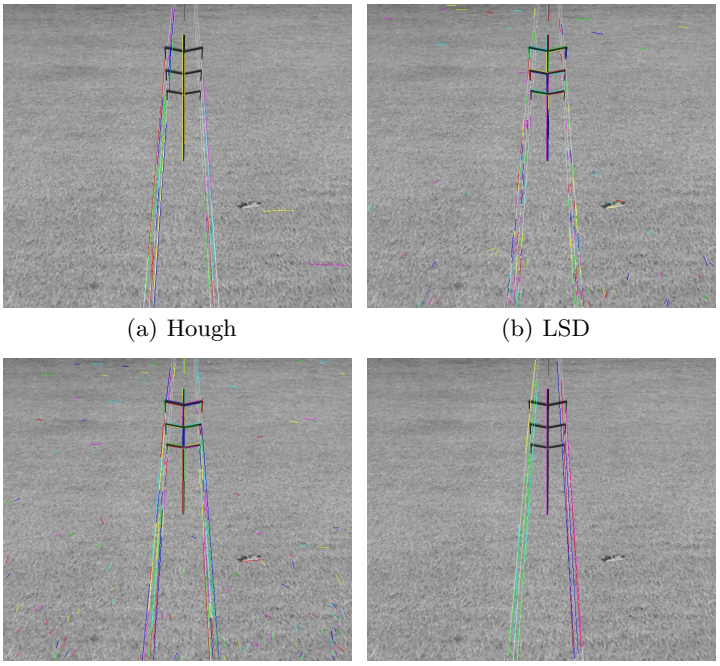
(a) Original (b) Valid points

**Fig. 4.** Obtainment of valid points.

## 5 Results achieved and their validity

A new method for line detection based in the computer graphics approach for drawing circles was developed. This method allows to obtain key points of lines by using a symmetry analysis in a circle based search.

This method is validated with synthetic images (see Figure 5) and real images (see Figure 6), obtaining satisfactory results.



(a) Hough (b) LSD

(c) EDLines (d) Proposed

**Fig. 5.** Comparison between different methods for the image in the Figure 3(b).

One advantage of this method, is the use of integer arithmetic in most of the process. The time of response of the line detection is enough for real time processing 10 to 60 milliseconds, approximately, in images of 800 × 600 pixels.

A three dimensional model of the power lines was built in order to generate synthetic images of power lines with different points of view; see Figure 3.

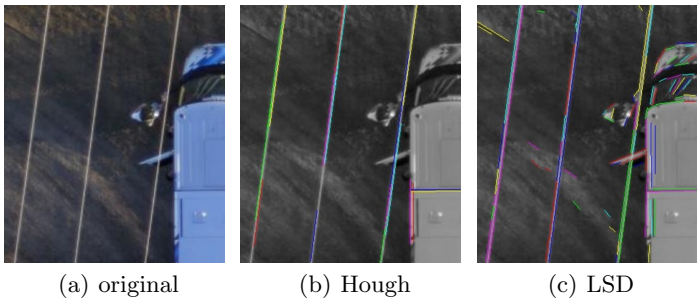
It is important to mention that by using virtual environments, it is possible to create different configurations of the scene for validating computer vision techniques.

Finally, an integration of the virtual environment of simulation with the computer vision techniques is obtained. In Figure 7, the process of segmenting and line detection is shown.

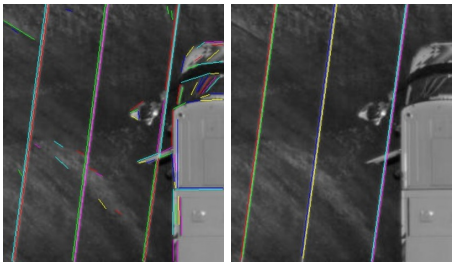
## References

1. Li, Z., Walker, R., Hayward, R., Mejias, L.: Advances in Vegetation Management for Power Line Corridor Monitoring Using Aerial Remote Sensing Techniques. In: Proceedings of the First International Conference on Applied Robotics for the Power Industry (CARPI), Ieee (2010) 1–6
2. Heer, P.: Framework for Vision-Based Power Line Inspection with an UAV. Technical report (2012)

3. Lu, M., Sheng, G., Liu, Y., Jiang, X., Nie, S., Qu, G.: Inspection Based on Unmanned Aerial Vehicle. In: Power and Energy Engineering Conference (APPEEC). (2012)
4. Wang, B., Chen, X., Wang, Q., Liu, L., Zhang, H., Li, B.: Power line inspection with a flying robot. In: 2010 1st International Conference on Applied Robotics for the Power Industry (CARPI 2010), Ieee (2010) 1–6
5. Li, Z., Bruggemann, T.S., Ford, J.J., Mejias, L., Liu, Y.: Toward Automated Power Line Corridor Monitoring Using Advanced Aircraft Control and Multisource Feature Fusion. Journal of Field Robotics **29** (2012) 4–24
6. Hough, P.: Method and means for recognizing complex patterns. US Patent: 3,069,654. (1962)
7. Duda, R.O., Hart, P.E., Park, M.: Use of the Hough Transformation To Detect Lines and Curves in Pictures. Graphics and Image Processing **15** (1972) 11–15
8. Burns, B., Hanson, A.R., Riseman, E.M.: Extracting Straight Lines. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-8** (1986) 425–455
9. Grompone von Gioi, R., Jakubowicz, J., Morel, J.M., Randall, G.: LSD: a Line Segment Detector. Image Processing On Line (2012)
10. Akinlar, C., Topal, C.: EDLines: A real-time line segment detector with a false detection control. Pattern Recognition Letters **32** (2011) 1633–1642
11. Yao, X., Guo, L., Zhao, T.: Power Line Detection Based on Region Growing and Ridge-Based Line Detector. In Sun, Z., Deng, Z., eds.: Chinese Intelligent Automation Conference. Volume 255 of Lecture Notes in Electrical Engineering., Berlin, Heidelberg, Springer Berlin Heidelberg (2013) 431–437
12. Li, Z., Liu, Y., Hayward, R., Zhang, J., Cai, J.: Knowledge-based power line detection for UAV surveillance and inspection systems. 2008 23rd International Conference Image and Vision Computing New Zealand (2008) 1–6
13. Zhang, J., Liu, L., Wang, B., Chen, X., Zheng, Q.W.: High speed Automatic Power Line Detection and Tracking for a UAV-Based Inspection. In: International Conference on Industrial Control and Electronics Engineering. (2012)
14. Liu, Y., Mejias, L., Li, Z.: Fast power line detection and localization using steerable filter for active uav guidance. In: In 12th International Society for Photogrammetry & Remote Sensing (ISPRS2012). Volume XXXIX. (2012) 491–496
15. Cheng, W., Song, Z.: Power Pole Detection Based on Graph Cut. 2008 Congress on Image and Signal Processing (2008) 720–724
16. Campoy, P., Correa, J.F., Mondrag, I., Mart, C., Olivares, M., Mejias, L., Artieda, J.: Computer Vision onboard UAVs for civilian tasks. Ingellingent and Robotics Systems (2009)
17. Artieda, J., Campoy, P., Correa, J.F., Mart, C., Olivares, M.: Visual 3-D SLAM from UAVs. Journal of Intelligent & Robotic Systems **55** (2009)
18. Mondragón, I.F.: On-board visual control algorithms for Unmanned Aerial Vehicles. PhD thesis, Universidad Politecnica de Madrid (2011)
19. Weiss, S.M.: Vision Based Navigation for Micro Helicopters (PhD Thesis - Weiss 2012). PhD thesis, ETH (2012)
20. Dijkshoorn, N.: Simultaneous localization and mapping with the AR.Drone. (2012)
21. Meyer, J., Sendobry, A., Kohlbrecher, S., Klingauf, U., Stryk, O.V.: Comprehensive Simulation of Quadrotor UAVs Using ROS and Gazebo. In: Simulation, Modeling, and Programming for Autonomous Robots. (2012) 400–411
22. Dijkshoorn, N.: Integrating Sensor and Motion Models to Localize an Autonomous AR.Drone. International Journal of Micro Air Vehicles **3** (2011)
23. Hearn, D., Baker, M.P.: Computer Graphics. Prentice Hall (1996)

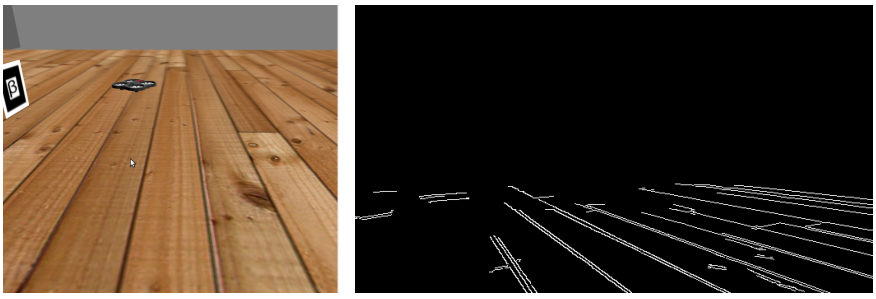


(a) original (b) Hough (c) LSD

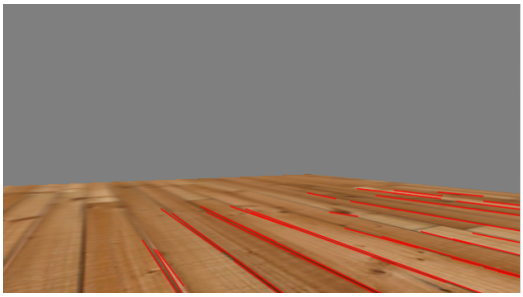


(d) EDLines (e) Proposed

**Fig. 6.** Result for detecting lines in real images segmented using canny and different methods for detection.



(a) Virtual environment (b) Segmentation in real time



(c) Line detection in real time

**Fig. 7.** Results of integration of simulator with computer vision techniques.

# Gait Analysis of the Inferior Articulations of Healthy People and with Locomotion Problems

Sergio Rodolfo Cruz Gómez and Manuel Martín Ortiz

Autonomous University of Puebla
Faculty of Computer Science
14 Sur y Av. San Claudio, C.P. 72570 Puebla, Mexico
sergio.cruz.g@hotmail.com, mmartin@cs.buap.mx

**Abstract.** Gait analysis is the study of the animal locomotion, more precisely for this paper, the human locomotion. Several studies have been made by many researchers, however, there is limited information about the unique characteristics of the direction change in the gait. In this paper we present a different approach to get the information we need to do this analysis, by being non invasive, because most of the ways researchers do have complicated procedures and often require the patience to be static for long periods of time. The objective of this work is to give an alternative non invasive help to medics and physiotherapists when evaluating a patience state.

**Key words:** Kinect, computer vision, gait analysis.

## 1 Introduction

The knowledge of the human locomotion is the base of the systematic treatment and of the manage of the pathological gait, specially when prosthesis and orthosis are used. The normal human locomotion is being described as a series of alternating and rhythmic movements of the extremities and the chest which determine a movement forwards from the center of gravity. More specifically, the normal human locomotion can be described enumerating some of it's characteristics. Even thou there are small differences in the gait from an individual to another, these differences are between small limits. The gait's cycle starts when the foot touches the ground and ends when with the next touch of the ground from the same foot. The two greatest components of the gait's cycle are: the stand phase and the swing phase. A leg is in stand phase when it's in touch with the ground and it's in the swing phase when it does not touch the ground. The length of the full step is the lineal distance of the sequential points of contact of the heel of the same foot. [1]

## 2 Gait's cycle

In it's study, the gait's cycle starts when a foot makes contact with the ground and ends with the same contact of the same foot, to the distance between those points we are going to call a full step. The gaits cycle is divided in two different components: the stand phase, and the swing phase. These two phases are alternating from one leg to another during the gait. In a full step, the simple stand refers to the period when there is only one leg touching the ground. The period of double stand occurs when both feet are in touch with the ground simultaneously. The difference between walking and running is the absence of a period of double stand. For this project is important to know the time that each of the gait's phases take, this way we can have a general idea to corroborate with the prototype. The relative quantity of time spent during each phase of the gait's cycle is as follows [1]:

- Stand Phase: 60% of the cycle
- Swing Phase: 40% of the cycle

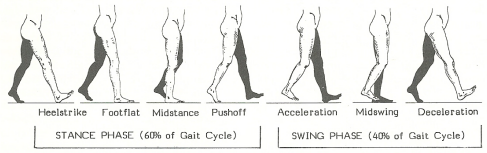



**Fig. 1.** Gait Phases

## 3 Cinematic analysis of the human gait on the sagittal plane

For this project is necessary to know what happens in each of the segments of the leg when walking. The analysis is divided in three intervals, in which are described how the ankle, knee and hip act in the sagittal plane for each one of the phases.

### 3.1 Interval I

Movement of the articulations in the sagittal plane between the contact of the ankle with the ground and the stand point. [1]

- The ankle

—

| | |
|---|---|
| Moment with the contact of the ankle with the ground | The articulation of the ankle is in neutral position (0 degrees) |
| Simultaneously in contact with the ground | The articulation of the ankle starts to move in direction of the plantar flexion |
| Moment in which the foot's sole makes contact with the ground | The ankle's articulation moves 15% from the neutral position to the plantar flexion. |
| In the middle phase | The ankles articulation goes fast to approximately 5 degrees |

– The knee

—

| | |
|---|---|
| Immediately after the contact of the ankle with the ground | The articulation of the knee is in complete extension |
| Simultaneously the ankle in contact with the ground | The articulation of the knee starts to flex and does so until the foot's sole completely touches the ground |
| Immediately after reaching flat position of the foot | The knee has approximately a 20 degrees flexion and starts to stretch |
| In the middle stand | The knee articulation has approximately a 20 degrees angle of flexion and continues stretching |

– The hip

—

| | |
|---|---|
| Simultaneously the ankle in contact with the ground | The hip is approximately in 30 degrees of flexion |
| Immediately after the contact of the ankle with the ground | The articulation of the hip starts to stretch |
| Flat position of the foot | The angle of flexion decreases around 20 degrees |
| Between flat position and middle stand | The articulation of the hip goes to it's neutral position (0 degrees) |

### 3.2 Interval II

Movement of the articulations in the sagittal plane between the middle stand and the takeoff of the foot from the ground. [1]

– The ankle

—

| | |
|---|---|
| In the middle stand | The ankles articulation goes fast to approximately 5 degrees |
| The moment the ankle takes off the ground | The ankle's articulation is approximately 15 degrees |
| In the interval of the ankles elevation and the foot's takeoff | The ankle's angle moves fast to 35 degrees, and when taking off the foot the articulation is approximately at 20 degrees from the plantar flexion |

– The knee

—

| | |
|---|---|
| In the middle stand | The knee has approximately a 10 degrees flexion and continues stretching |
| immediately after the ankle looses touch with the ground | The knee is at 4 degrees from full extension |
| Between ankle's takeoff and toes takeoff | The articulation of the knee moves from almost full extension to a 40 degrees flexion |

– The hip

—

| | |
|---|---|
| In the middle stand | The articulation of the hip is at neutral position (0 degrees) and starts to stretch |
| Immediately after the ankle looses touch with the ground | The hip reaches the maximum stretch of 20 degrees |
| The moment the toes takeoff | The hip is almost at neutral position and moves towards flexion |

### 3.3 Interval III

Describes the movement of the articulations in the sagittal plane in the swing phase. [1]

- The ankle

| | |
|---|---|
| During swing phase | The foot moves from it's initial plantar flexion to an essentially neutral position (0 degrees) and stays like that during all the phase |

- The knee

| | |
|---|---|
| Between the foot's takeoff and the middle part of the swing phase | The knee flexes from an initial position of approximately 40 degrees to a maximum angle of approximately 64 degrees |
| Between the middle part of the swing phase and the touch of the ankle with the ground | The knee stretches almost fully until the last instant of the swing phase. |

- The hip

| | |
|---|---|
| During swing phase | Starting from a neutral position, the hip's articulation flexes approximately 30 degrees and keeps that position |

With this information we can know the correct gait that a healthy person should have. [1]

## 4 Diseases of the lower limbs in walking

To define the diseases in this paper we contacted Doctor Jaime Rebollo Vázquez, from the medical clinic of the Medicine Faculty of the Meritorious Autonomous University of Puebla, so he could advise us. Dr Rebollo did a statistic of the patiences that go to the clinic and gave us a list of the most common diseases:

- Ankle sprain
- Chondromalacia patella
- Gonarthrosis

### 4.1 Ankle sprain

The ankle sprain is the most frequent traumatic lesion of the lower limbs. This lesion changes the posture, which is why the treatment is a important to the normalization of the posture. The ankle sprain is the forced inversion of the ankle. It is a lesion that occurs more frequently between the age of 21 and 30 years old, possibly due to the increase of sport activities in that age. When this disease occurs in younger or older persons, the lesion tends to be worse. Up to 44% of the people that suffered an ankle sprain present some kind of sequel.

Depending of the damage of the lesion we can classify it in three types:

- Grade 1: A stretch is produced, the patient can walk, there is low pain and in general there are a few symptoms. The swelling is minimum and the patient can start a sport activity in two to three weeks.

– Grade 2: There is medium pain with a minor articular instability. There is swelling. The patient walks with a pain position and the symptoms are more evident.
– Grade 3: There is severe pain, deformity and swelling. The patient can't walk nor rest the foot in the ground.

From the three types we are going to be able to study only one, the second one, because the grade 1 does not have enough symptoms and the grade 3 does not allow the patient walk. [14]

### 4.2 Chondromalacia patella

The Chondromalacia patella presents in the body when there is cartilage wear. The cartilage is the best shock absorber we have in our articulations. We have Chondromalacia patella when the cartilage stops being smooth and white, and changes to a wrinkled surface, even being able to disappear, leaving the bone exposed. The Chondromalacia patella presents the following symptoms:

– Patients frequently complain about pain in the back of the knee, in some cases the knee can be stiff and is difficult to stretch.
– The Q angle is usually increased (angle formed by Ankle-knee-Hip).

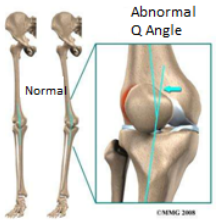



**Fig. 2.** Q Angle

– The ankle is slipped towards the inside.

In this case we will approach the disease by noticing if the knee moves towards the inside of the legs, as seen in figure 2. [15]

### 4.3 Gonarthrosis

The Gonarthrosis is a disease that affects normally the adult population, characterized by the pain, claudication, deformity and functional incapacity. Affects mainly women. The symptoms of the Gonarthrosis are:

– Pain: Is the most frequent symptom, can be located in the back, front and medial part of the knee .

– Stiffness: The duration of this symptom is less than 30 minutes, which makes this disease different from the other swelling diseases.
– Swelling incapacity: The affected articulation presents difficulty to move.

**Fig. 3.** Displacement of the knee towards outside

In the Gonarthrosis, exists an external ligament decomposition, which makes the knee to slip towards the external part of the leg, this movement is the most common of the Gonarthrosis as seen in figure 3. [16]

## 5 Motion Capture (MoCap)

Motion Capture or MoCap is a technique of digitalization recording the movements of the entities, persons or animals. Traditionally the computer animation techniques are used to create movements of the entity. There are three types of techniques of motion capture (MoCap). The first technique is called optical motion capture where the photogrammetry is used to establish the position of an object in the 3D space based in it's observed place in the fields in 2D from a certain number of cameras. [4][5] The second technique is called magnetic motion capture, where the positions and orientations of magnetic sensors are calculated with a transmitter.[8][9] The last technique is called electro mechanic motion capture, and involves motion modeling using a suit with glued sensors. [1][2][3][6][10]

Another form of motion capture is the kinect, which is a combination of motion captures. The kinect sensor is an horizontal bar connected to a small base with a motorized pivot. The device has an RGB camera, depth sensor and a multi array of microphones. The depth sensor is made by the PrimeSense company.[12]

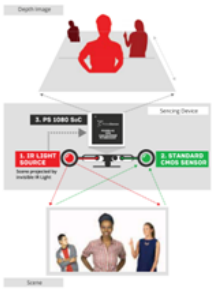


**Fig. 4.** Use of the kinect's depth

The technology that PrimeSense uses gives the ability to digital devices to observe a scene in three dimensions. Translates it's observations in a synchronized stream of images to information, such as people identifier, it's body properties, movement and gestures, and objects classifier, as seen in figure 4.

This is due to the infrared light, which is invisible to the naked eye. This solution uses a standard images sensor CMOS to read the codified light from the scene. This process allows the depth acquisition and is what makes the PrimeSense solutions so accurate. [13]

This solution can be coupled to the customer needs, because it can change it's preferences, such as range, field vision, depth resolution and frames per second.

### 5.1 Software development

The first part of the software is focused in data capture, so we can define if a person has any disease of the previously mentioned. We are going to use the kinect device because it has many advantages, like no being invasive, it does not need specialized devices as sensors and external transmitters and it is also a very known device, which gives us access to the device support and wide information.

When it comes to the articulations needed for this project we only need the ankle, toe tip, knee and hip, because the diseases that we are focusing into are concentrated in these articulations, and their symptoms are presented in the same articulations.

**Fig. 5.** Human skeleton given by the kinect device

We capture the articulations using the kinect, which can give us a full body point map as seen in figure 5 . Now the following question is if the kinect device can be used to capture the human body accurately. Given that it is an non invasive system it does have errors when it comes to accuracy. The error changes depending on how far the object is.

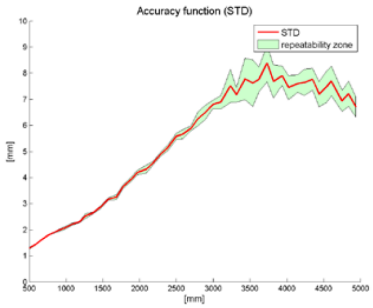



**Fig. 6.** Kinect error

As we can see in figure 6 the error increases depending on the depth of the object. Due to the error we can not go too far away from the device, the distance where the error is low enough would be from 1 meter to 3 meters. [17]

Now, the next step is to evaluate the data in order to say if a person is healthy or not, we will approach this matter by creating a classifier with the recollected data, we will teach our classifier the healthy states of the person's gait and then the non healthy ones.

The data we will classify is the angles that the joints of the leg form, given that the diseases manifest in specific angles we don't have to address all the joint's angles, instead we will focus on the angles needed to evaluate the diseases. For example the angle formed by the hip,knee and ankle joints, a healthy person would have a 180 angle, since the joints are aligned, having a disease like in the figure 2 . This way we can tell if a person has this disease.

The classifier we want to use is clustering due to the fact that is versatile, this classifier needs numerical data, and since the data we analyze is the joint's angle then it fits perfectly, and depending on the classifier's parameters we can adjust it's accuracy. [18]

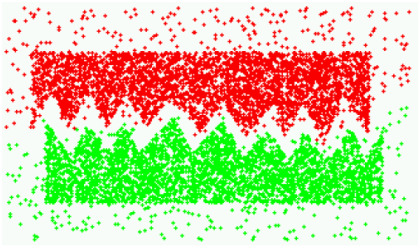


**Fig. 7.** Clusters example

We can see the adaptability of the classifier in figure 7 , which is one of the reasons we chose the clustering classifier.

### 5.2 System limitation

This system has some limitations that are very known in the computer vision filed, and it is occlusion, which is when an object passes in front of another from the camera or sensor. This makes the software to not know what to do and throw incorrect data. Another limitation is the accuracy, because it can be less precise if the device is far from the object, but if it is close enough from the object it can be almost as accurate as the other motion capture methods.

## 6 Expected contributions

We want to provide a non invasive classifier that will help medics evaluate if a person has a disease in it's gait. We want to present a software that will classify accurately the angles of a person's gait and return a accurate enough diagnosis.

## References

1. Biomecánica de la marcha, `http://www.slideshare.net/maxifox/biomecnica-de-la-marcha`
2. D. J. Rose: Equilibrio y movilidad con personas mayores, First Editorial, Paidotribo
3. R. Budiman, M. Bennamoun, and D.Q. Huynh: Cost Motion Capture. The University of Western Australia, School of Computer Science and Software Engineering
4. L. Ballan and G. Maria Cortelazzo: Marker-less motion capture of skinned models in a four camera set-up using optical ?ow and silhouettes. in 3DPVT08, Georgia Institute of Technology, Atlanta, GA, USA, 2008, p. 8.
5. N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall, and H.-P. Seidel: Markerless Motion Capture with Unsynchronized Moving Cameras. TNT, Hannover University, BIWI, ETH Zurich.

6. T.-C. Feng,, P. Gunawardane, J. Davis , and B. Jiang: Motion Capture Data Retrieval Using an Artists Doll. University of California, California, Santa Cruz.
7. J. Shotton, A. Fitzgibbon, and M. Cook: Real-Time Human Pose Recognition in Parts from Single Depth Images. Microsoft Research Cambridge and Xbox Incubation, p. 8.
8. D. Roetenberg, H. Luinge, and P. Slycke: Xsens MVN: Full 6DOF Human Motion Tracking Using Miniature Inertial Sensors. XSENS TECHNOLOGIES , p. 9, Apr. 2013.
9. Mocap Data, http://www.mocapdata.com/
10. Impulse X2 Motion Capture System, http://www.phasespace.com/impulse_motion_capture.html
11. Vicon MX Hardware, System Reference Revision 1.4, Vicon MX System, 20042006, http://www.vicon.com/products/viconmx.html
12. Asus, Xtion Pro, http://www.asus.com/Multimedia/Xtion_PRO/
13. PrimeSense, PrimeSenseTM3D Sensors, PrimeSense Natural Selection, http://www.primesense.com/
14. Klgo. Herrera, Agustn Luis Angel: Esguince de Tobillo. Enfoque kinsico-osteoptico de evaluacin y tratamiento. Magazine Kinsico Number 10, November-December 1999
15. Javier Avendao , Jos Ramn Malvar: CONDROPATIA ROTULIANA. Medical Services, Spain Real Federation of Athletics
16. Rogelio Ramos Naranjo: Caractersticas clnicas de la gonartrosis primaria tratada con limpieza articular por artroscopa. Medicine Faculty, University of Colima, Colima, Col., Septiembre 2008
17. B. Molnr, C. K. Toth, A. Detrekoi: ACCURACY TEST OF MICROSOFT KINECT FOR HUMAN MORPHOLOGIC MEASUREMENTS, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXIX-B3, 2012 XXII ISPRS Congress, 25 August 01 September 2012, Melbourne, Australia
18. George Karypis, Eui-Hong (Sam) Han, Vipin Kumar: CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling, Department of Computer Science and Engineering, University of Minnesota, IEEE Computer: Special Issue on Data Analysis and Mining, Technical Report No. 99-007

# Automatic Linguistic Pattern Identification Based on Graph Text Representation

Helena Gómez-Adorno[1], Grigori Sidorov[1], David Pinto[2],and Darnes Vilariño[2]

[1] Centro de Investigación en Computación,
Instituto Politécnico Nacional, Mexico
helena.adorno@gmail.com, sidorov@cic.ipn.mx
[2] Facultad de Ciencias de la Computación,
Benemérita Universidad Autonóma de Puebla, Mexico
{dpinto,darnes}@cs.buap.mx

**Abstract.** In this paper it is presented a model of text representation based on graphs. The model is applied in the particular case study of authorship attribution. The experiments were performed by using a corpus made up of 500 documents written by 10 different authors (50 documents per author). The obtained results highlight the benefit of using text features at different levels of language description in tasks associated to automatic processing of information. In particular, we have obtained a performance of 57% of accuracy for the authorship attribution task.

**Key words:** Graph representation, writing style patterns, authorship attribution.

## 1 Introduction

There is a huge amount of digital information produced daily in the form of natural language written texts, such as, magazines, books, web pages, newspapers, reports, etc. The exponential growth of these documents requires the development of new approaches in order to explore, analyse and discover knowledge from them. The development of new tools for discovering new, previously unknown information is one of the goals of the Text Mining (TM) research field. It is a non-trivial task to find accurate knowledge in text documents that are helpful to tackle a determined task. Some TM task are text categorization, text clustering, sentiment analysis, document summarization, authorship attribution, information retrieval, tagging/annotation, among others.

The representation of natural language text is needed in order to represent human knowledge in computers. Conventional text representation models observed in TM task are: Boolean models [1], probabilistic models [2] and vector space models [3]. The features used for text representation usually are Bag-Of-Words (BOW) [4] or $n$-grams models [5–7]. The majority of these text representations are based on the BoW representation, thus ignoring the words' sequentiality and, hence, the meaning implied or expressed in the documents as well. This deficiency generally results in failure to perceive contextual similarity

of text passages. This may be due to the variation of words that the passages contain. Another chance is perceiving contextually dissimilar text passages as being similar, because of the resemblance of their words.

For many problems in natural language processing, a graph structure is an intuitive, natural and direct way to represent the data. This paper presents an innovating graph-based framework for automatic analysis of texts, which was proposed in our previous work [8] for the specific task of document understanding. The new model will include linguistic features of different levels of language description, which provide important information about the knowledge that is expressed in a natural language written text.

The aim of the present research work is the construction of a suitable text representation model based on graphs, that can facilitate discovering of important text patterns from it. We show that the features (text patterns) in this manner discovered can be used in various tasks associated to document understanding (such as for document classification, information retrieval, information filtering, information extraction, authorship attribution).

The text pattern discovering technique proposed here is based on the traversal of the graph representation of documents, using the shortest paths. This text pattern discovery is used in our experimental case study for estimating similarities between pairs of texts. The case study of authorship attribution presented here demonstrates how our framework works and its' efficacy. The results of the experimental work reported here are analyzed, and current findings are discussed.

There exist several research works that have employed graphs for representing text. A comprehensive study of the use of graph-based algorithms for natural language processing and information retrieval can be found in [9]. It describes approaches and algorithmic formulations for: (a) synonym detection and automatic construction of semantic classes using measures of graph connectivity on graphs built from either raw text or user-contributed resources; (b) measures of semantic distance on semantic networks, including simple path-length algorithms and more complex random-walk methods; (c) textual entailment using graph-matching algorithms on syntactic or semantic graphs; (d) word-sense disambiguation and name disambiguation, including random-walk algorithms and semi-supervised methods using label propagation on graphs; and (e) sentiment classification using semi-supervised graph-based learning or prior subjectivity detection with min-cut/max-flow algorithms. Although the work described in [9] covers a wide range of algorithms and applications, there exist other relevant works in literature worth mentioning. A great interest has grown in the computational linguistic community for using this kind of text representation in diverse tasks of natural language processing, such as in summarization [10], coreference resolution [11], word sense disambiguation [12–14], word clustering [15, 16], document clustering [17], etc.

The majority of the approaches presented in literature use well known graph-based techniques in order to find and exploit the structural properties of the graph underlying a particular dataset. Because the graph is analysed as a whole,

these techniques have the remarkable property of being able to find globally optimal solutions, given the relations between entities. For instance, graph-based methods are particularly suited for disambiguating word sequences, and they manage to exploit the interrelations among the senses in the given context. Unfortunately, most of the research works that use graph-based representations propose ad-hoc graph-structures that only work with the particular problem they are dealing with. It is, therefore, imperative to attempt to propose a general framework that may be used in different contexts with a minimum amount of changes.

In summary, this research work presents a new text representation schema useful for mining documents, exploiting their lexical, morpho-syntactic and semantic information. The representation schema is built over a syntactic analysis developed through a dependency parser for all the sentences in the document, including further morphologic and semantic information. The final result obtained is an enriched output in the form of a graph that represents the input document in the form of a multiple level formal representation of natural language sentences. The graph-based representation schema and the similarity measure proposed here, enables a more effective and efficient text mining process.

## 2 Research Methodology

The proposed text representation schema utilizes multiple linguistic levels of formal definition of natural language texts. The motivation for the schema is to capture most of the features present in a document, ranging from lexical to semantic level. By including lexical, syntactic, morphologic and semantic analysis in the representation, we attempt to represent how different text components (words, phrases, clauses, sentences, etc.) are related.

The representation of each linguistic level together with their association with the graph components is described as follows.

**Lexical level** At the lexical level we deal with words, one of the most basic units of text, describing their meaning in relation to the physical world or to abstract concepts, without reference to any sentence in which they may occur. Lexical definition attempts to capture everything that a term is used to refer to and, as such, is often too vague for many purposes.

**Morphological level** At the morphological level we deal with the identification, analysis and description of the structure of a given language's morphemes and other linguistic units, such as root words, affixes and Parts of Speech (PoS).

**Syntactical level** At the syntactical level we deal with rules and principles that govern the sentence structures. Usually, the lexical parser (or simply: the parser) can read various forms of plain text input and can output various analysis formats, including part-of-speech tagged text (morphological level), phrase structure trees, and a grammatical relations (typed dependency) format.

**Semantic level** At the semantic level we deal with the meaning of sentence, i.e., human expression stated through language. In general, semantic level refers to interpretation of signs or symbols used in agents or communities within particular circumstances and contexts. In written language, things like paragraphs, words and punctuation symbols bear semantic content. The most popular semantic relationships are: antonym, synonym, class inclusion, part-whole, and case [18].

### 2.1 Formalization of the Graph-Based Multi-Level Linguistic Representation

Given a text $T = \{t_1, t_2, \cdots, t_{|T|}\}$ with $t_i$ a word in the document. Let $PoS(t_i)$ be the PoS tag of $t_i$, $Lem(t_i)$ be the lemma of $t_i$, $Sem(t_i)$ be a term semantically related with $t_i$, and $Dep(t_i, t_k)$ be the dependency tag obtained by some syntactical parser over the sequence "$t_i t_k$". The graph-based multi-level linguistic representation of $T$ can be formally expressed by a di-graph $G = \{V, E, L_V, L_E, \alpha, \beta\}$, with:

- $V = \{v_i | i = 1, ..., n\}$ is a finite set of vertices, $V \neq \emptyset$, and $n$ is the number of vertices in the graph.
- $E = \{(v_i, v_j) | v_i, v_j \in V, 1 \leq i, j \leq n\}$. Note that the notation $(v_i, v_j)$ indicates that a given order is established.
- $L_V = \{\bigcup_{i=1,\cdots,|T|}(Lem(t_i) \bigcup Pos(t_i))\}$
- $L_E = \{\bigcup_{i,j=1,\cdots,|V|} Dep(v_i, v_j)$ with $v_i, v_j \in V,$ and $(v_i, v_j) \in E\}$
- $\alpha : V \rightarrow L_V$
- $\beta : E \rightarrow L_E$

Here, we say that $L_E$ represents the dependency tag between a pair of words. However, it is more practical to have a numeric value as edge label in addition to the dependency tag. We, therefore, extend the graph-based representation using the following definition of $L_E$.

$L_E = \{\forall_{i,j=1,\cdots,|V|}(Dep(v_i, v_j) : frec(Dep(v_i, v_j)) + frec((v_i, v_j)))\}$
with $v_i, v_j \in V$, and $(v_i, v_j) \in E$

Where $frec(x)$ is a function that counts the occurrences of $x$ in the entire graph.

Figure 1 shows the final constructed graph for an example sentence. To illustrate the graph representation, let us consider the following example: *Text mining searches patterns in texts.*

### 2.2 Pattern Extraction Technique for Discovering Text Features

The graph may represent one sentence, one paragraph, one document, or even a collection of documents. We assume that the graph uses the representations we discussed in the previous section. The technique proposes to find features in the graph by counting text components (word lemmas, PoS tags, grammatical tags) when different paths are traversed. These components would seem to be

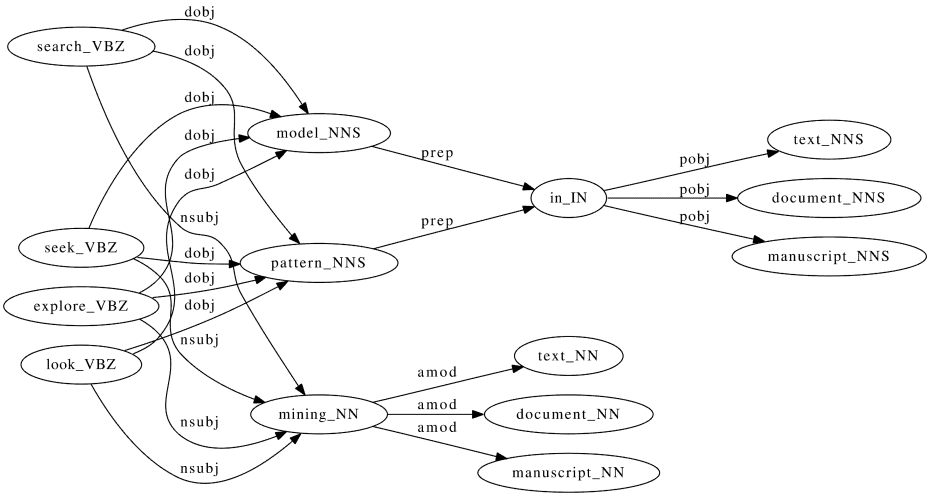



**Fig. 1.** Semantical representation of texts using word lemmas, PoS tags, dependency tags and word synonyms

isolated elements of the graph, however, counted over a path of interest they are considered to be textual patterns.

Let us consider the semantic representation shown in Figure 1, the minimum path from the node $search\_VBZ$ to the node $text\_NNS$ will have the following features at different language description levels:

- Lexical level: *search*, *model*, *text*, *in*.
- Morphological level: $VBZ$, $NNS$, $IN$, $NNS$.
- Syntactical level: *dobj*, *prep*, *pobj*.

Those features may be further used (perhaps as a bag of words or a vector space model based vector) for some particular task to be carried out. Thus, a textual document represented by a graph may provide a set of features for each of the minimum paths found in that graph. These features can be used for encoding a meta-representation of the text.

## 3 Case Study

In order to analyse the performance of the graph-based multi-level linguistic representation and the pattern extraction technique, we present their application in a particular problem of text mining known as Authorship Attribution.

The authorship attribution is often seen as an issue of computational linguistics and aims to identify the author or the original writer of a given text. To accomplish this task, it is necessary to identify characteristics (features) or profiles that identify the target author. This is not a trivial task because writing styles are often similar. Many different techniques have been develop in order to

solve this problem, from mathematical and statistical applications to computational linguistic approaches.

The particular interest in this research field lies in the huge amount of existing documents and the need to develop automatic and reliable models that contribute to the optimal solution of the problem.

In order to determine the performance of the text representation proposed in this work in a real scenario, we used a corpus gathered by Efstathios Stamatatos(ccat C10) [19] from the RCV1 Reuters collection [20]. This corpus is made up of 500 documents (10 authors each one with 50 different documents) for both, the training and test dataset.

### 3.1 Applying the Proposed Methodology to the Authorship Attribution Task

By using the methodology proposed, we can construct ten graphs that include all the document collection. In this case, each graph will represent only one author (50 documents).

In order to identify the authorship of a given test document, we need first to represent the test document with the proposed graph-based representation 2. Thereafter, we use the pattern extraction technique introduced in Section 2.2 for obtaining numeric vectors and subsequently to calculate the similarity between the test graph and each of the ten author's graphs (which contains the graph-based representation of the fifty documents of each author). Thus, following the features extraction technique, the root node of the test graph is fixed as the initial node whereas the final nodes selected correspond to the rest nodes of the test graph. This leads to diminish the computational time to $O(n)$, with $n$ equal to the number of nodes in the test graph. We have used the Dijkstra algorithm[21] for finding the minimum path between the initial and each final node. Thereafter, we count the occurrences of all the multi-level linguistic features considered in the text representation, such as part-of-speech tags and dependency tags found in the path. The same procedure is performed with the training graph by using the pair of words identified in test graph as initial and final nodes. As a result of this procedure, we obtain two set of feature vectors: one for the test graph, and one for the training graph.

The technique extracts a set of vectorial patterns ($\overrightarrow{f_{t,i}}$) for each text $t$, with $V$ equal to the total number of lexical, morphological and syntactical features. Thus, the training graphs $Tr$ will now be represented by $m$ feature vectors ($Tr^* = \{\overrightarrow{f_{d,1}}, \overrightarrow{f_{d,2}}, \cdots, \overrightarrow{f_{d,m}}\}$), as well as the test graph $Te$ ($Te^* = \{\overrightarrow{f_{h,1}}, \overrightarrow{f_{h,2}}, \cdots, \overrightarrow{f_{h,m}}\}$). Here, $m$ is the number of different paths that may be traversed in both graphs, using the "ROOT-0" vertex as the initial node and each word appearing in the test graph as the final node.

Since each path of the test graph contains exactly the same number and types of components as that of the training graphs, it is possible to calculate the degree of similarity among each path traversed. For the purposes of this case study, we have used the cosine similarity measure, which is calculated as in Eq.(1).

$$
\begin{aligned}
Similarity(h^*, d^*) &= \sum_{i=1}^{m} Cosine(\overrightarrow{f_{h,i}}, \overrightarrow{f_{d,i}}) \\
&= \sum_{i=1}^{m} \frac{\overrightarrow{f_{h,i}} \ . \ \overrightarrow{f_{d,i}}}{||\overrightarrow{f_{h,i}}|| \ . \ ||\overrightarrow{f_{d,i}}||} \\
&= \sum_{i=1}^{m} \frac{\sum_{j=1}^{|V|} (f_{(h,i),j} * f_{(d,i),j})}{\sqrt{\sum_{j=1}^{|V|} (f_{(h,i),j})^2} * \sqrt{\sum_{j=1}^{|V|} (f_{(d,i),j})^2}}
\end{aligned} \tag{1}
$$

After obtaining all the similarity scores for the test graph with respect to the ten training graphs, the training graph obtaining the highest score is selected as the correct author for that test graph.

### 3.2 Obtained Results

In Figure 2 we show the results obtained for each one of the ten authors. We have executed the following four different representation methods:

- *MinText_WithoutLem:* It represents the text in the graphs by using the document words (not lemmatized).
- *MinText_LemFrec:* It includes the word lemmatization plus the frequency count of the nodes in the entire graph.
- *MinText_LemFrecLexMor:* It includes the counting of lexical and morphological features.
- *MinText_LemFrecLexMorVocSuf:* It includes, besides all the previous features, the combination of vowels count, permutation of vowels count and suffix count.

As can be seen, there exist some authors that were benefited of using more complex components of text representation, for example, lemmatization and use of PoS tags. However, there are other authors who did not. We consider this behaviour is due to the coherency of writing style of some authors. However, this is an hypothesis that we need to analyse more into detail by reviewing each one of the documents written by each author. Some authors have a writing style that is much more easy to be discovered and modeled, such as "Alan Crosby". In these cases, the inclusion of new characteristics did not contribute to a better classification accuracy because the simple use of their words is enough for detecting the writing style. For those authors that did not use a consistent writing style, the inclusion of new characteristics allowed to discover their possible textual patterns.

In Table 1, we present a summary of the results obtained by each representation model. In average, the *MinText_LemFrecLexMorVocSuf* model obtained the best performance.

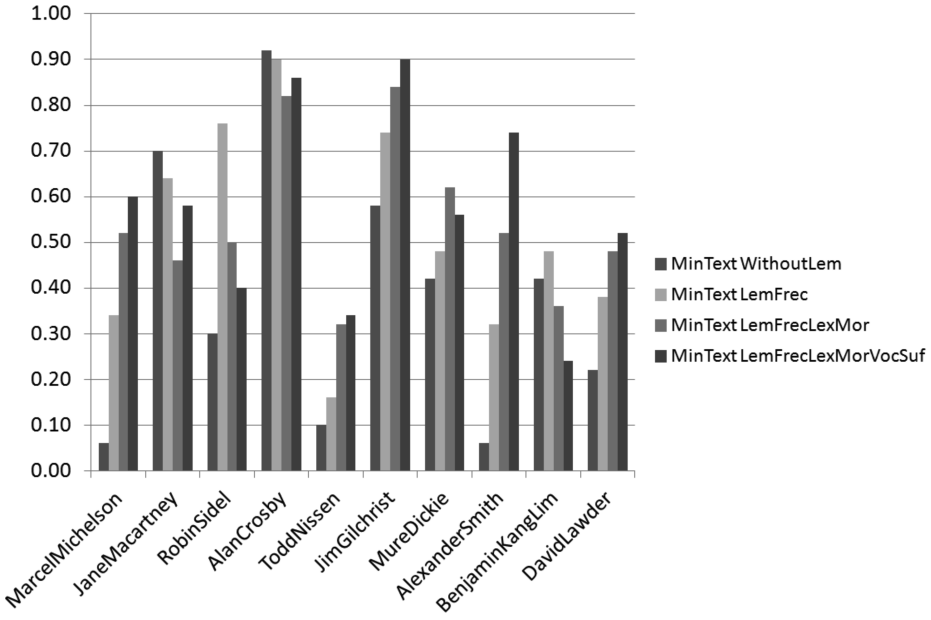



**Fig. 2.** Evaluation of accuracy for each one of the ten authors

**Table 1.** Comparison of the results obtained in the Authorship Attribution task

| **Evaluated approach** | **2011** |
|---|---|
| MinText_WithoutLem | 37.80 |
| MinText_LemFrec | 52.00 |
| MinText_LemFrecLexMor | 54.40 |
| MinText_LemFrecLexMorVocSuf | **57.4** |

## 4 Main Contribution

At the end of this research project we will count with a new methodological framework that can be used in various tasks of MT and PLN. In particular, the new methodology will consist of:

1. A model for text representation based on graph structures, using the different levels of natural language formal description.
2. A set of linguistic pattern extraction techniques useful for solving specific tasks of MT and PLN.
3. Metric for evaluating similarity between patterns discovered on different datasets.
4. Classification models for various PLN and MT tasks: text classification, authorship attribution and profiling, analysis of polarity, among others.

There exist, however, a gap in the process of selecting the best features to be included in the graph that will represent the target text, because, some features

may help or harm the final task to be tackled. Despite this issue, we consider very important to propose a representation that takes into consideration features beyond the traditional ones (such as bag of words).

## References


1. Mauldin, M.L.: Retrieval performance in ferret a conceptual information retrieval system. In: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '91, New York, NY, USA, ACM (1991) 347–355
2. Croft, W.B., Turtle, H.R., Lewis, D.D.: The use of phrases and structured queries in information retrieval. In: Proc. of the 14th SIGIR conference, New York, NY, USA, ACM (1991) 32–45
3. Salton, G., ed.: Automatic text processing. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1988)
4. Mladenic, D., Grobelnik, M.: Word sequences as features in text-learning. In: In Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98). (1998) 145–148
5. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Computer-based authorship attribution without lexical measures. In: Computers and the Humanities. (2001) 193–214
6. Keselj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution (2003)
7. Sidorov, G., ed.: Non-linear construction of n-grams in computational linguistics: syntactic, filtered, and generalized n-grams. Sociedad Mexicana de Inteligencia Artificial (2013)
8. Pinto, D., Gómez-Adorno, H., no, D.V., Singh, V.K.: A graph-based multi-level linguistic representation for document understanding. Pattern Recognition Letters **41**(0) (2014) 93 – 102
9. Mihalcea, R., Radev, D.: Graph-based natural language processing and information retrieval. Cambridge university press (2011)
10. Zha, H.: Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In: SIGIR, ACM (2002) 113–120
11. Nicolae, C., Nicolae, G.: Bestcut: a graph algorithm for coreference resolution. In: Proc. of the EMNLP 2006 Conference, Stroudsburg, PA, USA, Association for Computational Linguistics (2006) 275–283
12. Dorow, B., Widdows, D.: Discovering corpus-specific word senses. In: EACL, The Association for Computer Linguistics (2003) 79–82
13. Veronis, J.: Hyperlex: lexical cartography for information retrieval. Computer Speech & Language **18**(3) (2004) 223–252
14. Agirre, E., Martínez, D., de Lacalle, O.L., Soroa, A.: Two graph-based algorithms for state-of-the-art wsd. In Jurafsky, D., Gaussier, É., eds.: EMNLP, ACL (2006) 585–593
15. Matsuo, Y., Sakaki, T., Uchiyama, K., Ishizuka, M.: Graph-based word clustering using a web search engine. In: Proc. of the EMNLP 2006 Conference, Stroudsburg, PA, USA, Association for Computational Linguistics (2006) 542–550
16. Biemann, C.: Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In: Proc. of the 1st Workshop on Graph Based Methods for Natural Language Processing, Stroudsburg, PA, USA, Association for Computational Linguistics (2006) 73–80

17. Zhong, S.: Generative model-based document clustering: a comparative study. Knowledge and Information Systems **8** (2005) 374–384
18. Bejar, I., Chaffin, R., Embretson, S.: Cognitive and psychometric analysis of analogical problem solving. Recent research in psychology. Springer-Verlag (1991)
19. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In Euzenat, J., Domingue, J., eds.: Artificial Intelligence: Methodology, Systems, and Applications. Volume 4183 of Lecture Notes in Computer Science., Springer Berlin Heidelberg (2006) 77–86
20. Lewis, D.D., Yang, Y., Rose, T.G., Li, F., Dietterich, G., Li, F.: Rcv1: A new benchmark collection for text categorization research. Journal of Machine Learning Research **5** (2004) 361–397
21. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische mathematik **1**(1) (1959) 269–271

# Physically-based Modeling of Virtual Content for a Natural way of Interaction in AR

J. Adrián Leal-Meléndez and Leopoldo Altamirano-Robles

Instituto Nacional de Astrofísica, Óptica y Electrónica,
Computer Science Department
Luis Enrique Erro No. 1, Tonantzintla, Puebla, México
{jalm,robles}@ccc.inaoep.mx

**Abstract.** Nowadays the Augmented Reality (AR) has been spread into several practical areas of applications. Nonetheless, in many cases the interaction between the real and virtual content is in somehow limited. In this regard, most of the work has devoted to the study of AR using physical controls and gesture recognition. In recent years, several works have moved their attention to research more natural ways of interaction. One of the main problems of interest is the physically-based modeling of real objects. This is, export the physical properties of real objects to the virtual objects. Providing a user interface with such physical properties makes possible to improve the user experience. In this proposal, we are interested in modeling three physical properties (weight, center of gravity and material) into the virtual objects and also modeling the environment. Preliminary results, show that using a 3D sensor to generate a point-cloud-based mesh, it is possible to model a suitable environment that improve the spatial-awareness by allowing occlusion handling. Such modeling seems to be an ideal scenario to perform the physically interaction.

**Key words:** Physically-based modeling, natural interaction, augmented reality, 3D sensors.

## 1 Introduction

Augmented Reality (AR) has the potential to bring new user interfaces, in which space is not restricted to a screen and controls, even they could be unnecessary. The classical AR aims to superimpose synthetic information on a view of the real world where both elements, real and virtual, look as a part of the same 3D scene. In this way, a user can receive useful information in real time and in a most adequate place (real environment) in order to be guided in a determined task. In the last years, with the progress of portable technologies such as smartphones, cameras, and sensors for tracking, many AR applications have been developed in several fields. However, the interaction methods provided for such applications are restricted to the use of common input devices: mouse and keyboard. This situation produce a poor user-computer interaction environment.

In this regard, an ideal application of AR should have two main features: 1) allowing the user to interact with the virtual content in a more natural and intuitive way, and 2) the system must provide a kind of feedback mechanism to ensure the spatial-awareness (mixed space) of the user while navigation is performed in the environment. The aforementioned two points are key procedures to improve the immersion and user experience. In this context, different approaches have been used to give more natural ways of interaction with virtual content. There are two main approaches focused on improving the user experience: i) the use of physical objects to interact with the virtual content (tangible user interfaces, TUI), and ii) the use of gesture recognition. The first approach is based on using physical objects as a tool for interacting with virtual objects, this approach was one of the first that removed the dependence of the use of controls along the manipulation phase [11, 12, 15]. The main idea behind this approach, is to manipulate the real object and then reproduce this behavior in the virtual object, thereby when the physical object is manipulated the associated virtual one does too. In this way the user has the feeling of interacting with a real object. Because of the rise of the use of mobile devices, this approach adopted a new form; in this case, the manipulation of virtual objects is based on the use of tactile screens and sensors for tracking and tracing (GPS, accelerometers and gyroscopes). Researches in this approach provide a way of interaction with virtual content through the device, using it as an interface [21, 10, 7, 8, 6]. In spite of providing more intuitive ways for interacting, these works, still remain far from provide an enhanced user experience. The second approach is the one based in gesture recognition, in [13, 3, 2, 14] the authors propose more natural methods for interacting by matching gestures with actions applied to the virtual model. In these works the gestures do not imply contact with the virtual object. On the other hand, in [18, 9, 1] the authors consider the contact with the virtual object and they propose a modeling of that. Finally, the above two main approaches (the use of physical objects and gesture recognition) can be combined into an hybrid approach, in order to consider both of them. This approach is best known as multi-modal interfaces [4].

In order to provide a more natural and intuitive mechanism for interaction in AR, it is necessary going beyond the gesture recognition. In real life, while interacting with real objects there are a number of physical constraints which determine the behavior of such object, these restrictions also determine the user's behavior necessary to manipulate the object. Thus, according to the literature, a way to improve the user experience in AR is modeling the virtual content in such a way it can react to the real world. This is achieved by modeling physical properties into the virtual object, but also modeling the way in which the environment interact with the virtual content. In this way, the user experience is improved by giving to the user the feeling that virtual objects are part of the real world. Furthermore, this condition might allow a natural interaction directly with the user's hands. Having said that, this research is closer to the second approach. The research lies in designing methods to give to the user a

physically-based way of interaction in AR with the purpose of improving the user experience. The main problem addressed in this research is deepen next.

**Problem**: gestures are the most expressive form of communication between humans and computers. Moreover, despite of the approach based on gesture recognition has been used extensively to interact with virtual content in AR, most of these works usually do not consider physical properties to model its behavior. The latter result in a poor manipulation experience due to the lack of feedback; this is there are no feeling of working with real objects. On the other hand, it has been shown that considering physical properties while modeling the virtual content improved the immersion of an AR system, mainly along the interaction phase [16, 5, 18, 17].

**Main objective**: Proposing an strategy for natural interaction with virtual objects in an augmented reality environment that considers physical properties (weight, center of gravity, material) of real objects to model virtual objects.

**Main contribution**: the main contribution in this research is the design of algorithms for modeling real and virtual objects considering physical properties to provide a mechanism for a natural user-experience in AR.

Following the above three points, this research lies on a physically-based interaction in augmented reality; mainly in modeling physical properties to emulate the behavior of a real object when this is pushed by another one and, then apply this model to a virtual object. We are focused in modeling properties of rigid objects directly involved in the interaction task. As a first approach we have considered to model the rigid virtual objects through the use of the finite element method (FEM), The latter is because its capabilities could reproduce real-world behavior [22, 20, 23]. For the rest of the scene we have considered to build a simple mesh from a point-cloud[1] obtained with a 3D sensor. We are interested in modeling only certain specific-target objects, the rest of the scene can be treated as a simple mesh. The rest of this paper is organized as follows: in Section 2 describe the research methodology. In Section 3 we show the work done so far. Finally, in Section 4 an analysis of the obtained results is presented and future avenues of inquiry.

## 2 Research Methodology

The proposed methodology is divided into five main parts, where the main contributions are listed in 1(b), 2 y 3:

1. **Designing an spatial representation of the work area to allow collision detection and occlusion handling**: collision detection and occlusion handling among objects are important features to improve the user experience, while the interaction is performed through the virtual content. To make

[1] Point-cloud: in a three-dimensional coordinate system, is a set of points in which the points are usually defined by $[X, Y, Z]$ coordinates, and often are intended to represent the external surface of an object. A point-cloud may contain additional components as color and direction.

virtual objects react to a collision as a result of an applied force on them, it is needed to model both the virtual object and also the environment. Moreover, the modeling allows detect occlusion relationship, i.e., after reacting to a force, a virtual object could be behind of a real one. The modeling of the environment allow this condition.

(a) **Creating a general model (mesh) in which the virtual object could be attached**: the main idea is to use a point-cloud-based representation using a 3D sensor to create a mesh geometry, which will cover the entire work area. The main problems working with a cloud point obtained from a 3D sensor are: i) there are areas that are not covered by the sensor that produces holes on the mesh, in which the virtual content can not be attached, and ii) there is an overlapping between the mesh and the rgb image.
(b) **Modeling an specific horizontal flat surface, in which the virtual object should slide when a force is applied to it**: once the general model of the work area was created, the next step is to segment an specific flat surface, on which the virtual object will be slipped. We are focused in modeling this surface with two different materials: wood and glass.
(c) **Modeling the obstacles in which the virtual object should collide**: in this point we are interested in select vertical surfaces, which will represent obstacles on the general geometry and model them like a rigid bodies.

This stage include an study of 3D reconstruction and physical modeling approaches as well as point-cloud correction techniques.

2. **Physically-based modeling of virtual objects**: in this stage we are focused in reproduce the behavior of an object, mainly when a perpendicular force is applied on one side. The physical properties considered to model the virtual object are the following: i) weight, ii) center of gravity (contact area which is divided into six zones, see Fig 1), and iii) material (resistance and friction). Initially, we propose to work with three objects with well-defined shapes, which are: cube, pyramid, and sphere. We propose to combine two approaches: i) to use a Finite Element Method to describe the physical constraints and the geometry itself, and ii) to use semantic rules to describe the behavior. Both approaches are studied along this stage and an analysis of the advantages and disadvantages of each approach will be made.
3. **Modeling the tool to interact with the virtual content (user's hand)**: once the work area and the virtual objects have been modeled, the last element to complete the interaction phase, is the tool used to interact with the virtual content. We are focusing in only one interaction way: push an object; so we propose to model the hand as a rigid body, where only the tip of the index finger will have a slight deformation when the contact occurs. Figure 2 shows the shape of the model. In order to match the model with the real hand, we propose a color-based approach to segment the hand in a real video sequence. Then a template-matching approach is used to select

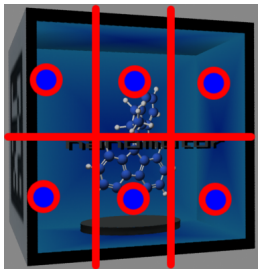


**Fig. 1.** Contact area

a more appropriate shape according to the size. During the interaction task we assume that the shape is constant.

**Fig. 2.** Push model hand

4. **Designing a feedback mechanism to guide the user during the interaction task**: when working with virtual objects in an interaction task is necessary to provide some sort of feedback to advise the user that is handling with a virtual object or maybe to guide him/her to the right position to be touched. We are interested in forms of haptic feedback; to this end, we will study different approaches on visual, auditory and tactile feedback. The integration of such mechanism to the interaction process will also be made in this stage.
5. **Evaluating the proposed strategy for physical-interaction in AR**: two applications will be designed in order to evaluate the proposed strategy. The first one will be designed so as to measure the difference of the final positions between the real and virtual objects after a force is applied on each one, as well as the difference of the paths followed to reach the final positions. This application will take place in the early stage of the research, in which the user will be asked to apply a force on a virtual object to show its behavior. In the later stage, we will focus on the second application, which aims to evaluate the user experience and usability. In order to achieve that a user study will be performed through this application.

We expect to achieve the goals of this research by following this methodology. In the next section we describe the preliminary results obtained so far.

## 3 Preliminary Results

This section describes the preliminary results of our research. It also describes the work preformed following the proposed research methodology.

### 3.1 Modeling a spatial representation of the work area

As a first approach, we propose to use a point-cloud from a 3D sensor to create a representation of the environment. Figure 3 shows a common point-cloud obtained from the sensor. Figure 3(a) shows the original point-cloud, which has over 307,000 points, composed by $[x, y, z]$ and color data. We can see the more black pixels being holes in which the virtual object cannot be attached. Figure 3(b) shows a correction of the point-cloud to fill holes. To do this, we used the inpainting-telea algorithm [19].

The original cloud have many points and it is impractical to be used in a real-time application, Figure 3(c) depicts an example of reducing the point-cloud density. By using thresholding, we can delete points out of the work area; however, the cloud remaining still has unnecessary points (there is not a significant difference in the $z$ component of many grouped points). In this point, we are interested in methods that allows reducing the redundant information without sacrificing significant geometries. After the cloud is reduced, the normal[2] of the faces (triangular faces) must be calculated and added to it; so, now the cloud has five components. A first candidate to be used in this point is the moving least squares method (MLS), which is used to reconstruct a surface from a set of points and to calculate the normal of the faces. This produces a smoothed cloud.

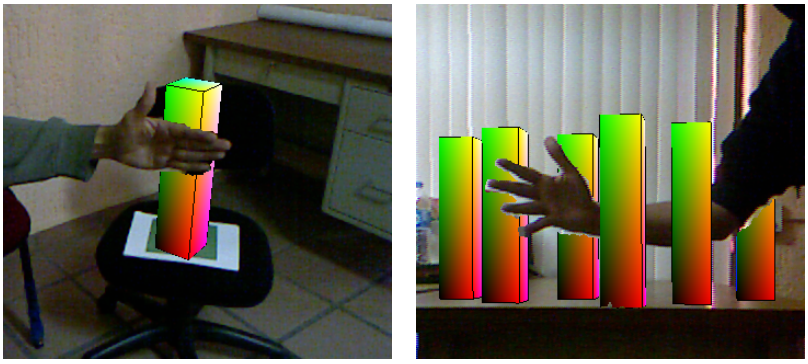
(a) raw data (b) inpainting (c) thresholding

**Fig. 3.** Point-cloud from kinect

Once we have the five components, the surface can be reconstructed. Figure (4) shows only the hand area that is reconstructed using the greedy triangulation

[2] In geometry, the normal is an object such as a line or vector that is perpendicular to a given object. In this case it can be seen as the direction of the faces.

algorithm, provided by PCL library[3]. The reconstructed surface enables the occlusion handling and lays the foundation for collision detection. Only the hand area was reconstructed because we are only using the threshold filter for reducing the cloud. Thus, this is still computationally expensive.



(a) (b)

**Fig. 4.** Enabling occlusion

### 3.2 Evaluating the proposed strategy

The evaluation will be conducted within the context of an application. In this point of the research, we have already defined a user study, which will be driven through this application. Furthermore, this study will help us to define the features that should be present in the application itself.

**Objective**: with this study we want to evaluate the proposed strategy for interacting in AR, in terms of realism and intuitiveness. The two main factors which we are interested in measuring are:

1. **User experience**: involves the person's behaviors, attitudes, and emotions about using our application.
   - Intuitiveness: involves the feeling of naturalness and human-likeness in interacting with the virtual content.
   - Captivation: involves the feeling of being immersed and captivated in the interaction with the AR.
   - Responsive: involves the feeling about the response time of the system to a particular action.
   - Tangibility: involves the feelings spatial-awareness and coherence of the content according with the environment.
2. **Usability**: involves the distribution of content and clarity with which the interaction is designed.

[3] `http://www.pointclouds.org`

- Learnability: involves the easy to accomplish a task the first time the user has contact with the application.
- Efficiency: involves the speed with which a task is performed.
- Errors: involves the number of errors made by the user in a task.

**External factors**: there are some external factors that can influence in the evaluation, we are interested in control the next two.

- **User experience**: the user experience with AR systems is a key factor in performing a task.
- **Incentives**: the willingness of the user to perform a task influenced in the result, a change in user motivation usually results in a change in what he/she does. In this point, we are studying the best way to encourage the participants without causing bias.

**Users**: the users play an important role in this evaluation, in order to cover as many aspects as possible and considering that the application is a game, the participants will be divided into three groups: 1) inexperienced users in augmented reality systems and video games, 2) inexperienced users in augmented reality but experienced in video games and 3) experienced users in augmented reality systems and video game.

**Setting**: each participant will be given a tutorial before performing the task. Then he/she will be asked by the system to perform certain tasks. While the user is performing the tasks, an observer will note for feedback purpose. When all tasks are completed, the participant will be asked to answer a poll covering all aforementioned aspects.

A more accurate evaluation also will be addressed in order to give foundations that allows the comparison with other works. This include measuring the differences between the final positions (translation and rotation) of the real and virtual objects after an interaction task, as well as measuring the differences between the paths followed to get the final position.

## 4 Conclusions

In this paper, we present our research proposal to address the problem of natural interaction in augmented reality. The main motivation to do this work is to advance in the state-of-the-art with respect to interaction approaches in AR using a physical-based modeling to reproduce the behavior of real objects into virtual objects. To achieve a realistic interaction we consider modeling both the virtual and the real objects. Although our research is in the early stage, the preliminary results showed in Figure 4 give evidence that it is possible to improve the spatial-awareness of the user. The user can determine at all time the right position of the objects in a mixed space, that suggest evidence that the modeling of both real and virtual elements is a right direction to achieve and realistic interaction in an AR environment. The future paths of our research are focused on exploring video-games techniques of physical modeling and propose methods to bring them to a real interaction. We are also interested in propose novel feedback mechanisms to guide the user along the interaction.

## References


1. Akman, O., Poelman, R., Caarls, W., Jonker, P.: Multi-cue hand detection and tracking for a head-mounted augmented reality system. Machine Vision and Applications 24(5), 931–946 (2013), http://dx.doi.org/10.1007/s00138-013-0500-6
2. Bai, H., Gao, L., El-Sana, J., Billinghurst, M.: Markerless 3D gesture-based interaction for handheld Augmented Reality interfaces. In: Mixed and Augmented Reality (ISMAR), 2013 IEEE International Symposium on. pp. 1–6 (Oct 2013)
3. Bellarbi, A., Benbelkacem, S., Zenati-Henda, N., Belhocine, M.: Hand gesture interaction using color-based method for tabletop interfaces. In: Intelligent Signal Processing (WISP), 2011 IEEE 7th International Symposium on. pp. 1–6 (Sept 2011)
4. Billinghurst, M., Kato, H., Myojin, S.: Advanced Interaction Techniques for Augmented Reality Applications. In: Shumaker, R. (ed.) Virtual and Mixed Reality, Lecture Notes in Computer Science, vol. 5622, pp. 13–22. Springer Berlin Heidelberg (2009), http://dx.doi.org/10.1007/978-3-642-02771-0_2
5. Corbett-Davies, S., Green, R., Clark, A.: Physically Interactive Tabletop Augmented Reality Using the Kinect. In: Proceedings of the 27th Conference on Image and Vision Computing New Zealand. pp. 210–215. IVCNZ '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2425836.2425880
6. Ha, T., Woo, W.: ARWand: Phone-Based 3D Object Manipulation in Augmented Reality Environment. In: Proceedings of the 2011 International Symposium on Ubiquitous Virtual Reality. pp. 44–47. ISUVR '11, IEEE Computer Society, Washington, DC, USA (2011)
7. Ha, T., Woo, W.: ARWand for an augmuented world builder. In: Lécuyer, A., Steinicke, F., Billinghurst, M. (eds.) 3DUI. pp. 207–208. IEEE (2013), http://dblp.uni-trier.de/db/conf/3dui/3dui2013.html#HaW13a;http://dx.doi.org/10.1109/3DUI.2013.6550251;http://www.bibsonomy.org/bibtex/2874b307c0dda8c24479f4bab168ba97f/dblp
8. Ha, T., Woo, W.: Poster: A pilot study on stepwise 6-DoF manipulation of virtual 3D objects using smartphone in wearable augmented reality environment. In: 3D User Interfaces (3DUI), 2013 IEEE Symposium on. pp. 137–138 (March 2013)
9. Hürst, W., Wezel, C.: Gesture-based interaction via finger tracking for mobile augmented reality. Multimedia Tools and Applications 62(1), 233–258 (2013), http://dx.doi.org/10.1007/s11042-011-0983-y
10. Kasahara, S., Niiyama, R., Heun, V., Ishii, H.: exTouch: Spatially-aware Embodied Manipulation of Actuated Objects Mediated by Augmented Reality. In: Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction. pp. 223–228. TEI '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2460625.2460661
11. Kato, H., Billinghurst, M., Poupyrev, I., Imamoto, K., Tachibana, K.: Virtual Object Manipulation on a Table-Top AR Environment, ISAR. Presence pp. 111–119 (2000)
12. Kato, H., Tachibana, K., Tanabe, M., Nakajima, T., Fukuda, Y.: MagicCup: a tangible interface for virtual objects manipulation in table-top augmented reality. In: Augmented Reality Toolkit Workshop, 2003. IEEE International. pp. 75–76 (Oct 2003)
13. Lee, B., Chun, J.: Interactive Manipulation of Augmented Objects in Marker-Less AR Using Vision-Based Hand Interaction. In: Information Technology: New Generations (ITNG), 2010 Seventh International Conference on. pp. 398–403 (April 2010)

14. Lee, S., Lim, Y., Chun, J.: 3D interaction in Augmented Reality with stereo-vision technique. In: Advanced Communication Technology (ICACT), 2013 15th International Conference on. pp. 401–405 (Jan 2013)
15. Maier, P., Klinker, G.: Evaluation of an Augmented-Reality-based 3D User Interface to Enhance the 3D-Understanding of Molecular Chemistry. In: Foley, O., Restivo, M.T., Uhomoibhi, J.O., Helfert, M. (eds.) CSEDU. pp. 294–302. SciTePress (2013), http://dblp.uni-trier.de/db/conf/csedu/csedu2013.html#0002K13;http://dx.doi.org/10.5220/0004349502940302;http://www.bibsonomy.org/bibtex/2ce5a6bb4c6284a3f8f59db9c77f30655/dblp
16. Piumsomboon, T., Clark, A., Billinghurst, M.: Physically-based Interaction for Tabletop Augmented Reality Using a Depth-sensing Camera for Environment Mapping. In: Proc. Image and Vision Computing New Zealand (IVCNZ-2011). pp. 161–166. Auckland (Dec 2011), http://www.ivs.auckland.ac.nz/ivcnz2011_temp/uploads/1345/3-Physically-based_Interaction_for_Tabletop_Augmented_Reality_Using_a_De.pdf
17. Song, P., Yu, H., Winkler, S.: Vision-based 3D Finger Interactions for Mixed Reality Games with Physics Simulation. In: Proceedings of The 7th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry. pp. 7:1–7:6. VRCAI '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1477862.1477871
18. Talvas, A., Marchal, M., Lécuyer, A.: The god-finger method for improving 3D interaction with virtual objects through simulation of contact area. In: 3DUI. pp. 111–114 (2013), http://dx.doi.org/10.1109/3DUI.2013.6550206
19. Telea, A.: An Image Inpainting Technique Based on the Fast Marching Method. journal of graphics, gpu, and game tools 9(1), 23–34 (2004)
20. Umetani, N., Takayama, K., Mitani, J., Igarashi, T.: A Responsive Finite Element Method to Aid Interactive Geometric Modeling. Computer Graphics and Applications, IEEE 31(5), 43–53 (Sept 2011)
21. Wagner, D., Schmalstieg, D.: First steps towards handheld augmented reality. In: Wearable Computers, 2003. Proceedings. Seventh IEEE International Symposium on. pp. 127–135 (Oct 2003)
22. Wu, J., Dick, C., Westermann, R.: Efficient collision detection for composite finite element simulation of cuts in deformable bodies. The Visual Computer 29(6-8), 739–749 (2013)
23. Xu, G.: Finite Element Methods for Geometric Modeling and Processing Using General Fourth Order Geometric Flows. In: Chen, F., Jüttler, B. (eds.) Advances in Geometric Modeling and Processing, Lecture Notes in Computer Science, vol. 4975, pp. 164–177. Springer Berlin Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-79246-8_13

# Image Classification through Text Mining techniques: a Proposal

A. Pastor López-Monroy[1], Manuel Montes-y-Gómez[1], Hugo Jair Escalante[1], and Fabio A. González[2]

[1] Instituto Nacional de Astrofísica, Óptica y Eletrónica (INAOE)
LabTL, Computer Science Department
Luis Enrique Erro No. 1, C.P. 72840, Tonantzintla, Puebla, México
{pastor,mmontesg,hugojair}@inaoep.mx
[2] National University of Colombia
MindLab, Computing Systems and Industrial Engineering Department
Cra 30 No 45 03-Ciudad Universitaria, Bogotá DC, Colombia.
fagonzalezo@unal.edu.co

**Abstract.** Image classification is an important task for the organization and analysis of visual information. According to the literature one of the most important concepts is the *visual word*; a visual element that represents a set of visual-similar regions. The Bag-of-Visual Words (BoVW) is one of the most widely used approaches in High Level Computer Vision (HLCV). The BoVW is an histogram of the occurrence of visual words in each image, which is in some way inspired by the Bag-of-Words (BoW) used in Natural Language Processing (NLP). In spite of the success of BoVW, it has the same limitations of BoW (e.g., the overlook of the spatial context). In this research proposal we bear in mind the successful evidence of visual words in HLCV, and we take the analogy of visual-textual words to a new higher level. This is, by designing methods inspired in NLP, we aim to consider contextual (e.g., spatial, sequential), and high level (e.g., semantic) information among visual words. However, bringing NLP like approaches pose several nontrivial problems, for example: i) the definition of analogous attributes (visual-textual), ii) a suitable strategy to interpret images; documents can be read only in one direction, but in images we have a 2D plane without an specific way to *read* them, iii) the way to extract high level information (e.g., semantic). This paper presents the proposed research methodology and through preliminary results, we provide strong evidence of the feasibility of this research. For this, a popular NLP technique is used to improve the BoVW; the Bag-of-Visual $n$-grams (BoVN). The idea is evaluated in the challenging task of Histopathology image classification overcoming the BoVW and an state-of-the-art approach based in language models.

**Key words:** Visual words, n-grams, spatial context, image classification, histopathology.

## 1 Introduction

Nowadays there is a huge amount of images available through different media sources. In many situations all this information is useless without appropriate tools for analysis. In this regard, image classification is one of the most important tasks for the organization and exploitation of visual information for different areas. The representation of images is one of the key procedures for successful models in classification. Currently one of the most widely used approaches in the state-of-the-art of High Level Computer Vision (HLCV) tasks is the Bag-of-Visual Words (BoVW). The BoVW is somehow inspired by the Bag-of-Words (BoW) representation of text mining (see e.g., [16]). Under the BoW formulation, vocabulary vectors representing documents are built, and each element of the vector indicates the presence or absence of each word in the document. Similarly, in HLCV tasks a vocabulary of visual word is generated (clustering feature vectors representing image regions and taking the centroid of each cluster as a visual word) in order to represent images through vectors that accounts for the occurrence of visual words in each image (see Figure 3). The BoVW has been successfully used in several HLCV task including: medical image categorization [2, 3], texture and object classification [20], video retrieval [15], image retrieval [18], human activity recognition [19], etc.

**Problem to solve:** Notwithstanding the fact that visual words approaches (like BoVW) are widely used, they usually do not exploit the contextual (spatial relationships) and high level (e.g., semantic) information among visual words. Spatial context has proven to be useful to increase the performance of several HLCV tasks (see e.g., [5, 9]). In this direction, contextual and high level information among visual words could be captured taking the analogy visual-textual words into a complete new higher level using Natural Language Processing (NLP) approaches.

**Main Objective:** Designing and developing methods for image classification, which based on the concept of visual word and inspired by NLP approaches, can model contextual and high level information to improve the classification.

**Main Contribution:** the design of novel and effective HLCV methods inspired by NLP, that consider the properties of the image domain to exploit effectively contextual and high level information among visual words. For example, novel methods based on: $n$-grams (sequences of $n$ elements), weighting schemes (weight functions for the visual elements), semantical distributional analysis, etc.

In this context, the interest of this research lies in a relatively young area, the intersection of the fields of NLP and HLCV, which has been the main subject of study of different forums and works [1,14,17,18]. To design a successful approach we focus on techniques that have proven to be highly useful in NLP. To figure out whether the best approaches in NLP have the opportunity to improve visual words methods, we begin exploring one basic, intuitive, yet effective idea of NLP; $n$-grams. $n$-grams are sequences of $n$ elements which have proven to be very useful in text categorization tasks for capturing the context [16]. Through the achieved results, we show the feasibility of this proposal giving two main contributions: i) a method to extract $n$-grams from visual words, and ii) the way to effectively

use $n$-grams as attributes for a classifier. The proposal overcomes the traditional BoVW and an state-of-the-art approach based on language models. The rest of this paper is organized as follows. Section 2 describes the proposed research methodology. Section 3 describes the dataset. Section 4 introduces our approach. Section 5 shows and discusses preliminary results. Finally, Section 6 shows our conclusions and indicates the paths of future research.

## 2 Research Methodology

The research methodology is as follows (main contributions are steps 3 and 4):

1. **Identifying and obtaining copora:** To find datasets with challenging peculiarities that NLP methods could handle (e.g., contextual information).
2. **Analyzing and developing methods to extract visual words:** To identify methods to get a better analogy between visual-textual words, our initial paths considers: i) Extraction of regions through regular grids [13], and ii) Extraction of regions through key points [12].
3. **Proposing a set of new representations inspired in NLP to capture contextual information:** For this, we consider the following three approaches as the best candidates, which aims to capture contextual information at different levels:
   (a) **Sequences of visual words to capture the pure local context**: The general idea is to use sequences of elements similar to the $n$-grams (sequences of $n$ words) for text mining [6]. A challenge here consists in defining a suitable way to extract such $n$-grams. This is because in contrast to text documents, in images we have a 2D plane, and the way to read the elements is not defined.
   (b) **Locally weighted bag of words to capture local-global context:** Using this approach a higher level of contextual information can be capture [4]. Through this representation it is possible to assign different weighs to several *parts* of a document, which in analogy could facilitates to focus in relevant image *regions*.
   (c) **$n$-gram graphs to capture the pure global context:** In the classic text mining graphs of words [7], nodes would represent visual words, which are connected among them by edges modeling the co-occurrence, frequency and order. Such graphs allow to capture global information about the elements in the target object (images).
4. **Proposing a set of new representations inspired in NLP to capture high level information:** For this we consider the following three approaches as the best candidates, which aims to capture high level information (e.g., semantic) when bringing to the image domain.
   (a) **Concise Semantic Analysis (CSA) of visual words:** This is an special distributional semantical representation, which through the use of low-dimensional vectors, allows to capture relationships among documents and the target classes [11]. In analogy, adapting this approach is

possible to build image vectors that highlights discriminative relationships with each target class.

(b) **Knowledge based hierarchies (*is-a* hierarchies):** These kind of hierarchies are widely used in text mining to represent semantic relationships among words of specific domains. Similarly, having a hierarchy for visual words would makes possible to capture different information. Building such hierarchies for visual words is challenging, but could be achieved in different ways (etc. using the hierarchy produced by a hierarchical clustering, using distributional semantical representations, etc.).

5. **Designing and implementing a method to integrate the information extracted by the NLP inspired approaches:** This last step involves to take advantage of different spaces of information. There are several ways in the literature to achieve this combination (e.g., classifier ensemble techniques, multiple kernel learning, etc.) [8].

In the following section, we present the work done so far. For this, we describe the initially dataset, then we explain the process to build the visual words and extract visual $n$-grams to improve BoVW (item 3(a) of the proposed methodology). Finally, we discuss the obtained results and future avenues of inquiry.

## 3 Description of the Image Collection

The proposed methods and representations in this research will be evaluated using several image collections ranging from natural to medical images (selected collections depend of the Step 1 of the methodology). For the evaluation of the sequences of visual words we initially perform experiments using an Histopathology image collection. We decide to use this kind of images, because their visual tissues structures (healthy or pathological) make them challenging. In this images, classification is related to pathological lesions and morphological-architectural features which can be captured by our proposed visual $n$-grams (see Figure 1).

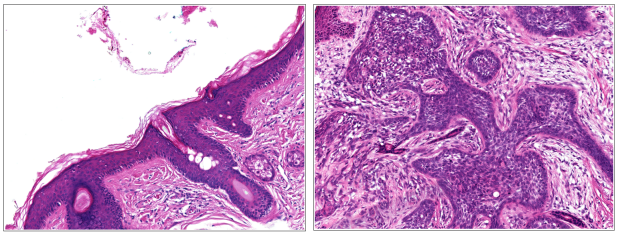


Fig. 1: Example of histopathology images from skin biopsies with healthy and pathological tissues (basal-cell carcinoma), left and right respectively.

In the evaluation we use a dataset of 1417 histopathology images, annotated by a pathologist, describing the presence of architectural features, and pathological tissues [3]. Each image might belong to one or more of 7 categories (see

distribution in Table 1). To evaluate our approach we built a binary classifier using an standard one-vs-rest approach.

| **Histopathology class** | **positives** | **negatives** |
|---|---|---|
| 1. basal-cell carcinoma | 518 | 899 |
| 2. collagen | 1238 | 179 |
| 3. epidermis | 147 | 1270 |
| 4. hair follicle | 118 | 1299 |
| 5. eccrine glands | 126 | 1291 |
| 6. sebaceous glands | 136 | 1281 |
| 7. inflammatory infiltrate | 99 | 1318 |

Table 1: The seven binary problems of the 1417 Histopathology image collection. The positive instances are images belonging to a target category.

## 4 Sequences of Visual Words: Visual *n*-grams

Our first approximation to capture the local context of visual words is through the use of one popular technique in NLP: the $n$-grams. For this we propouse the use $n$-grams of visual words to improve the BoVW. In other words, we focus in the Bag-of-Visual $n$-grams (BoVN). In Figure 2 we outlines each step of the process for generating the BoVN. In the first step, the training collection is used to generate the dictionary of visual words (codebook) (explained in Section 4.1). In the second step, each patch of each image is replaced by the nearest visual word in the codebook. The second step also involves the extraction of $n$-grams in order to build our visual $n$-gram codebook (explained in Section 4.2). The third step combines the visual words codebook and the visual $n$-gram codebook. The final codebook is used to build histograms of the visual $n$-grams in each image. We explain in detail the latter steps in the following subsections.

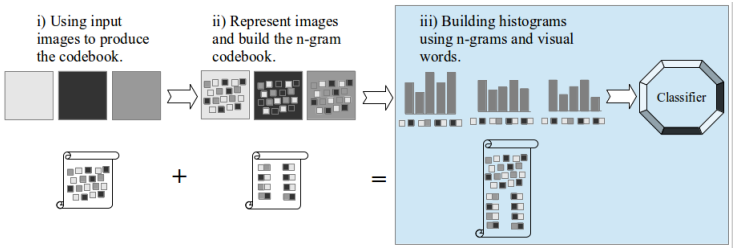



Fig. 2: Image Representation through Bag-of_Visual-Ngrams.

### 4.1 Construction of the Visual Words Codebook

In Figure 3 (A), we show the extraction of visual words using a standard BoVW formulation. In the step 1 and 2, we use a regular-grid-based patch extraction. In step 3, we represent each patch using the discrete cosine transform (DCT) applied to each channel of the RGB color space. We merge the 64 coefficients from each of the three channels to get the final descriptor. In the last step, the codebook is built using cluster centroids (using a 400-Means algorithm) of the training patch descriptors. Those settings are supported by previous closely-related studies using Histopathology images, showing better performance than other configurations (including SIFT and raw-patches, and several $k$ values) [2].

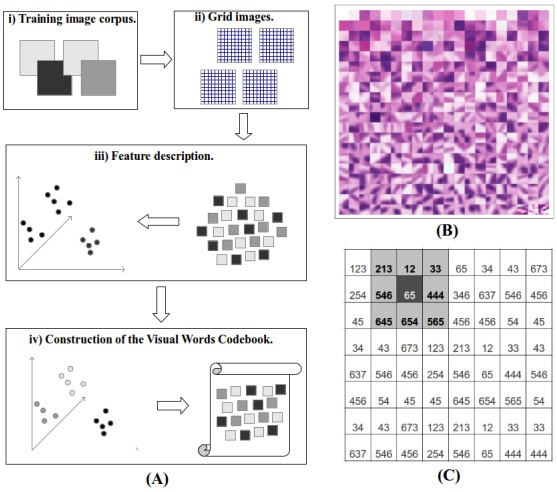



Fig. 3: (A) The process to build a visual word codebook. (B) Example of a generated visual word codebook. (C) The process to build visual $n$-grams using an sliding window. For the dark path (65) the extracted $n$-grams are: 65-12, 65-213, 65-546, 65-645, 65-654, 65-565, 65-444, 65-33.

### 4.2 Extraction of visual n-grams

To capture spatial relationships among visual words, we inspired our idea in the use of word $n$-grams for text classification tasks. They are sequences of $n$ consecutive words that helps to maintain semantic relationships between words, which allows to represents as one attribute concepts like "cold war". Nonetheless in image domain, the extraction of visual $n$-grams face some additional issues. For example, a document can be read only in one direction, but sequences of image descriptors can be extracted horizontally, vertical, at an angle of $\theta$ degrees, etc.). Another problem is to determine the right direction to interpret each visual

$n$-gram. For example, 3-grams in text normally can be interpreted correctly only in one direction (say, "the human being", but not "being human the"). On the other hand, visual 3-grams can have the same order but different orientation if the image is rotated. Therefore, the two descriptor sequences $d_a$-$d_b$-$d_c$ and $d_c$-$d_b$-$d_a$ might be the same pattern. In this work, we consider such patterns the same, making them rotation invariant.

To construct visual $n$-grams we apply the following effective approach. Consider a document containing the codeword matrix for each image (see (C) in Figure 3), the main idea is to produce $n$-grams ignoring the orientation in which they appear. For this, we iterate over each item $a_{i,j}$ of the matrix A and we extract neighbors in a straight fashion. Thus, we build horizontal,vertical and diagonal sequences using items between the current item $a_{i,j}$ and $a_{i+k,j+h}$, if and only if they are part of the straight line joining $a_{i,j}$ and $a_{i+k,j+h}$. This approximation to text $n$-grams produces eight possible $n$-grams for each position in the matrix. Finally, each $n$-gram is normalized to be "read" only in one direction. For example, in the visual $n$-gram codebook, a trigram 21-61-73 is indexed as the same item than 73-61-21. For image classification we build feature vectors using a BoVN. This is, using the learned codebook, each image is represented by an histogram of the occurrence of found visual $n$-grams. We use a Support Vector Machine (SVM) using the default settings of Weka framework. We used a SVM because it has shown to be effective in similar histology image problems finding visual patterns [2, 3].

## 5 Preliminary Results on Image Clasification using Sequences of Visual Words

In the evaluation we use an stratified 10 fold cross validation (10FCV) and report the averaged F-Measure and Area Under roc Curve (AUC) of the seven binary problems. We have used several scenarios, for term weighting; binary (BIN), and term frequency (TF). The former focuses only in the presence/absence of the element, the latter in the weighted presence/absence. For size patches we have: 8x8, and 16x16. Finally, as text mining evidence suggest [16], $k$ $n$-grams includes $k$ $(n-1)$-$grams$, $k$ $(n-2)$-$grams$, $\ldots$, $k$ (1)-$grams$. Finally, it is worth knowing that we perform a set of specific experiments that highlights the general usefulness of visual $n$-grams, but a detailed study of the parameterization and other applications can be found in [10].

**First Experiment:** Table 2, thought the first row, shows the results of evaluating the traditional BoVW (Unigramas). There the 8 size patch using TF weighting obtains the best results. The 8x8 size patch seems to be a good size of resolution to cover the biological structure of cells, which confirms results reported of other works in this dataset [2]. On the other hand, the TF weighting best results suggest that, in general is a good choice the accounting of visual patterns rather than focusing only in their presence/absence.

In Table 2 we also present results of BoVN and an state-of-the-art approach under the same conditions. Since the number of possible $n$-grams are of hun-

dreds of thousands, we analyze the performance influence in the performance of BoVN (testing with values from one thousand to ten thousand of features), getting that 2500 bigrams are a good balance between the dimensionality and the performance of our approach. The results show evidence of the usefulness of $n$-grams, showing that, every experiment using visual bigrams outperforms unigrams. Also in Table 2 we show results of another classical approach for visual words in the literature; language models. In this paper we have implemented a Language Model Classifier (LMC) as the one used in [17]. The goal is to compare the BoVN with other approaches in HLCV that also take advantage of the contextual information. To train language models, we have used exactly the same software and parameters (*Carnegie Mellon Statistical Language Modeling Toolkit*). As can be seen from Table 2, at least for this problem and under the same conditions, LMC does not get better performance than BoVN. In part, this might be because language models rely in probabilistic, where the unbalanced data represents a common problem to build accurate models for positive classes.

| *Performance of BoVN* | | | | | |
|---|---|---|---|---|---|
| | | *8x8* | | *16x16* | |
| *Approach* | *Criteria* | *Bin* | *TF* | *Bin* | *TF* |
| BoVW (1grams) | FM | 48.27 | 58.90 | 47.63 | 52.33 |
| | AUC | 67.74 | 72.27 | 67.56 | 68.89 |
| BoVN (1+2grams) | FM | 59.50 | **64.31** | 56.67 | 56.09 |
| | AUC | 72.54 | **76.03** | 70.46 | 71.17 |
| LMC (1+2+3grams) | FM | 53.0 | | 48.31 | |
| | AUC | 69.89 | | 72.21 | |

Table 2: Experiments using Uni-Bi-grams (sequences of visual words), two kinds of term weighting (TF and BIN) and two different size patches (8 and 16).

In other set of experiments we also analyze the performance considering higher order $n$-grams (e.g., 3grams, 4grams, etc.). Results show that the best setting is $1+2$grams, which is somewhat expected because it is well known that for a higher $n$-grams more instances are required to find those sequences [16]

**Second Experiment:** In Table 3 using an 10FCV, we analyze the detailed performance for each class. Thus, for BoVW and BoVN, we use the best settings for each method in Table 2. Results in Tables 3 shows that 1+2grams overcomes 1grams in classes 1, 3, 4, and 5. The class 1 is the most important, because it is the only one related with cancer diagnosis. Images in class 1 present structural tumor cells having large and darker nuclei, which are accurately characterized by visual bigrams. Visual words (1-grams) are competitive or better in classes 2, 6 and 7 (none of them related with cancer diagnosis). Such classes are in opposite ends, either by the lack of structured spatial visual elements (classes 2 and 6) that make bigrams to lose their advantage, or because the contextual information of visual words are much more global rather than local (class 6). We

think those problems need more instances and explore other parameters (e.g. patch sizes, size of sequences, or alternative descriptors).

| *Detailed F-Measure by class* | | | |
|---|---|---|---|
| *Class* | *(a) 1grams* | *(b) 1+2grams* | *(b-a) gain/loss* |
| 1 | 86.10 | 90.70 | 4.6 |
| 2 | 94.80 | 95.50 | 0.7 |
| 3 | 74.40 | 83.40 | 9.0 |
| 4 | 36.80 | 50.80 | 14.0 |
| 5 | 35.80 | 52.50 | 16.7 |
| 6 | 48.00 | 43.60 | -4.4 |
| 7 | 34.20 | 33.70 | -0.5 |

| *Detailed AUC by class* | | | |
|---|---|---|---|
| *Class* | *(a) 1grams* | *(b) 1+2grams* | *(b-a) gain/loss* |
| 1 | 89.00 | 92.50 | 3.5 |
| 2 | 76.40 | 79.20 | 2.8 |
| 3 | 84.00 | 90.90 | 6.9 |
| 4 | 62.60 | 68.80 | 6.2 |
| 5 | 62.80 | 71.60 | 8.8 |
| 6 | 68.70 | 66.90 | -1.8 |
| 7 | 62.40 | 62.30 | -0.1 |

Table 3: Detailed experiments per class using Unigrams versus Uni-Bi-grams. In column"(b-a) gain/loss" we show the gain or loss caused by the use of bigrams.

## 6 Conclusions

The interest of this research lies in the fields of HLCV and NLP. The underlying motivation is to improve state-of-the-art visual words approaches (such as the BoVW) through methods that takes the analogy visual-textual words into a new higher level. For this, we consider contextual (spatial) and high level (semantic) information, which is overlooked by several approaches like BoVW. Since the use of the contextual information is a common factor in NLP tasks, as an initial approach, we propose the natural extension to BoVW; the use on visual $n$-grams as attributes (the BoVN). Our results suggest strong evidence of the usefulness of BoVN in Histophalogy images. To the best of our knowledge, $n$-grams have never been extracted as we propose; in a similar way they boost NLP tasks, and subsequently use them as feature vectors for a classifier. Future research paths include bringing ideas to capture contextual information in a more global way, and extracting high level information among visual words.

## References

1. Bruni, E., Tran, K.N., Baroni, M.: Multimodal distributional semantics. Journal of Artificial Intelligence Research (49), 1–47 (2014)
2. Cruz-Roa, A., Caicedo, J.C., González, F.A.: Visual pattern mining in histology image collections using bag of features. Artificial intelligence in medicine 52, 91–106 (2011)
3. Díaz, G., Romero, E.: Micro-structural tissue analysis for automatic histopathological image annotation. Microscopy Research and Technique 75, 343–358 (2012)

4. Escalante, H.J., Solorio, T., Montes-y Gómez, M.: Local histograms of character n-grams for authorship attribution. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. pp. 288–298 (2011)
5. Galleguillos, C., Belongie, S.: Context based object categorization: A critical survey. Computer Vision and Image Understanding 114, 712–722 (2010)
6. García-Hernández, R.A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A.: A new algorithm for fast discovery of maximal sequential patterns in a document collection. In: Computational Linguistics and Intelligent Text Processing, pp. 514–523. Springer (2006)
7. Giannakopoulos, G., Mavridi, P., Paliouras, G., Papadakis, G., Tserpes, K.: Representation models for text classification: a comparative analysis over three web document types. In: Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics. p. 13. ACM (2012)
8. Kuncheva, L.: Combining pattern classifiers. Wiley Press, New York pp. 241–259 (2005)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006. vol. 2, pp. 2169–2178. IEEE (2006)
10. López-Monroy, A.P., Montes-y Gómez, M., Escalante, H.J., Cruz-Roa, A., González, F.A.: Bag-of-visual-ngrams for histopathology image classification. In: IX International Seminar on Medical Information Processing and Analysis, vol. 8922, p. 89220P. SPIE (2013)
11. López-Monroy, A.P., Montes-y-Gómez, M., Villaseñor-Pineda, L., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: A new document author representation for authorship attribution. In: Mexican Conference in Pattern Recognition. pp. 283–292. Springer (2012)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2), 91–110 (2004)
13. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Computer Vision–ECCV 2006, pp. 490–503. Springer (2006)
14. Quack, T., Ferrari, V., Leibe, B., Van Gool, L.: Efficient mining of frequent and distinctive feature configurations. In: IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)
15. Sivic, J., Zisserman., A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision (2003)
16. Tan, C.M., Wang, Y.F., Lee, C.D.: The use of bigrams to enhance text categorization. Information processing and management 38, 529–546 (2002)
17. Tirilly, P., Claveau, V., Gros, P.: Language modeling for bag-of-visual words image categorization. In: ACM Proceedings of the 2008 international conference on Content-based image and video retrieval. pp. 249–258 (2008)
18. Tirilly, P., Claveau, V., Gros, P.: A review of weighting schemes for bag of visual words image retrieval. Tech. rep., Technical report, TEXMEX - INRIA - IRISA (2009)
19. Wang, H., Ullah, M.M., Klaser, A., Laptev, I.: Evaluation of local spatio-temporal features for action recognition. In: Proceedings of the British Machine Vision Conference. pp. 1–11 (2009)
20. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision 73, 213–238 (2007)

# Graph-Based Representations for Computational Linguistics Tasks

Nahun Loya[1], Iván Olmos[1], David Pinto[1], and Jesús González[2]

[1] Benemérita Universidad Autonóma de Puebla
Faculty of Computer Science
Av. San Claudio y 14 Sur, C.P. 72570, Puebla Pue., Mexico
{nahun.loya,iolmos,dpinto}@cs.buap.mx
[2] National Institute of Astrophysics, Optics and electronics
Department of Computer Science
Luis Enrique Erro No. 1, C.P. 72840, Tonantzintla, Puebla, Mexico
jagonzalez@inaoep.mx

**Abstract.** In this work is introduced a methodology for discover hidden patterns in texts based on a graph-based representation. This methodology is divided on three main steps: first, a pre-processing phase, where texts are cleaning of unuseful data; second, we propose three different graph-based representations, where texts can be mapped to labeled graphs; third, the data mining phase, which can be performed with the SUBDUE system. Also, we show how could be interpreted the outputs patterns. The advantage of our methodology in opposite with others techniques is that allow extract structured patterns from texts.

**Key words:** Graph-based representations, PLN, authorship attribution, author profiling identification.

## 1 Introduction

Nowadays, there is a huge amount of data based on texts. Therefore users need modern tools capable to analyze and extract useful information over data, because is not possible a manual analysis since the huge volume of data. Text mining is a new burgeoning research field that attempts to glean meaningful information from natural language texts, with the aim to extract useful information (knowledge) [5].

A key factor to have success in the extraction knowledge process is the text representation. In the Natural Language Processing (NLP) has been proposed different text representations, such as bags of words [3], analysis with "n-grams" (detection of contiguous sequeces of words) [4]. More sofisticated representations include the use of ontologies, which represents knowledge as a set of concepts within a domain, and the relationships between those concepts [2], or techniques based on the Latent Semantic Indexing, where queries and documents are represented into a space with latent semantic dimensions [1].

Based on the above mentioned, there are different ways to represent texts, each of one with advantages or disadvantages (in terms of space complexity, time

complexity, or capability of preserve patterns that exist in the original texts). In this sense, this work introduce the use of labeled graphs as text representation, because is possible to include many linguistics levels of the original text such as, lexical, morphological, syntactical, etc. Furthermore, graphs let us express concepts and relations using their nodes and their edges.

The main contribution of this work is the introduction of a methodology that define how input texts can be represented with labeled graphs, with the aim to discover common patterns (common graphs) in a data mining task, taking advantage of the flexibility of the graphs to represent complex structures.

The rest of the paper is structured as follows: Section 2 introduced some notation useful to represent labeled graphs. Section 3 presents the proposed methodology for discovering patterns in text based on a graph representation. Section 4 introduce general concepts with the aim to understand the graph-based data mining phase. Finally, in section 5 is presented the conclusions.

## 2 Graph definitons and basic concepts

Several researchers have been proposed different graph notations. The most widely used graph notation for graphs is $G = (V, E)$, where $V$ is a set (not empty) of vertices, and $E$ is a set of edges, $E \subseteq V \times V$. However this notation is not suitable for text mining, because data information is represented through labels that are attached to the vertices and edges. Based on the above mentioned, is introduced a graph notation for labeled/unlabeled graphs.

**Definition 1** *A labeled graph $G$ is a 6-tuple $G = (V, E, L_V, L_E, \alpha, \beta)$, where:*

- $V = \{v_i | i = 1, ..., n\}$ is a finite set of nodes, $V \neq \emptyset$, and $n = \#vertices$ in $G$.
- $E \subseteq V \times V$ is a finite set of edges, $E = \{e = \{v_i, v_j\} | v_i, v_j \in V, 1 \leq i, j \leq n\}$.
- $L_V$, is a set of nodes labels.
- $L_E$, is a set of edges labels.
- $\alpha : V \rightarrow L_V$, is a function that assigning labels to the nodes.
- $\beta : E \rightarrow L_E$, is a function that assigning labels to the edges.

For unlabeled graphs and without loss of generality, it is possible to assign a same label for all vertices and all edges. Based on this notation, it is possible to introduce different graph topologies useful for our work.

**Definition 2** *A chain topology is defined as the set of all graphs $G = \{V, E, L_V, L_E, \alpha, \beta\}$, where: $V = \{v_j | j = 1, \ldots, n\}$; $E = \{e = \{v_j, v_{j+1}\}: 1 \leq j \leq n-1\}$; $L_V$ = Set of label of vertices; $L_E$ = Set of label of edges; $\alpha : V \rightarrow L_V$.; $\beta : E \rightarrow L_E$..*

**Definition 3** *The ring topology is the set of all graphs $G = \{V, E, L_V, L_E, \alpha, \beta\}$, where: $V = \{v_j | j = 1, \ldots, n\}$; $E = \{e = \{v_j, v_{j+1}\}: j = 1 \ldots n-1\} \bigcup \{e = \{v_n, v_1\}\}$; $L_V$ = Set of label of vertices; $L_E$ = Set of label of edges; $\alpha : V \rightarrow L_V$,; $\beta$: $E \rightarrow L_E$.*

**Definition 4** *The star topology is the set of all graphs* $G = \{V, E, L_V, L_E, \alpha, \beta\}$, *where:* $V_1 = \{v_0\}$; $V_2 = \{v_j | j = 1, \ldots, n\}$; $V = V_1 \bigcup V_2$; $E = \{e = \{v_0, v_j\}: 1 \leq j \leq n\}$; $L_V$ = *Set of label of vertices;* $L_E$ = *Set of label of edges;* $\alpha = \{\alpha(v) = etq, etq \in L_V, v \in V_2\} \bigcup \{\alpha(v_0) = inicial, v_0 \in V_1\}$; $\beta = \{\beta(e) = etq, etq \in L_E \ y \ e \in E\}$

For example, the sentence *"Abraham Lincoln fue presidente de EU"* can be mapped to a graph with a star topology as is shown in figure 1, which is defined as follows: $G = \{V, E, L_V, L_E, \alpha, \beta\}$, where: $V = \{v_{ini}, v_j | \ j = 1, \ldots, 6\}$; $E = \{e = \{v_{ini}, v_j\}: 1 \leq j \leq 6\}$; $L_V = \{Abraham, Lincoln, fue, presidente, de, EU\}$; $L_E = \{next\}$; $\alpha = \{\alpha(v_1) = Abraham, \alpha(v_2) = Lincoln, \alpha(v_3) = fue, \alpha(v_4) = presidente, \alpha(v_5) = de, \alpha(v_6) = EU\}$; $\beta = \{\beta(e_1) = next, \ldots, \beta(e_6) = next\}$

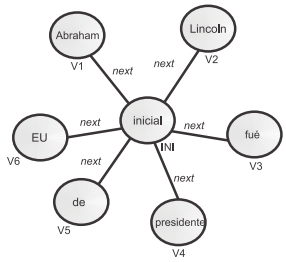



**Fig. 1.** Example of a sentence mapped to a graph with a star topology.

Important graph topologies were introduced in this section: chain, ring, and star. However, there exist more topologies (see figure 2), some of one based on the above described, such as: trees (derived from stars), backbone, backbone-tree, star ring, etc. Our work is based on trees and stars topologies, but it is possible to propose new representations with others topologies.

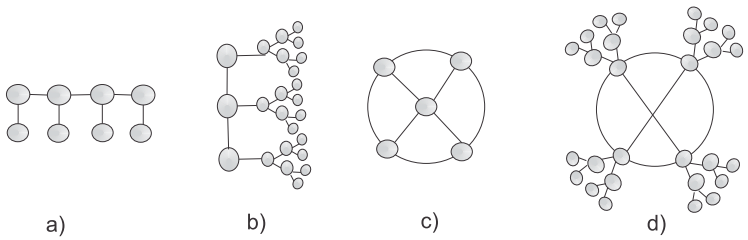



**Fig. 2.** Different types of topologies: a)Backbone, b)Backbone-tree, c)Star ring, d)Star ring with trees.

## 3 Methodology for text representation using labeled graphs

The main contribution of our work is the introduction of a methodology where it is possible to reduce texts to labeled graphs (transformation process). The metodology is presented in this section, which consist of three main steps, as is shown in figure 3:

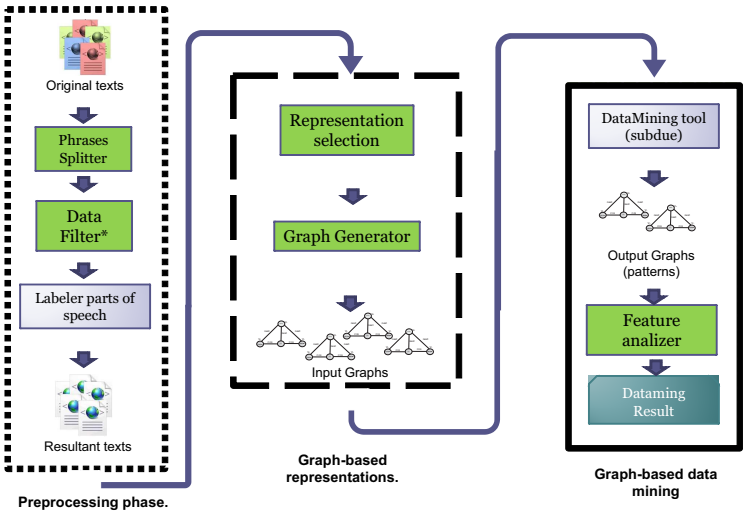



**Fig. 3.** Steps of the Proposed methodology.

### 3.1 Preprocessing phase

The preprocessing phase implements three steps, where the input texts are splitting, cleaning, and tagged. Consider that the input text has a plain format. Then, the text is divided in sentences, delimited by the puntuation mark ".". Clearly, there are some cases where this rule should not be applied. For example, if appear suspension points ("Dudé..", in Spanish), and abbreviations ("Dr. Juan Peréz", in Spanish). Consequently, it is necessary to detect " tokens", where words with a single mark "." at the end are considered as tokens. Next, stopwords are removed, because they not apport relevant information about the semantic of the text. Then, a morphological tagger is used[3] with the aim to detect features of the words, such as lemmas, dependency relations between words, and others. Thus, in this step every word in the text is labeled, depending of the roll that each

[3] In this work is used the Freeling tagger, www.freeling.org

word take into the sentence. As result of this, in the next step (process where texts are mapped to graphs) is possible to use different type of information of the input text, such as the original words of the text, lemmas, parts of the speach (grammatical categories of the words), sequence relations between words (the "next" relation), and type dependency relation between words.

### 3.2 Graph representations for texts

In this section, three graph-based representations are proposed, which are useful to represent different levels of information that the original texts have. As we mentioned before, the Freeling tool is used to generate the gramatical categories of each word. For the sake of simplicity, the set of all grammatical categories included in the input text is represented with POS, the set of all lemmas is represented with PALLEM, and the words of the input text is denoted with WORDS.

**Sequential three representation (STR).** This representation is focused on modeling the structure of a sentence. This representation is called "Sequential three representation (STR), based on a hybrid star topology. In STR the words are mapped to vertices, where vertices are labeled with the lemma of each word. Also, two types of relations are established between vertices: first, all vertices that represent neighborhood words (in sequence) are linked with an edge labeled with the word "next". Also, these vertices are linked to a vertex called "enun" with an edge labeled with the word "POS", as is shown in figure 4. This topology is focused to detect common structures (common graphs) between two or more sentences, including secuencial or not secuencial words.

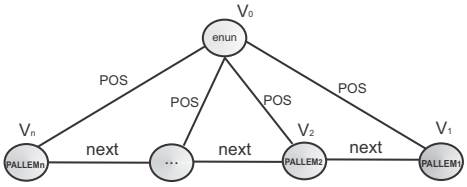



**Fig. 4.** Representation (STR).

Formally the representation is expressed as follows. Consider a sentence $S =< word_1, word_2, \ldots, word_n >$, where $word_i$ represents a word in the input sentence (that previously has been preprocessed). Consider that $POS(word_i)$ represents the grammatical class of the word $word_i$. A graph $G = \{V, E, L_V, L_E, \alpha, \beta\}$ is a STR representation of the sentence $S =< word_1, word_2, \ldots, word_n >$, if:

- $V = \{v_0\} \cup V_{PALLEM}$, where $V_{PALLEM} = \{v_1, \ldots, v_n\}$
- $E = \{\{v_0, v_i\}$, where $v_i \in V_{PALLEM}$, $i = 1, \ldots, n\} \cup \{\{v_i, v_{i+1}\}$, where $v_i \in V_{PALLEM}, i = 1, \ldots, n-1\}$

- $L_V = \{enun\} \cup L_{PALLEM}$, where $L_{PALLEM} = \{PALLEM(word_1) \ldots, PALLEM(word_n)\}$
- $L_E = \{next\} \cup L_{POS}$, where $L_{POS} = \{POS(word_1) \ldots, POS(word_n)\}$
- $\alpha : V \rightarrow L_V$ where $\alpha(v_0) = enun$, $\alpha(v_i) = PALLEM(word_i)$ if $v_i \in V_{PALLEM}$
- $\beta :\rightarrow L_E$, where $\beta(\{v_x, v_{X+1}\}) = next$ if $v_x \in V_{PALLEM}$ and $x = 1, \ldots, n-1$, either $\beta(\{v_x, v_y\}) = POS(word_i)$ if $v_x = v_0$ and $v_y \in V_{PALLEM}$

For example, consider the following two sentences: S1="Abraham Lincoln fue presidente de Estados Unidos"; S2="Estados Unidos tuvo un presidente llamado Abraham Lincoln". After the preprocessing phase, the original sentences are lemmatized: S1'="Abraham_Lincoln ser presidente Estados_Unidos"; S2'="Estados_Unidos tener presidente llamar Abraham_Lincoln". Using Freeling, it is possible to calculate the set of grammatical categories (POS). The results of this phase are: S1"="NP VS NC NP"; S2"="NP VM NC VM NP". Finally, the STR representation of S1" and S2" is shown in figure 5.

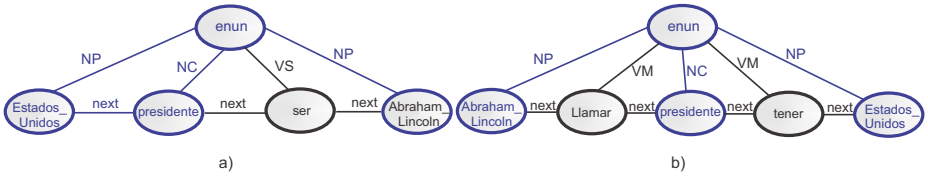



**Fig. 5.** Example of STR representation.

Based on these graph representations of the original sentences, with a GBDM tool such as SUBDUE, it is possible to discover common structures between these graphs (represented with no-black lines). Note that the common patterns includes words that are not in a strict order. Due to the structure of STR, the graph-based data mining process is capable to discover such patterns. As consequence, this representation can be used to discover common structures between sentences, including: common lemmas (of the words), common lemmas that appear in sequence, and common structure of the sentences that are similar in the input sentences.

**Representation based on word sequence (RBWS).** The next representation is called "Representation based on word sequence (RBWS)". The aim of this representation is to detect the relationship between words in each sentence. For each different word in the input sentence, is created a node in the graph, labeled with its lemma. Then, and edge between two nodes of the graph is created with the label "next", if and only if the words associated with these vertices appear in sequence in the original sentence. Therefore, this representation is not based on a specific graph topology, instead it is defined by the structure of the sentence. As consequence, it is possible to discover $n$-grams, but unlike the original $n$-gram

model, in RBWS is not defined a particular value for $n$. This representation is shown in figure 6.

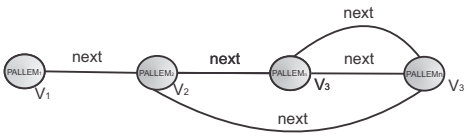



**Fig. 6.** Representation (RBWS).

For example, consider the following two sentences: S1="Abraham Lincoln fue presidente de Estados Unidos. El presidente dio los fundamentos de la libertad de Estados Unidos"; S2="Los fundamentos de la libertad fueron dados por el presidente de los Estados Unidos". After the lemmatization process of the preprocessing phase, the result sentences are: S1'="Abraham_Lincoln ser presidente Estados_Unidos presidente dar fundamento libertad Estados_Unidos"; S2'="Fundamento libertad ser dar presidente Estados_Unidos". The RBWS representation for these two sentences is shown in figure 7.

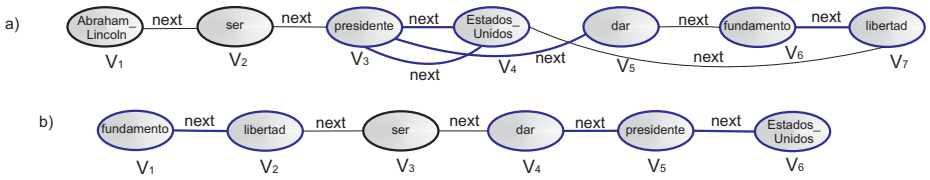



**Fig. 7.** Example of RBWS representation.

In this example, it is possible to see that the common graph represents words that are not only in sequence, but also in not sequence. Moreover, since this repesentation have not a specific topology, then it is possible to discover complex patterns, which represents words that have common neighborhood words.

**Representation based on dependence type (RDT).** In the same way of the previous representation, RDT is a representation without a specific topology. The aim of this representation consist of the construction of a graph based on the analysis of type dependencies. The graph is defined with vertices, where their labels include the lemma of the word (PALLEM), and the grammatical category, denoted by POS. For example, consider the spanish word "comió", its corresponding vertex is labeled with "comer_vb", where "comer" is the lemma and "vb" represents the grammatical category. Finally, the dependency types define the edges of the graph. Clearly, the labels of the vertices are more restrictive at the moment of the mapping process, but is an interesting differentiation

criterion for words with different grammatical categories. This representation is illustred in figure 8.

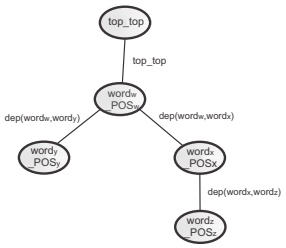



**Fig. 8.** Representation based on dependence type(RDT).

As an example, consider the sentences: S1="Abraham Lincoln fue presidente de los americanos"; S2="Michel Obama será presidenta americana". With the reprocessing phase, the original sentences are transformed to: S1'="Abraham_Lincoln fue presidente americanos", and S2'="Michel_Obama será presidenta americana". After that, the lemmas of each word are calculated: S1"="Abraham_Lincoln ser presidente americano", and S2"="Michel_Obama ser presidente americano". Additionally, all typed-dependency relations are calculated, as in shown in table 1.

| Relation | Token 1 | Token 2 |
|---|---|---|
| top_top (Root Node) | top_top | ser_VS |
| subj (Topic of the sentence) | Abraham_Lincoln | ser_VS |
| att (Verb Attribute) | presidente_NC | ser_VS |
| adj_mod (Modifier adjective ) | americano_AQ | presidente_NC |

**Table 1.** Dependence typed relations of the example RDT.

Finally, based on the RDT representation, the graphs shown in figure 9 are generated. It is evident that this graph representation reduce the possibilities of matching in the data mining phase, and if we analyze a set of sentences with few examples, then the probability to discover interesting patterns is reduced. However, this limitation can be useful in sets with a high number of instances.

## 4 Graph-based data mining phase

Based on the transformation process described above, it is possible to generate a set of graphs from a set of input sentences (texts). Now, it is necessary to explain how it is possible to discover hidden patterns in graphs, called the Graph-Based

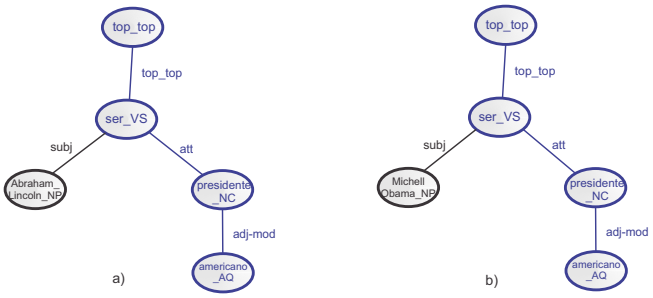



**Fig. 9.** Example of type dependencies (RDT).

Data Mining phase. One of the most important tools in the GBDM area is SUBDUE. In this section is introduced a brief description of this tool[4].

In a general way, this tool admit as input a set of labeled graphs, and as output report common substructures (subgraphs) of the input graphs. For the case of text mining, the output represents common patterns of the sentences, such as common words, characteristics and relations. In this sense, the output can be analyzed in two ways:

- The characteristics reported into the graphs. In this sense, it is possible to extract all words, lemmas, etc. that graphs reports. As consequence of this, interesting patterns can be found, such as: words or lemmas with a high number of instances, n-gramms (words that appear together), words or lemmas with a high probability to appear into the text, but not one beside the other, etc. This information can be used in the PLN area such as characteristics of the texts.
- The structure reported into the graphs. This part is one of the most important results of this approach, because not only is possible to discover common words, lemmas, etc., but also interesting relations between these elements, which represent structured patterns of the text. This part is an important contribution, because many tools in PLN are not capable to report this type of results.

As an example, in this work was performed a set of experiments in a set of documents written in Spanish by different authors: CS Lewis, Darren Shan, J. K. Rowling, Justin Somper, Jules Verne and Rick Riordan (translated into Spanish)[5]. The objective of the task was to identify particular patterns of each author. These results can be used in tasks called authorship attribution. In this experiment, the RDT representation was select, because is capable to represent structure of the input text, which could represent writing styles. One example of interesting patterns discovered in this process is shown in figure 10.

[4] For more info consult http://ailab.wsu.edu/subdue/
[5] These books can be downloaded from http://espanol.free-ebooks.net/

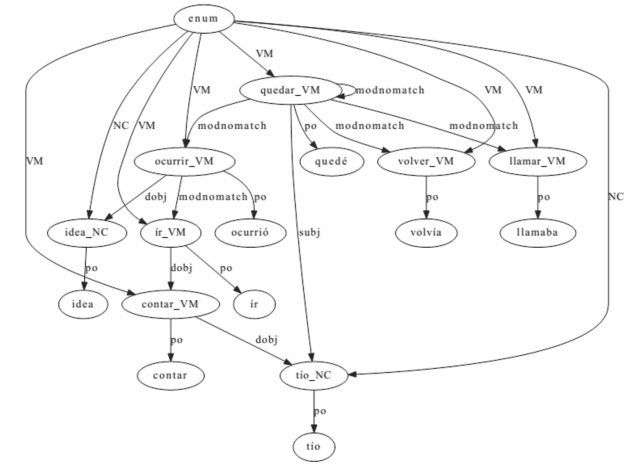



**Fig. 10.** Example of a pattern discovered with the RDP Representation for the authorship attribution task.

## 5 Conclusions

In this paper was proposed a methodology useful to represent texts with labeled graphs, with the aim to discover hidden patterns in text with a graph-based data mining process. The methodology is divided in three main steps: a preprocessing phase, the mapping process (where text are transformed to graphs), and the data mining process.

The preprocessing phase is focused to clean, remove unuseful words, splitte the text in sentences, and assign a set of labels that allow to determine the role of each word in the text. Based on this information, in this work is introduced three graph-based representation useful to map texts to graphs: STR, RBWS, and RDT. Each of them is capable to represent different information of the input text, such as word frequency, relations between words and lemmas, structure of the sentence, etc. Also, is presented a brief discussion of what kind of patterns can be discovered with each representation. Finally, is presented an example where was used this methodology to discover styles of writing in a set of text of different authors.

This paper only expose the idea of the authors of how it is possible to discover knowledge from text with a new alternative method with respect to the techniques used in the PLN area. Then, as future work, we need to prove that this proposal is useful to analyze and discover hidden patterns in real applications. Also, it is necessary to perform a theoretical analysis of these representations, with the aim to prove interesting properties such as time complexity and space complexity costs.

## References

1. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)
2. Harish, B.S., Guru, D.S., Manjunath, S.: Representation and classification of text documents: A brief review. IJCA,Special Issue on RTIPPR (2), 110–119 (2010), published By Foundation of Computer Science
3. Harris, Z.: Distributional structure. Word 10(23), 146–162 (1954)
4. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, MA, USA (1999)
5. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005)

# Computer Assisted Diagnosis of Upper Limbs Using Markerless Mocap and Discriminant Bands of Normal Movement

V. Javier Romano M., Manuel Martin O., and Arturo Olvera L.

Faculty of Computer Science
Autonomous University of Puebla
Av. San Claudio y 14 sur, C.P. 72570, Puebla Pue., Mexico
{victor.romano,mmartin,aolvera}@cs.buap.mx

**Abstract.** Computer vision has become and important tool in several fields of science, like Medicine and Biomechanics. In this work we made an analysis of upper limb movemnts from a kinematic strategy.The information is captured through Markerless Mocap, we obtain the spatial position of the most important joints of the body. We take this numeric values to process and translate them into specific information for the specialist in the physiotherapy field. As a complementary task we create discriminant bands of normal human movement using several curves of behaviour of the local population in an specifc group with similar physical characteristics. An specific support to the diagnosis is the quatitative data obtained from the curve of movement in a standard motion of a patient and the results of the comparisons with the discriminant bands. The system is programmed using Opensource tools with an easy installation without demanding special physical conditions and as a not expensive approach.

**Key words:** Computer vision, markerless MoCap, computer assisted diagnosis.

## 1 Introduction

Movement is one of the vital tasks of human lifestyle, it help us in our working activities, walking, lifting objects, practicing sports and even eating or doing recreational activities. Most of this movement involves the torso and the extremities, these parts of the body are vulnerable of different factors including diseases that can alter joints, muscles, bones or nerves that are involved in some movements, another factor are injuries limiting the range of motion and the third factor is an unavoidable problem every human has: Age, as we grow older the muscles loose their flexibility, the joints their elasticity and we are more exposed to external pathogens factors. As the average of life span is growing bigger we need more specialized tools in detection and treatment of patients with motion problems [1].

Computer vision is growing fastly because of the great improvements in proccesing speeds, increased amount of storage and the increasing level of detail that videocameras now have. We introduce a system that helps physiotherapists when a diagnosis is made as a great potential tool, with a non-expensive cost and easy usage, only using a optical capture device, the tool offers a non-invasive Markerless Mocap that has several advantages because is not as demanding as wearing exoskeletons like the ones used by Xsens [2] or the vest proposed by the Ergonomics' division of the Zaragoza University [6]. Using a capture and a processing subsystem we obtain numerical data about the position of the main body joints through a basic movement routine that streams the depth and RGB video, then with this information a Physiotherapist can use a Results subsystem to compare the motion of a patient with the normal preestimated movement of a certain local population and also can be compared with its own motions of previous medical checkups; the purpose of this is to evaluate the progress of a rehabilitacional therapy, to reinforce a diagnosis of improvement or deterioration of a patient's motion made by the specialist , to route the diagnosis on the path of an specific pathology. The assisted diagnosis is more complete because it includes the experience, intuition and knowledge of the specialist and it is based on numerical values, graphics and videos provided and stored in a computer.

## 2 Methodology

The target of the present project is to help on diagnosis tasks and tracking of rehabilitational therapies, so it should offer simple and organized interfaces. It offers the capture of movement in a kinematic approach of upper limbs, it registers the behavior through time, the diagnosis for each appointment and graphical results that help to interpret the numerical data. Also it stores the depth streams and the three-channels (RGB) video and links them with the medical history to use them at the study of a rehab progress, in a general way the available information, processed, visualized and interpreted helps a better medical practice. This work helps in the inspection of active, passive movements and movements against resistance that are common stages of a kinesio therapy appointment; the most conclusive data obtained using this proccess are the recognition of motion range limitations and the recognition of movement stages with the presence of pain.

The study is targeted to the five basic moves of the shoulder joint that are Abduction-Adduction, Flection- Extension, Hyperextension, Internal-External Rotation and Horizontal Abduction- Adduction, we selected the shoulder because it is the most affected upper body joint in the specialized clinic in physical rehabilitation of the physiotheraphy school in Autonomous University of Puebla. Focusing to the shoulder joint, if we get a limitation of the motion range it could be related to one of these pathologies: Reflex sympathetic dystrophy, recurrent dislocation, subluxations, Periarthritis scapular - humeral ankylosing. If the patient presents pain in a specific range of angles could be related with: Brachial plexus syndrome, Narrowing syndrome (Pain from 60 to 120 on Abduction),

Impingment syndrome (Pain from 80 to 120 on Abduction), Anterior cruciate ligament injury or hemiplejia among many other diseases that could be detected visually [7]. Regarding the capture system we use an optical method that allows depth streams over an electromagnetic, mechanical or an ultrasonic device because it satisfies the accuracy requirements when measuring the position of joints, its approachability, it is easy to install and set up. We are using Markerless mocap over a skeleton model so we can proccess the information in real time.

To develop the system we used the unified proccess model that allow us to divide the entire process in different stages so we can detect errors at early stages creating in an iterative way a more sturdy solution. The analysis of requirements, the first stage, was based on contributions of an expert of physiotheraphy to obtain the medical background, the technical part was stablished by the working group.

## 3 System description

Motion Capture (Mocap) is the process to capture the events of real movement, measures the change of position, the angle of an object in a scene to traduce it on usable mathematics values obtained from the tracking of marks through time [8] [9]. We use optical Mocap so the adquisition of information is only visual, the system uses a depth sensor to capture the information according to the Markerless Mocap type, with this approach we do not require external elements attached to a person or to his or her clothes, this is the most important advantage of markerless mocap because it allows the captured subject to wear its own clothes and without a specific suit or gear, it also helps to move as "natural" as possible. To avoid the auto-oclusion of parts of the body, one of the biggest problems with markerless, we use a starting position in a frontal angle with the hands separated from the rest of the body (Figure 1) and during the lenght of the motion routine there should not be any external obstruction.

Another advantage of Markerless Mocap with optical devices is the close relationship with the Kinematic approach, so we study the movement only as a change of position or angle of any joint without taking into consideration the forces that provoke it [5]. The system is divided into three subsystems, the first

**Fig. 1.** Initial Position to avoid the auto-oclusion

one is Capture subsystem that was implemented with Visual C++ and uses the Openni Library [3] which contains many programmed routines including calibration and basic use of cameras and depth sensors, additionally we use Nite middleware [4] that help us to obtain the tracking of a person and the numerical values of specific joints through time, in addition we use OpenGL and Glui for the visual elements of the interface. The joints that we use are: *Head, Neck, Chest, Left-Right Shoulder, Left-Right Elbow, Left-Righ Hand and Left-Right edge of the Hip*. Connecting each pair of adjacent point we create a line, the set of lines creates a basic skeleton that represents the captured subject, this skeleton-like structure has low cost -computational talking- so it shows the results in real-time.

We chose the Microsoft Kinect Sensor as the main capture device, it allows to store the color stream and depth stream separately and with its 30 Frames Per Second of capture rate, we can obtain the whole motion curve without losing any information. As an additional task of this subsystem we also use a PS3 Eye Camera to show in a television, a mirror-like video for the patient to look himself and an animation of what curve of motion he has to follow. To use this, the patient has to stand in the front of the cameras in the initial position facing them, in Figure 2 we see the assembly of the capture subsystem.

**Fig. 2.** Capture Subsystem Assembly

The second subsystem is the Processing one, here we take the files obtained from the capture, that have the spatial positions of each of the joints mentioned before, to create curves of motion that define the movement through time and we process them to get the final result. The first thing of the process, is reducing the noise that the sensor or any exterior factor had created, to do that we apply the rolling average and the simple exponential smoothing.

$$RollingAverage = \frac{P_M + P_{M-1} ... + P_{M-(n-1)}}{n} . \quad (1)$$

Where each *Pi* represents a joint position in a i-time and *Pi+1* the same point in the next time instant, $n$ represents the amount of points selected to create the average. The second step is to decide how many humps does the curve has,

according to the movement and the selected axis, this is shown in Table 1.

| Movement | X | Y | Z |
|---|---|---|---|
| Abduction | 2 Humps | 1 Hump | No Hump |
| Horizontal Adduction | 1 Hump | No Hump | 2 Humps |
| Flection | No Hump | 1 Hump | 2 Humps |
| Hyperextension | No Hump | 1 Hump | 1 Hump |
| Rotation | No Hump | 1 Hump | 2 Humps |

**Table 1.** Type of movement, depending on the amount of humps

The third step is removing the static parts, the time lapses where there is no change, usually this happens at the starting point or the end of each curve, these timeslots are not important to this work. They are cleaned selecting the point near the maximum and mininum range. The next step is create discriminant bands of normal movement, to do that we need to normalize curves regarding amplitude and time, this normalization needs to be done because the physical differences and times of movement of each subject. The normalization is done using the inverse linear interpolation for each separated half to correct the asimmetry of different upward and downward phases.

$$f_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0). \; . \tag{2}$$

Where $Xi$ represents the two possible spatial positions in which $X$ can be located and *f(x)* is to evaluate a point in a line equation. To create the discriminant bands we calculate the arithmetic mean, then we add the average standard deviation and substract it to create both the upper and lower discriminant bands.

The third subsystem is the Results one, here we calculate the maximum angle of each routine, this defines the range of motion, to calculate the angle we create two vectors both start at shoulder, one ends at the hip and the other ends at the hand, with this two vectors we calculate the angle between them in any moment of the movement (shown in Figure 3) , also we detect if there was unwanted movements of the patient when performing a certain kind of movement, this is estimated comparing the shoulder position with the other joints, this kind of natural movements are the compensation. There is another kind of result in which we compare the subject's curve of motion with the pre- estimated pattern of normal movement to indicate if there is an angle interval where the patient's move is out of the normal range, it calculates the percentage error of that interval and the exact angles of it. All the results include the regular videos, the depth streams and we generate a PDF report file with the results of each study.

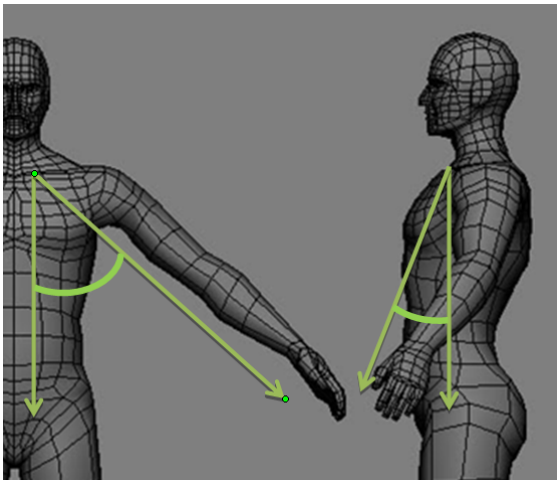


**Fig. 3.** Vector to calculate the range of motion

## 4 Results

We made captures among the local population with fifty three persons performing the five basic movement routines. They all share the same age range between 19 and 32 years old, the 84.90% were men and the remaining 15.01% were women. There was a wide range of weight from 47 kg to 160 kg with an average body mass index of 25.77. The height was from 1.45m to 1.98m. Some of the captured movements did not satisfy the needs to be included in the discriminant bands because of either incomplete movement routines or too much compensation, with the rest we created two discriminant bands for each of the movements. In Figure 4 and 5 we show the discriminant bands for the flection movement, the first one is for Y-Axis with only one hump and the second is for Z-Axis with two different humps. Normalizing the curves of movement was

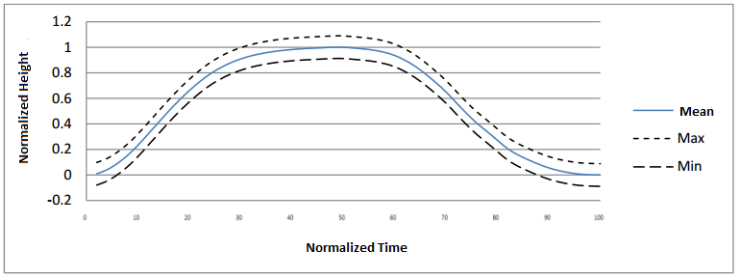



**Fig. 4.** Discriminant band for Y-Axis of Flection-Extension

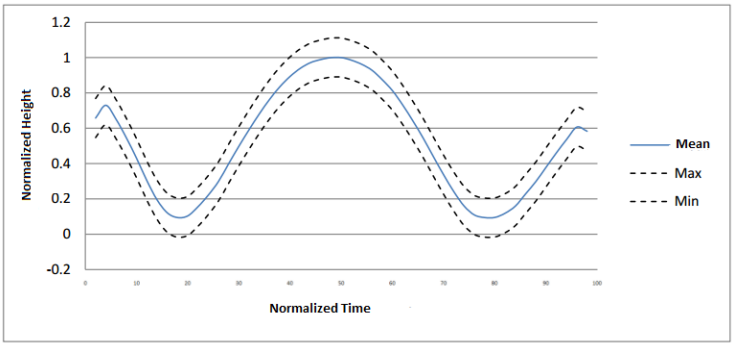



**Fig. 5.** Discriminant band for Z-Axis of Flection-Extension

a necessary task to do because it allowed us to compare people with different heights (in Figure 6 we see the comparison between the tallest and the smallest person of the group, having a difference of more than 50 centimeters, without the normalization we could not compare them, once that the two curves were normalized we can do comparison between them, see Figure 7), by this way we can compare different arms lenghts, weights, body mass index, disposition of motion depending of work or sport activity, different global times on the routine, different upward and downward time and different resting times. We generated

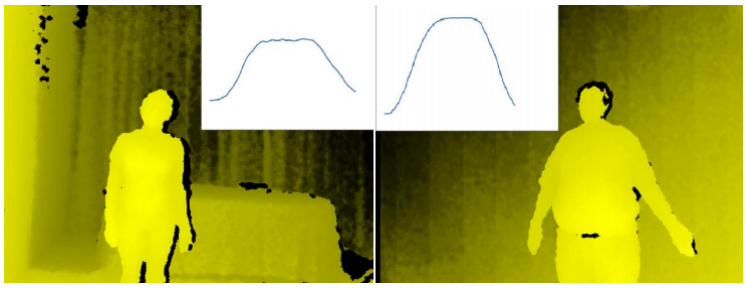


**Fig. 6.** Raw Motion curves of different sized people

results reports for different persons to validate the precision of arm lengths and height, the amount of error was around milimiters, it is important to mention that this precision is obtained when we have controlled illumination in the scene in other case the depth sensor has not this accuracy. We prove the results subsystem with regular data to see whether it fits to normal movement bands (Figure 8) and we also prove curves of movement out the regular pattern (FIgure 9), the result was possitive and they match with the expected medical analysis.

These previous results are done in inspections of active movements (movements made by ourselves), we also prove the system to work with passive inspections

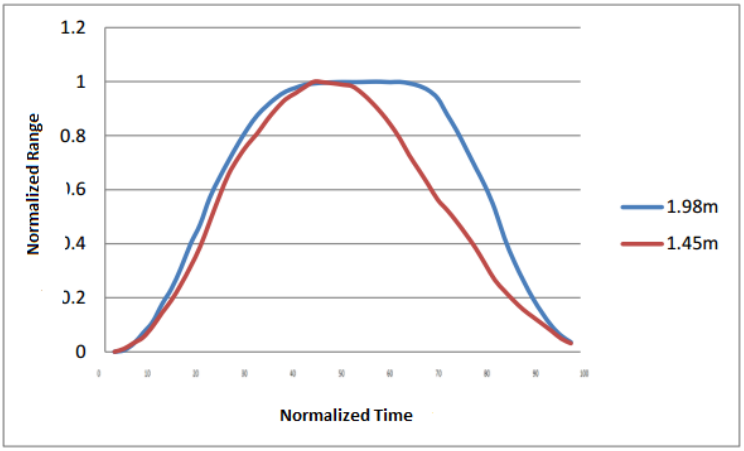



**Fig. 7.** Normalized Motion curves of different sized people

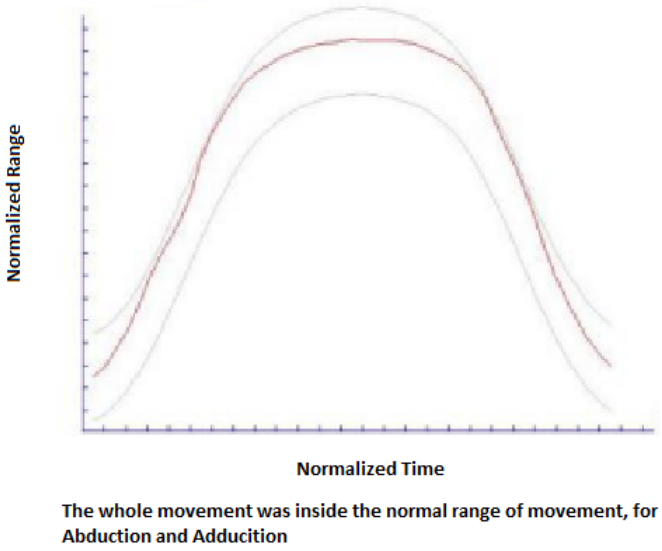



**Fig. 8.** Motion curve that fits the Normal movement pattern

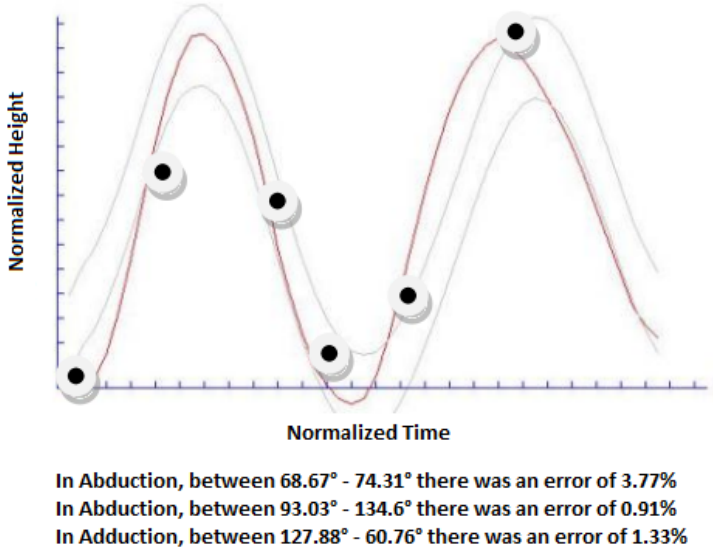



**Fig. 9.** Motion curve that does not fit the Normal movement pattern

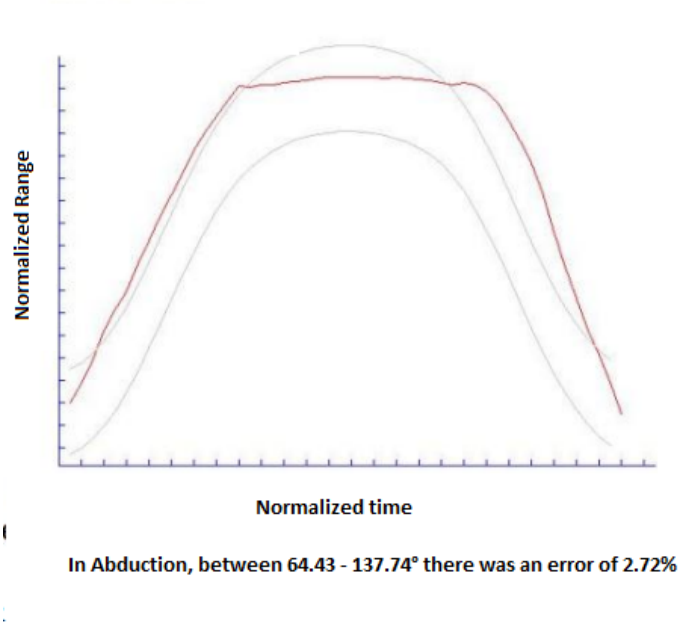



**Fig. 10.** Motion curve of an inspection of movements against resistances (holding a weight)

and movements against resistance. In the passive inspection we find possitive results but the physiotherapist has to be behind the patient, hidden from the depth sensor so the system do not detect the specialist or a part of him as an extension of the subject, if it is done right the system works perfectly. When working in movements against resistance we have better results because if the capture subject holds a weight or a stick (in kinesiotherapy are the most used objects in basic routines) the depth sensor does not recognized as an extension of the subject, so it does not modify the reading and recognition of the joints (Figure 10).

## 5 Conclusion

We show in this work, a system programmed with OpenSource tools, theoretically supported and validated by physiotherapy experts. The system allow us to capture patients with motor problems in a non-demanding, natural way, without the need to use heavy or unconfortable suits. It obtains numerical data, aided with mathematical and statistical processing, to deliver graphics and easy-to-understand values that help when creating a diagnosis. The used tools allow obtaining results in real time, the stored information has a small weight to be stored, the assemblage of the cameras and the software installation are easy and fast. The devices used have a low cost compared with the commercial solutions, so in social security could be more reachable. We create local discriminant bands of normal movements, this is an achievement because there is not enough information of local range of motion or specific values, most of the information used was obtained from different countries not including Latin American people. As future work, the system can be adapted to study different joints of the body. The tool can be used in different hospitals and health centers, with all the gathered information we could create more accurate discriminant bands. Also with enough captures, we can create discriminant classes of specific characteristics, like age, gender or weight, that could help on a more focused study.

## References

1. Freese, J.: Fitness terapeutico. Paidotribo, Barcelona, Espaa, (2006)
2. Xsens, www.xsens.com.
3. OpenNi, www.openni.org.
4. Middleware Nite v2.0, www.openni.org/file/nite.
5. Bouchrika, L. and Nixon, M.S.: Gait recognition by dynamic cues, Pattern Recognition ICPR 2008, pp 1-4, Tampa Florida, U.S., (2008)
6. Zaragoza University.: Research and Development in Ergonomics, http://i3a.unizar.es/en/content/idergo.
7. Andriacchi, T and Alexander, J.: Studies of human locomotion: past, present and Future, Deparments of Mechanical Engineering and Functional Restoration, Division of Biomechanical Engineering, Stanford University, (2000)
8. Menache, A.: Understanding motion capture for computer and videogames, Academic Press, San Diego, U.S. (2000)
9. Dyer, S.M.: Motion capture white paper, http://reality.sgi.com/employees/jamsb/mocap/.

# Towards a Surrogate-Assisted Multi-Objective Full Model Selection

Alejandro Rosales-Pérez[1], Jesus A. Gonzalez[1], Carlos A. Reyes-Garcia[1], and Carlos A. Coello Coello[2]

[1] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Computer Science Department
Luis Enrique Erro No 1, Sta. Ma. Tonantzintla, C.P. 72840, Puebla Pue., Mexico.
{arosales,jagonzalez}@inaoep.mx
[2] Centro de Investigación y de Estudios Avanzados del IPN (CINVESTAV-IPN)
Computer Science Department
San Pedro Zacatenco, Mexico City, Mexico
ccoello@cs.cinvestav.mx

**Abstract.** This research approaches the full model selection problem. The full model selection problem is defined as a method in which, given a pool of pre-processing methods, feature selection and learning algorithms, to choose from, we select a combination of them, together with their hyper-parameters,in such a way, that we can provide the "best" generalization performance on a given dataset. We propose to face this as a multi-objective optimization problem, where the classification-error and the model complexity are defined as the objectives to be minimized. We propose to use a surrogate-assisted multi-objective evolutionary algorithm approach to explore the models space. Our proposal derives from the fact that estimating the values of the objective function could be computationally expensive. Therefore, by using surrogate-assisted optimization we expect to reduce the number of full models that should be trained and tested so that we can reduce the total number of fitness function evaluations, without degrading, in a significant manner, the quality of the models. Our preliminary results give evidence of the validity of our proposed approach.

**Key words:** Full model selection, multi-objective optimization, ensemble methods, VC dimension.

## 1 Introduction

Classification is a mainstream in supervised learning. A large number of learning algorithms have been proposed so far, with the aim of constructing a classification model. However, there does not exist a single learning algorithm that is the best for all problems; this is sometimes known as the **no free lunch theorem** [23]. In addition to this lack of a universal best algorithm, the performance of many of them highly depends on the fine-tuning of a set of hyper-parameters. This raises the issue of model selection.

It is noteworthy that besides the learning algorithm, there exist methods for pre-processing the data and for feature selection, which could help to improve the model performance. For instance, $k$-nearest neighbor method is not robust to the way the features are scaled. Therefore, it could result beneficial if the data are first normalized or standardized. Moreover, it could also be favorable if the irrelevant/redundant features are previously filtered. Based on the above, one faces the issue of choosing a combination of these methods together with the hyper-parameters that improve the performance of the model. This is a problem known as *full model selection* [11].

In the literature, there are several studies on the model selection problem. Some of these have approached it as an optimization problem. They could be differentiated in two essential aspects: the criterion used and the search engine adopted for this task. Regarding the first aspect, this problem has been tackled both as a single criterion optimization problem and as a multiple criteria optimization problem. The single criterion approaches typically utilize the well-known $k$-fold cross validation to estimate the model performance [1,3–5,11]. On the other hand, multiple criteria approaches consider an estimation of the model performance and a measure of its complexity [2, 20]. Others have considered to minimize the error rates on positive and negative classes [6,14], estimates of the bias-variance model [18,19], or different estimates of the model performance [13].

Concerning the second aspect, authors have investigated the use of grid search [4, 21], gradient-based methods [1, 5], and bio-inspired meta-heuristics such as evolutionary algorithms [6,13,14,18–20], artificial immune systems [2], or particle swarm optimizers [3,11]. Grid search is the simplest approach to adjust the hyper-parameters values. Under this strategy, each combination of hyper-parameters is tested, which makes this approach suitable to adjust only a few number of hyper-parameters. In full model selection, several hyper-parameters need to be adjusted simultaneously, which could be unsuitable for this approach.

In spite of the fact that gradient-based methods are more efficient and they have been successfully applied to model selection problems, they still have several shortcomings. For instance, the objective function must be differentiable with respect to the hyper-parameters. Furthermore, the effectiveness of these kinds of methods highly depends on the initial search point. This makes that these methods are susceptible to getting trapped in a local optimal solution.

Evolutionary algorithms have also been used in previous studies for model selection. These kinds of algorithms could be less susceptible to local optimal solutions than gradient-based methods. Although they could be cheaper than grid search methods, their computational cost could still be high.

An alternative approach formulates the model selection problem as a supervised learning one by constructing a meta-model, which is in charge of making the suggestion for models. Recent studies have combined the ideas of treating model selection as supervised learning and optimization problems [12,15,17]. The main idea under these hybrid approaches is to use the meta-model for obtaining suggestions of potential models to be used as initial points in the search step. However, the quality of the meta-model depends on the quality of the samples

as well as on the number of problems which are learned and could be limited. These shortcomings could affect the convergence in the optimization step.

In spite of the considerable number of studies on model selection, most of them have focused on single model selection (i.e., the learning algorithm is fixed a priori and the task is performing the selection of its hyper-parameters), which could not be the most suitable for a particular problem. The studies on full model selection are still scarce, and they have been formulated as a single criterion optimization problem. Nevertheless, the advantages of multiple criteria over a single criterion on hyper-parameters tuning have been pointed out by several authors [6, 13].

Inspired by the above, in this research we propose to tackle both the full model selection problem as a multi-objective one (i.e., to consider multiple criteria) and the computational cost in this task. The latter is addressed by using surrogate-assisted optimization. The main motivation of this research is, precisely, to design an algorithm to perform a multi-objective full model selection emphasizing its efficiency, measured in terms of number of evaluations performed. Our working hypothesis is that, by minimizing simultaneously the error and complexity of a full model through surrogate-assisted optimization, it will be possible to obtain, in an efficient way, accurate full models that satisfy a good trade-off between the considered criteria. The estimation of the complexity should be generic in order to make it feasible to the full model selection problem, which is one of the main challenges in this research. The main contribution of this research is a general model selection framework, whose formulation makes it applicable to any learning algorithm and, in consequence, to the full model selection problem. Additional contributions are the following: (i) the multi-objective formulation of the full model selection problem (i.e., to choose a combination of pre-processing, feature selection methods, and learning algorithm together with its hyper-parameters); (ii) the hybridization with surrogate-assisted optimization to reduce the number of objective functions evaluations; and (iii) since the outcome of the multi-objective optimization is a set of solutions that satisfy a good trade-off between the objectives, the strategies to address the final model construction from such set would also be an additional contribution.

The remainder of this paper is as follows: Section 2 describes the basic concepts related to evolutionary multi-objective optimization. Section 3 describes the proposed research methodology. Section 4 shows the preliminary results of our research to give evidence of the feasibility of our proposal. Finally, Section 5 details some conclusions and indicate paths of future research.

## 2 Evolutionary Multi-Objective Optimization

A multi-objective optimization problem (MOP) is stated as follows [7]:

$$\begin{aligned} &\text{minimize } \mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_l(\mathbf{x})]^T \\ &\text{subject to } \mathbf{x} \in \mathcal{X} \end{aligned} \tag{1}$$

where $\mathbf{x} = [x_1, \ldots, x_n] \in \mathbb{R}^n$ is a vector of decision variables, $f_i(\mathbf{x})$, $i = 1, \ldots, l$, are the $l$-objective functions, and $\mathcal{X}$ is the set of feasible solutions.

In a MOP, the objectives could be in conflict. In such cases, the notion of optimum refers to finding good trade-offs among the objectives. The most accepted notion of optimality is the one proposed by Pareto. To describe the concept of Pareto optimality, we will introduce the following definitions:

**Definition 1. Pareto dominance:** *A solution* $\mathbf{x^{(1)}}$ **dominates** *a solution* $\mathbf{x^{(2)}}$ *(denoted by* $\mathbf{x^{(1)}} \preceq \mathbf{x^{(2)}}$*) iff* $\mathbf{x^{(1)}}$ *is better than* $\mathbf{x^{(2)}}$ *at least in one objective and it is not worse in the rest.*

**Definition 2. Pareto optimality:** *A solution* $\mathbf{x}^* \in \mathcal{X}$ *is a* **Pareto Optimal** *if there does not exist another solution* $\mathbf{x}' \in \mathcal{X}$ *such that* $\mathbf{x}' \preceq \mathbf{x}^*$.

The Pareto optimal definition does not produce a single solution, but a set of them, which represent the possible trade-offs among the different objectives. The set of trade-off solutions (in decision variable space) is known as **Pareto optimal set**.

**Definition 3. Pareto optimal set:** *The Pareto optimal set (PS) is defined as:*

$$PS = \{\mathbf{x} \in \mathcal{X} \mid \mathbf{x} \textit{ is a Pareto optimal solution}\}$$

The objective function values corresponding to the elements of the Pareto optimal set constitute the so-called **Pareto front**. Formally,

**Definition 4. Pareto front:** *The Pareto front (PF) is defined as:*

$$PF = \{\mathbf{f}(\mathbf{x}) \mid \mathbf{x} \in PS\}$$

Evolutionary algorithms have gained popularity to solve MOPs, mainly because they can obtain several elements of the Pareto optimal set in a single run. Furthermore, they are less sensitive to the shape and continuity of the Pareto front than mathematical programming techniques. In the literature, a large number of multi-objective evolutionary algorithms (MOEAs) have been reported so far. NSGA-II [9], PESA-II [8], and MOEA/D [24] are some of these MOEAs. A comprehensible review of MOEAs can be found in [7].

In the full model selection problem, both pre-processing, feature selection methods, and the learning algorithm together with its hyper-parameters have to be chosen, resulting in a vast search space. Furthermore, two criteria should be simultaneously optimized (the model performance and the model complexity). Thus, stochastic search techniques, such as MOEAs, are well suited for this. In spite of the MOEAs' advantages, they have to perform a relatively high number of fitness function evaluations to get a reasonable approximation to the Pareto front. This could be a shortcoming in the problem that we face, since the computation of the objective could require to train and to test a model a number of times. To overcome this handicap, in this research we propose to study the

surrogate-assisted multi-objective evolutionary optimization to address this issue. A surrogate is a cheaper approximation to the fitness function and it is used to approximate the fitness value of a given model. By using surrogate-assisted optimization, we expect to reduce the number of solutions evaluated with the fitness function and, in this manner, to reduce the computational cost of this task. Next, we explain the proposed research methodology.

## 3 Research Methodology

1. **Design an algorithm for multi-objective full model selection.** This stage of the research involves a review of the literature in order to find how to estimate the model complexity in a general fashion to any learning algorithm. This stage involves also analyzing the advantages and disadvantages of each approach and choosing one according to the previous analysis. It also involves the formulation of the full model selection problem as a multi-objective one, which implies the definition of how solutions are represented into the MOEA, the operators adopted to evolve the models, and the strategy for exploring the models space. The integration of these in an algorithm and its evaluation are also tasks in this stage. An analysis of the performance is used to propose improvements to the algorithm.
2. **Design a strategy for decision making in multi-objective full model selection.** This stage is mainly focused on analyzing the non-dominated front in order to determine what solutions should be chosen as the final classification model. We propose to explore two alternatives: the first one consists on choosing a single model from those generated during the optimization step. The second one considers an ensemble of models. In the first one, it is necessary to identify the regions on the Pareto front so as to find in which region is located the model with the best performance on unseen data. On the other hand, the second approach involves to study strategies to choose the subset of accurate and diverse models to be used in the ensemble. Both approaches are studied in this stage, and their advantages and disadvantages are also analyzed. The improvements are based on the results of the performed analysis. In case of being necessary, modifications to stage one are also performed.
3. **Integration with a surrogate-assisted optimization approach.** This stage includes the hybridization of the MOEA with a surrogate, which is used to approximate the fitness values of the models. Strategies to make such hybridization and an interaction with the expensive fitness functions are proposed in this stage. The integration of the proposed scheme with the multi-objective full model selection approach is also considered in this stage. We evaluate the performance of the proposed algorithm in terms of its accuracy-performance and the number of fitness function evaluated. Improvements to this stage are based on the performed evaluation. If necessary, modifications to the previous stages are also considered.

By following this methodology, we expect to achieve in a successful manner the goals of this research. The next section presents the preliminary results reached to date.

## 4 Results Achieved

In this section, we describe the preliminary results of our research. First, we present a brief description of the proposed method to deal with the model selection problem. Next, we present experimental results together with a statistical analysis.

### 4.1 Towards a Multi-Objective Full Model Selection

Following the proposed research methodology, we have formulated the model selection problem as a multi-objective optimization one. We consider different kinds of learning algorithms together with their hyper-parameters. For doing so, we first need to estimate the model complexity in a general fashion to any learning algorithm. We studied two approaches to do this. The first one is the model variance, due to the fact that a high complex model has a high variance. The second one is the VC-dimension, a measure of the capacity of the model, which is also related to the model complexity. The studies related to the variance as a measure of the model complexity are reported in [18,19]. Regarding the VC-dimension as a measure of the model complexity, we have proposed an approach for multi-objective model type selection (i.e., both a learning algorithm and its hyper-parameters are chosen). We compared three MOEAs widely used in the literature. These MOEAs are NSGA-II [9], PESA-II [8], and MOEA/D [24]. In the comparison, these algorithms reached, on average, a very similar performance in the problem at hand. However, the computational cost of MOEA/D was lower than the others. For this reason, we adopted MOEA/D.

In evolutionary algorithms, the solutions must be encoded in individuals. We propose to encode the solutions in a $D$-dimensional vector, where $D = 7$, as follows:

$$\mathbf{x}^i = \left[x^i_m, x^i_{hp_1}, \ldots, x^i_{hp_{D-1}}\right] \tag{2}$$

where $x^i_m$ controls the learning algorithm, and $[x^i_{hp_1}, \ldots, x^i_{hp_{D-1}}]$ represents the hyper-parameters for the learning algorithm.

The fitness function that we propose to estimate the merit of each model for a given dataset is as follows:

$$\begin{aligned} err &= \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(y_i, y_i^*\right) \\ complexity &= \underset{h}{\operatorname{argmin}} \sum_{i=1}^{k} \left[\overline{\xi}\left(n_i\right) - \Phi\left(n_i/h\right)\right]^2 \end{aligned} \tag{3}$$

where $N$ is the number of samples in the training set, $y_i$ is the class label, $y_i^*$ is the class predicted by the model, $\mathcal{L}(y_i, y_i^*)$ is a loss function, $\xi(n_i)$ is the experimental maximum deviation error rate of two observed independent labeled data sets, and $\Phi(n_i/h)$ is the expectation of the largest deviation error between two sets (we refer to [22] for details about complexity estimation). We used the 0/1 loss function because it is well suited for classification tasks.

These definitions correspond to the first step of our research methodology. Considering the second step, we have proposed three strategies for constructing a final classification model from those that are in the non-dominated set. The first strategy consists in choosing a single solution from the non-dominated front. The second and third strategies are based on the idea of combining the multiple models in the non-dominated front in an ensemble. For the first strategy, we analyzed the performance on test sets of each solution in the non-dominated front. We empirically noted that the solutions that are in the knee of the curve have the best generalization. We also noted that this solution in most cases corresponds to the one closest to the (0,0) point. Therefore, the objectives are first normalized and then the Euclidean distance is computed between each point and the (0,0) point. The one with the minimum distance is chosen. In Figure 1, the solution that was selected with this strategy is circled.

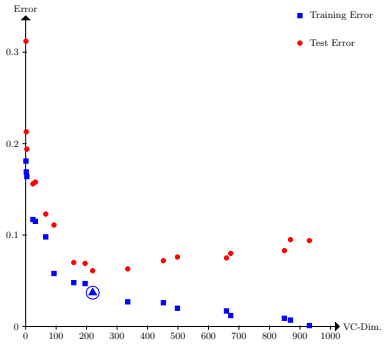



Fig. 1: Behavior of non-dominated solutions on training samples and test samples

The second strategy consists in considering all solutions in the non-dominated front and combining them in an ensemble. The final output of the ensemble is the weighted linear aggregation of the individuals predictions given by each model. The weight of each model is assigned based on the distance from such model to the (0,0) point, in objective function space. The third strategy considers to choose a subset of models in the non-dominated set taking into consideration the diversity among them. Next, we present the experimental results with our approach and the comparative study between these three strategies.

### 4.2 Experimental Results

We performed experiments using the IDA benchmark repository datasets. This benchmark has 13 datasets of binary classification problems. Table 1 shows some characteristics of these datasets. These datasets were previously pre-processed by [16], in which each data set was divided in 100 partitions for training and test (20 for the cases of image and splice data sets). We performed the model selection independently for each replication of each dataset.

Table 1: Details of the data sets used in our experiments.

| ID | Data set | Feat. | Training Samples | Testing Samples | Replications |
|---|---|---|---|---|---|
| 1 | Banana | 2 | 400 | 4900 | 100 |
| 2 | Breast Cancer | 9 | 200 | 77 | 100 |
| 3 | Diabetes | 8 | 468 | 300 | 100 |
| 4 | Flare Solar | 9 | 666 | 400 | 100 |
| 5 | German | 20 | 700 | 300 | 100 |
| 6 | Heart | 13 | 170 | 100 | 100 |
| 7 | Image | 20 | 1300 | 1010 | 20 |
| 8 | Ringnorm | 20 | 400 | 7000 | 100 |
| 9 | Splice | 60 | 1000 | 2175 | 20 |
| 10 | Thyroid | 5 | 140 | 75 | 100 |
| 11 | Titanic | 3 | 150 | 2051 | 100 |
| 12 | Twonorm | 20 | 400 | 7000 | 100 |
| 13 | Waveform | 21 | 400 | 4600 | 100 |

The performance of the proposed model selection method is assessed by means of the error rate attained on each data set. We compare the three strategies for the final classification model construction, and the best one is compared with PSMS, a full model selection method reported in the literature.

Table 2 shows the average error rates and standard error reached by our three strategies: single model selection (MOMTS-S1), ensemble of the whole non-dominated front (MOMTS-S2), and the ensemble of some solutions in the non-dominated front taking into consideration the diversity (MOMTS-S3). It also shows the performance reached by PSMS. From this table, one could note that the best results among the three strategies is reached by MOMTS-S2, the ensemble approach that combines all solutions in the non-dominated front. This is not entirely surprising, since the benefits of using the ensemble method for improving model performance are well known. For assessing the statistical difference between the three approaches for the final model construction over the different data sets, Demšar [10] recommends Friedman's test for comparing multiple classifiers over multiple data sets. This test is performed with a 95% of confidence, and the Nemenyi test as the post hoc test. According to these tests, the ensemble of the whole front approach is found to be statistically superior to the others.

Comparing with PSMS, we noted that MOMTS-S2 gets better performance in 12 out of 13 data sets. This shows the advantages of using a multi-objective

Table 2: Results obtained by the proposed approach, and those obtained by PSMS. The best result for each data set is shown in **boldface**.

| ID | PSMS [11] | MOMTS-S1 | MOMTS-S2 | MOMTS-S3 |
|---|---|---|---|---|
| 1 | $11.08 \pm 0.083$ | $14.34 \pm 0.105$ | $\mathbf{10.48 \pm 0.046}$ | $12.91 \pm 0.160$ |
| 2 | $33.01 \pm 0.658$ | $29.89 \pm 0.736$ | $\mathbf{25.61 \pm 0.593}$ | $27.82 \pm 0.676$ |
| 3 | $27.06 \pm 0.259$ | $28.34 \pm 0.318$ | $\mathbf{23.08 \pm 0.174}$ | $25.66 \pm 0.214$ |
| 4 | $34.81 \pm 0.173$ | $34.90 \pm 0.224$ | $34.59 \pm 0.189$ | $\mathbf{34.52 \pm 0.214}$ |
| 5 | $30.10 \pm 0.720$ | $28.30 \pm 0.274$ | $\mathbf{23.67 \pm 0.224}$ | $25.89 \pm 0.218$ |
| 6 | $20.69 \pm 0.634$ | $23.14 \pm 0.542$ | $\mathbf{16.48 \pm 0.241}$ | $18.75 \pm 0.351$ |
| 7 | $2.90 \pm 0.112$ | $3.79 \pm 0.226$ | $\mathbf{2.24 \pm 0.123}$ | $3.03 \pm 0.246$ |
| 8 | $7.98 \pm 0.660$ | $2.66 \pm 0.079$ | $\mathbf{2.49 \pm 0.074}$ | $3.02 \pm 0.164$ |
| 9 | $14.63 \pm 0.324$ | $7.43 \pm 0.373$ | $\mathbf{4.84 \pm 0.156}$ | $6.71 \pm 0.269$ |
| 10 | $4.32 \pm 0.235$ | $6.48 \pm 0.350$ | $\mathbf{4.00 \pm 0.194}$ | $6.11 \pm 0.347$ |
| 11 | $24.18 \pm 0.193$ | $26.53 \pm 0.127$ | $\mathbf{22.08 \pm 0.085}$ | $22.22 \pm 0.100$ |
| 12 | $\mathbf{3.09 \pm 0.127}$ | $5.21 \pm 0.555$ | $3.73 \pm 0.179$ | $5.70 \pm 0.679$ |
| 13 | $12.80 \pm 0.325$ | $11.34 \pm 0.180$ | $\mathbf{9.93 \pm 0.043}$ | $10.95 \pm 0.256$ |

approach over single-objective approaches for tackling the model selection problem. In order to statistically assess the performance of these two approaches over the suite of 13 benchmark data sets, the Wilcoxon signed rank test with a 95% of confidence was used. According to this test, MOMTS-S2 is statistically better than PSMS.

## 5 Conclusions

In this paper, we presented our research proposal on the full model selection problem. We proposed to approach it as a multi-objective optimization one. We have a general way for estimating experimentally the model complexity, by using the VC-dimension. Our formulation showed the following advantages: (i) the experimental way for measuring the VC dimension allows us to consider different learning algorithms in a general framework, and also allows making the method extensible to the full model selection problem; (ii) our proposal showed a competitive performance over different benchmark data sets, which makes it applicable to problems from diverse domains; and (iii) the multiple non-dominated solutions obtained through the multi-objective formulation facilitates its extension to ensembles of models.

The VC dimension is experimentally estimated, which implies that a model must be trained and tested a number of times. This makes it computationally expensive. As part of our future work, we want to explore the surrogate-assisted evolutionary computation to reduce the computational cost. We also want to extend our current approach to the full model selection problem, i.e., considering feature selection and data pre-processing into the model selection process. Studying more effective ways for constructing an ensemble (possibly) by using a second level of optimization would also be another interesting direction for this research work.

## References


1. Ayat, N., Cheriet, M., Suen, C.: Automatic model selection for the optimization of SVM kernels. Pattern Recogn 38(10), 1733 – 1745 (2005)
2. Aydin, I., Karakose, M., Akin, E.: A multi-objective artificial immune algorithm for parameter optimization in support vector machine. Appl Soft Comput 11(1), 120 – 129 (2011)
3. Bao, Y., Hu, Z., Xiong, T.: A PSO and pattern search based memetic algorithm for SVMs parameters optimization. Neurocomputing 117(0), 98 – 106 (2013)
4. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2(3), 27:1–27:27 (2011)
5. Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S.: Choosing multiple parameters for support vector machines. Mach Learn 46(1-3), 131–159 (2002)
6. Chatelain, C., Adam, S., Lecourtier, Y., Heutte, L., Paquet, T.: A multi-model selection framework for unknown and/or evolutive misclassification cost problems. Pattern Recogn 43(3), 815 – 823 (2010)
7. Coello Coello, C.A., Lamont, G.B., Veldhuizen, D.A.V.: Evolutionary Algorithms for Solving Multi-Objective Problems. Genetic and Evolutionary Computation, Springer US, 2 edn. (2007)
8. Corne, D.W., Jerram, N.R., Knowles, J.D., Oates, M.J., J, M.: Pesa-ii: Region-based selection in evolutionary multiobjective optimization. In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO2001. pp. 283–290. Morgan Kaufmann Publishers (2001)
9. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE T Evol Comput 6(2), 182–197 (2002)
10. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7, 1–30 (2006)
11. Escalante, H.J., Montes, M., Sucar, L.E.: Particle swarm model selection. J Mach Learn Res 10, 405–440 (2009)
12. Gomes, T.A.F., Prudencio, R.B.C., Soares, C., Rossi, A.L.D., Carvalho, A.: Combining meta-learning and search techniques to svm parameter selection. In: Proceedings of the $11^{th}$ Brazilian Symposium on Neural Networks. pp. 79–84 (2010)
13. Gorissen, D., Dhaene, T., Turck, F.D.: Evolutionary model type selection for global surrogate modeling. J Mach Learn Res 10, 2039–2078 (2009)
14. Li, W., Liu, L., Gong, W.: Multi-objective uniform design as a svm model selection tool for face recognition. Expert Syst Appl 38(6), 6689 – 6695 (2011)
15. Miranda, P.B.C., Prudencio, R.B.C., Carvalho, A.C.P.L.F., Soares, C.: Multi-objective optimization and meta-learning for svm parameter selection. In: Neural Networks (IJCNN), The 2012 International Joint Conference on. pp. 1–8 (2012)
16. Rätsch, G., Onoda, T., Müller, K.R.: Soft margins for adaboost. Mach Learn 42(3), 287–320 (2001)
17. Reif, M., Shafait, F., Dengel, A.: Meta-learning for evolutionary parameter optimization of classifiers. Mach Learn 87(3), 357–380 (2012)
18. Rosales-Pérez, A., Escalante, H.J., Gonzalez, J.A., Reyes, C.A.: Bias and variance optimization for svms model selection. In: The $26^{th}$ FLAIRS Conference (2013)
19. Rosales-Pérez, A., Escalante, H.J., Gonzalez, J.A., Reyes-Garcia, C.A., Coello Coello, C.A.: Bias and variance multi-objective optimization for support vector machines model selection. In: Sanches, J.a.M., Micó, L., Cardoso, J.S. (eds.) Pattern Recognition and Image Analysis. LNCS, vol. 7887, pp. 108–116. Springer Berlin Heidelberg (2013)

20. Suttorp, T., Igel, C.: Multi-objective optimization of support vector machines. In: Jin, Y. (ed.) Multi-Objective Machine Learning, Studies in Computational Intelligence, vol. 16, pp. 199–220. Springer Berlin / Heidelberg (2006)
21. Valentini, G., Dietterich, T.G.: Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. J Mach Learn Res 5, 725–775 (2004)
22. Vapnik, V., Levin, E., Le Cun, Y.: Measuring the VC-dimension of a learning machine. Neural Comput 6(5), 851–876 (1994)
23. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. Neural Comput 8(7), 1341–1390 (1996)
24. Zhang, Q., Li, H.: MOEA/D: A multiobjective evolutionary algorithm based on decomposition. IEEE T Evol Comput 11(6), 712–731 (2007)

# Reviewing committee

Vicente Alarcón
Maya Carrillo
Sergio Coria
Anilu Franco
René García
Félix González
Raudel Hernández
Manuel Lazo
Manuel Martín
Iván Olmos
Eber Orozco
Airel Pérez
David Pinto
Ansel Rodríguez
Roberto Rosas
Guillermo Sánchez

Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27, Centro Histórico, México, D.F.
Mayo de 2014.
Printing 500 / *Edición 500 ejemplares*.