Segmenting supervised activities in a video sequence based on handling of artifacts towards intelligent systems

Francisco E. Martínez-Pérez[1], Héctor G. Pérez-González[1], José A. González-Fraga[2], Juan Carlos Cuevas-Tello[1] and Sandra Edith Nava-Muñoz[1]

[1] Facultad de Ingeniería, Universidad Autónoma de San Luis Potosí, Av. Manuel Nava No. 8, Zona Universitaria, San Luis Potosí, S.L.P. 78290, México
eduardo.perez@uaslp.mx,hectorgerardo@acm.org,cuevas@uaslp.mx,senavam@uaslp.mx,
[2] Facultad de Ciencias, Universidad Autónoma de Baja California, Carretera Transpeninsular Ensenada-Tijuana No. 3917, Colonia Playitas, Ensenada, B.C. 022860, México
angel_fraga@uabc.edu.mx
Paper received on 11/22/13, Accepted on 01/19/14.

Abstract. Nowadays intelligent systems community has conducted research on human activity recognition. For example, in a healthcare environment, we would like to know what activities (feeding, blood pressure, hygiene and medication) are performed by a caregiver given a video sequence (recorded by a surveillance system). Specifically, it is complicated to infer those activities that are performed using one or several artifacts at different times, so the activity inference and video segmentation are complex tasks. Additionally, it is desirable to perform the video segmentation in an automatic fashion. Therefore, in this paper we present an intelligent system for video segmentation. We present an example in a realistic environment in which the analysis of video sequences per day was reduced by using video segmentation.

Keywords: video segmentation, activity recognition, roaming beat.

1 Introduction

Generally, an intelligent system is composed by several modules with the aim to support robustness and computational complexity. The system relies on signal processing and machine learning techniques [1]. It first applies preprocessing to remove noise from the signals and segment them. Next, the system extracts signal features to enhance the characteristics unique to each activity and to reduce data dimensionality, and then it uses classifiers to map these features to discrete activity or context classes. The goal of an intelligent system for human activity recognition consists of automatically analyzing and classifying activities with information gathered from different capture sources like video cameras or other sensors. However, human activity recognition is complicated. It is due to

Mireya García-Vázquez, Grigori Sidorov, Sunil Kumar (Eds.).
Advances in computing science, control and communications.
Research in Computing Science 69, 2014, pp. 108- 119

there exists a lot of way to perform an activity. For example, humans perform several physical activities such as walking, running and so on. In this kind of activity to get signal processing, some authors have reported physical recognition results using a set of templates or people silhouettes corresponding to each activity [2, 3]. A related work in which physical activities are analyzed, can be found in [4]. Other works of different kind are focused on activity classification, for example some works involve physical artifacts as shown in [5]. These authors show activity classification using several kinds of sensors. Moreover, the authors classify activities, that human performs, on three classes such as sequential, interleave and concurrent activities. For example, when someone brushes his teeth and continues to another activity like arranging his hair, this corresponds to a sequential activity. Interleaved activities correspond, for instance, to the case in which someone is eating dinner and then answers the phone, at the end both are performed at the same time. Finally, two activities are concurrent when these activities start at the same time, i.e. eating and watching TV, both activities carried out always at the same time.

For example, a comparison of inferring activities considering sequential, interleave and concurrent is done in [6]. They propose some methods to perform inference, comparing four techniques: the hidden Markov model (HMM), the conditional random field (CRF), a variant of CRF and emerging patterns (EP). They reported acceptable results using EP on inferring physical activities. Other work is presented in [5] in which they show an effectiveness rate of 66.13 percent in the inference of sequential, interleaved and concurrent activities. Its main disadvantage is that authors take into account sequential process and their validations require a controlled space. Therefore it is necessary to take into account an understanding such as presented in [7], which comprises three elements that correspond to a person, the use of devices or tools, and the purpose of the activity. The activities are never performed alone, these are interleaved with other activities with the same goals, or instruments used in producing the intended activity. In the development of a specific activity, the motion artifact may be viewed as a change of activity or a change in the purpose of the actions, or a grouping of actions. Therefore, it is established that the actions of the activities should never be labeled as unknown, according to [8], they create a list of ordered actions, the first one being the most likely.

For aforementioned reasons, activities must be seen as multiple flows of actions that are performed on different times, so these activities do not have a specific behavior in order to recognize them. Therefore, it is necessary to take into account all variants of actions when activity is started, performed and finished. In this sense, to solve this problematic, we use a concept and its methodology called Roaming Beat (RB) [9] (see section 2.1).

On the other hand, video sequence segmentation is a problem that activity recognition has tackled. For example, [10–13] show the work in which a video sequence segmentation is implemented using image processing. Background technique is used in [10]. Other authors treat video segmentation as non-local spatiotemporal structure using regions technique [11]. In the same direction, in [12] spatiotem-

poral technique is used; these authors show an algorithm for video segmentation using motion cues from past and future frames. Another way to make a video segmentation is presented by [14]. The process performed by these authors is based on a set of labels in the video sequence. These allow them to automatically discover a set of relevant tags and extract a set of key words correlated with an activity. However the proposed technique depends on the activities being related to video labels. Such labels consist of contextual information describing the activity of the video, i.e. the video was previously analyzed and labeled. Additionally activity recognition is not obtained in an automatic way. Unfortunately, all the video sequences in real life are not automatically labeled. Furthermore, these sequences involve a multitude of people both static and in movement. This allows us to give meaning to an activity as these movements are signals for the occurrence of an event, forming a set of activity events.
The main contribution of this paper is the development of an intelligent system for video segmentation. We describe the process to store an activity in a knowledge base and the process to recover an activity from the knowledge base. This paper is organized as follows: Section 2 introduces our methodology for activity segmentation in a video sequences. In Section 3, the results showing an example implemented in a healthcare environment. Finally section 4 provides our conclusions and future work.

2 Labeling a video sequence and reducing video analysis

The artifacts play an important role for activity inference, because the artifacts are a trigger within an event in the setting. It is caused through activity performance because an activity is mediated by one or more instruments and is directed toward a certain artifact[7].

2.1 Labeling a video sequence through Roaming Beat concept

In a previous work, for inferring human activities in a Healthcare Environment the Roaming Beat (RB) concept was used. A RB is defined as:
"The ability of an artifact to give a time stamp (hour and date) to change from motionless to a mobile state and change from the base location to any other" [15].
This concept is useful for describing the artifact behavior when an activity is performed. The behavior is obtained as a result of data conversion when the artifacts are recognized in a base location through a sensor [9, 15]. Also using RBs methodology it is possible to label video sequences through identifying artifacts.
Each artifact produces its own roaming beat when it is handled. The change of state is called "beat", so several changes of state can be observed as behavior of the artifact (see a, b, c, d in Fig. 1) as a result of the artifact recognition in a base location in a time span.

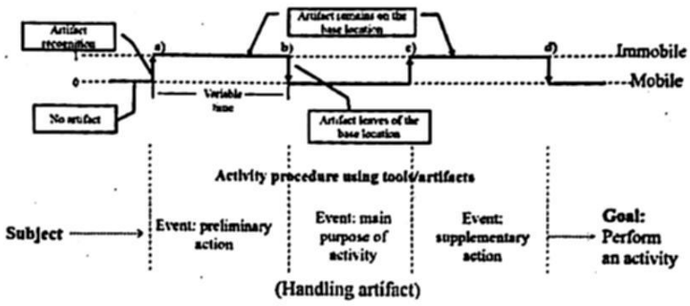

Fig. 1. Set of events related to an activity using an artifact. Artifact's behavior representation (or roaming beat) (a) first beat; (b) second beat; (c) third beat; (d) fourth beat

This concept is useful for describing the artifact behavior when an activity is performed. The behavior is obtained as a result of data conversion when the artifacts are recognized in a base location through a sensor 43. Also using RBs methodology it is possible to label video sequences through identifying artifacts. Each artifact produces its own roaming beat (RB) when it is handled. The change of state is called "beat", so several changes of state can be observed as behavior of the artifact (see a, b, c, d in Fig. 1) as a result of the artifact recognition in a base location in a time span. Each signal produced by the object recognition method, is converted into a train of pulses, as shown in Fig. 1.
Moreover, a camera is involved at the same setting. Each time that a beat is produced, an index of video sequence is related to it, so it is possible to get a specific video sequence segment related to the activity, i.e. in Fig. 1(a) starts the activity and Fig. 1(d) ends the activity. Additionally it is possible to get several images related to the activity using the indexes. It is important to note that an activity can be related to one or more artifacts, and each artifact produces its own beats when it is handled. Therefore, by each beat produced, there is a record in a database that it is composed by a specific index related to a video sequence (the first beat is equal to start index related to the activity and the last one beat is equal to the end index). Hence, it is possible to get a video segment related to the activity performed.

2.2 Identifying when activities or events happen

As mentioned above, each artifact produces its own beats when they are handled, this manipulation is detected because the artifact begins to have movement between the base location and any coordinate location in space (roaming), and can be a significant movement in the scenario when someone performed an activity. For this reason, a beat is interpreted as an event when the activity is performed. Therefore, each event is related to an index in a video sequence.
For activity recognition is necessary to define two terms: a) recognized activity, it is a set of arranged events produced by an artifact in a specific activity; and b) an event is when someone takes one or some artifacts and places them on the base location. The event can be an isolate movement.

An activity is recognized when the activity is completed by events produced by each artifact. The activity behavior is composed by three steps: 1) preliminary actions; 2) the main purpose of the activity; and 3) supplementary actions. These actions are accomplished when the all beats are obtained in the activity. The whole number of beats must be greater than or equal to 4 as shown in Fig. 1. Moreover, the beats or events number must complete a specific time t based on the activity 4. However, there are some events that may occur, so the artifact has not accomplished its established behavior in an activity. In this sense, there exist some events that can be postponed in an activity. Despite, the activity is not affected. Those events are called: preliminary and supplementary actions. These actions can be interrupted, but can be performed later in a time t+t1, and then perform the main purpose of the activity. Where t1 is the long time to take it into account in a whole activity. These kinds of events may be important; it depends of both the scenario and the user requirements. In this kind of event, the artifact can leave or remain on the base; therefore the main reason is that an activity is not finished so far.
In summary, the activity recognition and events are very important for the activity inference process. Because an activity is composed by a starting event; a behavior related to an activity is composed of a set of events; and an end activity. The end activity is performed within the activity in a time t. Those events that are started but the artifact has not accomplished, may be taken into account as important events and considered in future analysis for other intelligent systems.

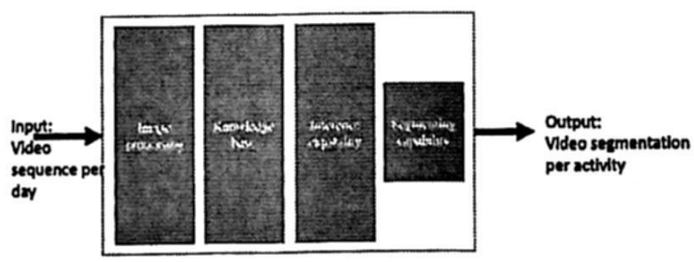

Fig. 2. General architecture of the intelligent system

2.3 Describing the activity recognition process

In Fig. 2, we present the architecture of the proposed intelligent system. The input is a video sequence and the output is a segmented video that represents an activity. The Image Processing module employs correlation filters for artifact recognition [16]; this process gives us the roaming beats. The knowledge base stores the RB of each artifact (see Fig. 3). The module of Inference Capability contains the process for recovering an activity from a knowledge base (see Fig. 4).

As aforementioned, the artifacts behavior produce a set of events within a time span, so an activity is recognized. The system shown in Fig. 3 is composed by two objects recognized, our example is based on video cameras (two video cameras for artifact recognition). The object recognition is just implemented in

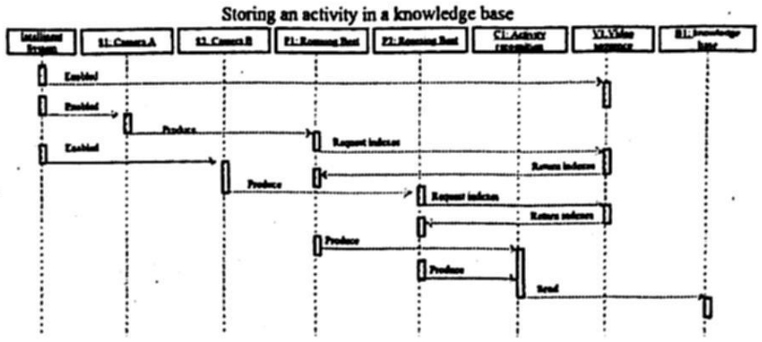

Fig. 3. Process for storing an activity in a knowledge base

a camera 1 but each beat produced is related to each index of both Camera A (S1) and Camera B (S2). Also the system includes a knowledge base in which all activities are recorded. To clarify the whole process, we divided it in two steps showed in Figs. 3 and 4. The first one is the process for storing each activity in a knowledge base and the second one is for recovering and showing an activity from the knowledge base.

Fig. 3 shows the activity recognition process 4, in which the system starts when it enables the cameras (S1 and S2) and the video sequence (V1). Each time a user handles one or several artifacts an event is produced (P1 and P2). Each event obtains an index related to image that it is requested to the video sequence and the index is returned (P1 and P2 through V1). Once all artifacts have performed an activity, the activity is sent and recorded into the knowledge base (P1 and P2 toward B1).

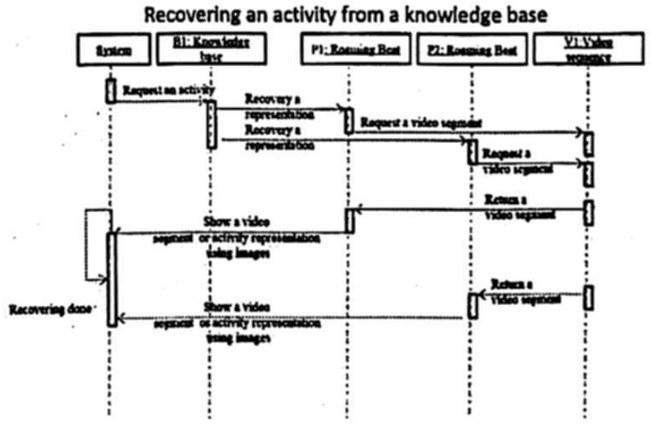

Fig. 4. Process for recovering an activity from a knowledge base

Fig. 4 depicts the process for recovering an activity from a knowledge base through a query. First, the users need to obtain information related to the performed activities and they request it to the system. It recovers an activity from the knowledge base (B1), but before that, the knowledge base joins the RBs involved in the activity requested. Each RB requests its own video segment (P1 and P2) from the whole video sequence (i.e. from current day). The video seg-

ment is returned to each RB. Once that all RB are obtained is possible to show an activity and can be in two ways: the first one is an activity representation using images obtained from each index related to an event or beat; and another is showed in a video segment. The system recoveries the indexes related to the start and end of and an activity.
In this section, we described the whole process related to activity recognition in which it was divided in two steps: the purpose of the first one is activity recognition through to obtain the artifacts' behavior; and the second one is to show the activity representation to the user in two ways: using images or video segments. However, it is necessary to evaluate our applications, so we present the evaluation protocol in the next section.

3 Results in situ

We proposed to evaluate our application in a healthcare environment. It is important to infer the activities that the caregivers perform during their workday with the aim to provide a proper care to the elders. The following sections describe the context of the environment and the main results of the evaluation.

3.1 Environment

The evaluation of the application was conducted in a private nursing home for a period of ten days during 12 hours per day. Two video cameras were installed in a room where an elderly patient with restricted mobility (ERM) was living. One camera was installed in the ceiling of the room and another was installed close to the base location (2 meters approximately). The cameras used were a wvc53gca Linksys model. The video was captured using MPGe-4 format having a resolution of 320 240 pixels, the frame rate was 2.4 per second. Storage and processing were done in a model T3500 Dell Precision workstation.

Table 1. Record of events and activities performed in a healthcare environment, Act: Activities, Ev: Events, D1 : Day 1, D2: Day 2... D10: Day 10, Tot: Total

Descrtiption	D1	D1	D2	D2	D3	D3	D4	D4	D5	D5	D6	D6	D7	D7	D8	D8	D9	D9	D10	D10	Tot	Tot
Activities/Events	Act	Ev	Act	Ev	Act	Ev	Act	Ev	Act	Ev	Act	Ev	Act	Ev	Act	Ev	Act	Ev	Act	Ev	Act	Ev
Hygiene	1	1	1	0	2	0	2	0	2	1	3	0	3	0	2	1	2	1	3	0	21	4
Blood Pre	1	0	2	0	2	0	1	0	1	0	1	0	1	0	3	0	1	1	1	1	14	2
Feeding	2	0	3	0	2	0	3	0	2	0	3	0	2	0	3	0	1	0	2	0	23	0
Medications	2	1	2	0	2	0	1	0	1	0	2	0	1	0	2	0	1	0	2	0	16	1
Total by day	6	2	8	0	8	0	7	0	6	1	9	0	7	0	10	1	5	2	8	0	74	7
Total																						81

There were a total of 109 hours recorded by camera which 941,760 images were analyzed in real time. The analysis consisted of applying 6 filters for recognition using image correlation methods for video sequence generated by the

camera that was on the base location 3 . There were a total of 81 video segments corresponding to 74 completed activities and 7 events are shown in Table 1.

3.2 Describing activities and events results

The application that was developed for this evaluation was able to infer a total of 74 activities and 7 events based on our classification as described (see Section 2.2). This assessment generates information of activities of the three types: sequential, concurrent and interleaved. The following describes the considerations taken into account with respect to the four activities as shown in Table 1:

1. The activity of taking blood pressure, was always performed independently, i.e. there was no additional activity or event when this event was performed. It was considered as a simple activity. The total number of simple activities correspond to 14 blood pressure activities, 10 of feeding activity and 3 medications, whose results were 27 activities and 2 events recorded.
2. The feeding and medication activities are sequential. These activities are sometimes performed in a concurrent way. The result was a total of 26 inferred activities and 1 registered event. These activities were considered concurrent because in healthcare environments, medicines are supplied during two feeding events corresponding to breakfast and dinner. Therefore, Table 1 shows a difference between these two activities (23 food and 16 drugs). These two activities are a total of 13 feeding activities performed concurrently and 10 performed in a sequential way. With respect to medication activity, 13 were performed concurrently and 3 were performed sequentially.
3. Hygiene activity is an activity with complex actions, because it includes a set of artifacts. Each artifact has its own behavior within the activity, so this activity was seen as interleaved activity, given the behavior of the artifacts and the people who performed the activity. We had 21 interleaved activities.

These three types of activities were successfully completed. Each behavior per artifact and per activity was established previously. However there were some events recorded that did not showed the expected behaviors. The results of these events were: two sequential activities were related to events (blood pressure event), an event associated with a concurrent activity (medicine) and four interleaved activities were related to hygiene events.
Subsequently, video segments generated by events were analyzed in order to determine the possible causes of the occurrence. The main reasons were identified: in the case of events of the blood pressure activity, it was observed that the artifact was removed from the base location for placement elsewhere; because the previous time of this activity had been placed on the base location. Another reason was the occlusion of the artifacts in its recognition, when reviewing the video related to medication; it was observed that this activity could be considered as a complete activity.
Finally, hygiene activity was the only one with a higher number of events. In

the analysis of the video segments generated by these events (4 events), it was observed that the completed behavior was not accomplished in this activity. However, it is possible to include these four events within another activity that correspond to the healing activity. This is because the video segments were very clear in this activity. Examples of movement indexes are presented below.

3.3 Showing activities based on roaming beat in a video segment

Following the activities in Table 1 and the results, let us show you the way how the events allow to recover from RB video clips or images that represent an activity.

Fig. 5 depicts how a complex activity is performed. This activity corresponds to the hygiene activity (interleaved activity). This figure shows the representation of video sequences captured by two cameras. Furthermore, this representation shows the behavior of the use of two artifacts that correspond to the paper towel and the physiological solution. Remember that, the image processing is performed on the video stream from the Camera A described in section 2.2.

Fig. 5 shows 16 beats produced by the paper towel (from 1 to 16). Each of these has an associated beat or event related to labeled picture with its number. There are also five beats produced by physiological solution (from 17 to 21) and displayed the images associated with their number.

The activity in the video segment is inferred from knowing the first and last beats corresponding to the relationship that is created when you registered the first beat with the index of the video stream into the knowledge base, so that the first beat was linked to the index 156 and the latter with the index 2160. These indexes are benchmarks that allow us to move to one side or the other. That is, index 156 tags the beginning of the sequence corresponding to the interaction between the caregiver and the patient. The caregiver enters to the room at the index number = 35 as shown in the image labeled "Initialized" in Fig. 5. It was possible to determine it by looking at the video stream from the Camera B starting at index 35. Specifically in this activity the preliminary actions of the activity start at index 35. But our application labeled the activity initialization at the index 156, also it was considered an activation of the activity from index 156 to 2160 and labeled the termination of the activity at the index 2160. The image labeled "Suspended" in Fig. 5 shows the time at which the caregiver leaves the room as a signal of completion of the activity, which corresponds to index 2220 of the sequence of video.

Another example from our application is shown in Fig. 6, which illustrates the execution of concurrent activities. Fig. 6a shows the feeding activity. This activity is related to the recognition of the tray. Figure 6b shows the medication activity. These activities were considered concurrent because when the tray arrives at the base location then the caregiver immediately performs preliminary actions to execute the activity of medication. It can be seen in Fig. 6 a small difference (in seconds) from the first beat corresponding to the activity of feeding and the first beat of medication activity. However, the activity of medication is the first one to be completed.

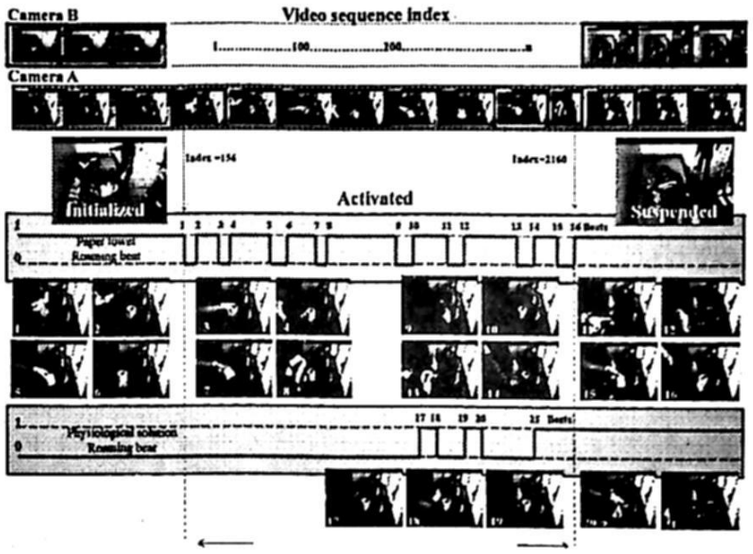

Fig. 5. Video segment related to hygiene activity using two artifacts.

As mentioned earlier, the indexes allow to move into the video stream backwards. As shown by the image obtained before the tray arrived to the base location, which was labeled as the image that initializes the activity and the beginning of the video segment. Fig. 6 shows images of the video stream from the Camera B. This is because each beat is related to the rate of each video sequence, i.e. for each camera.
On average 19 hours with 49 minutes were recorded per day. Using the activity inference application based on the manipulation of artifacts was possible to reduce the analysis of video sequences to only 2 hours 19 minutes which represents 11 percentage of the total hours in a day captured for analysis.

4 Conclusions and future work

We conclude that the activity inference is a complicated process, so it is necessary to take into account several details related to human activities. The role of the knowledge base is important to store and recover activities for obtaining a video segmentation in an automatic way. Therefore, performing activities using physical artifacts should be seen as multiple flows of actions at different times. For this reason, the technique of Roaming beat is used as a solution for the inference of sequential, interleaved and concurrent activities, in which multiple events take place at the same time. These events are part of the inferred activities from motion cues generated by artifacts while the activities are performed. Hence, it was possible to obtain video segmentation and an intelligent system in an easy way.
Each event provides a meaning of the activity based on artifacts handling. The meaning is related with three phases of the activity: start, center (main development), and end; we consider these phases as preliminary actions, actions in the activity performed, and complementary actions, respectively. In our case of

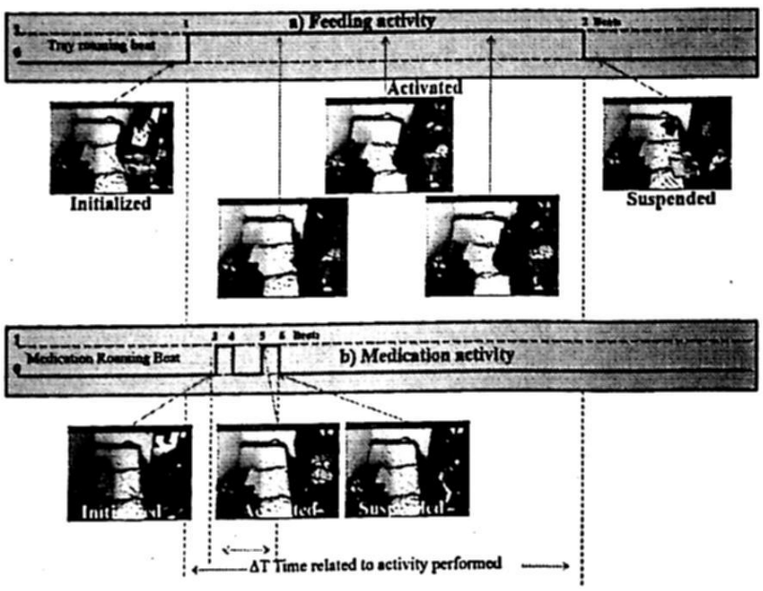

Fig. 6. Video segments related to two activities, feeding and medications activities.

study (healthcare environment), the meaning is related to the video sequence indexes; hence the indexes were used for video segmentation in an automatic way.

Moreover, each beat produced by artifacts allows us to recover video segment related to simple events or activities. These video segments are transformed to activity representation or in a means for decision making of the user.

We detailed our results for the purpose of linking multiple streams of actions. In this way it was possible to infer 27 sequential activities, 21 interleaved activities, and 26 concurrent activities. This produced 74 activity related, video segments, reducing the video sequence analysis to about 11 percentage of total daily video. Thus, the review that a person would have to make of a 12 hours video is reduced to only 2 hours.

In this work, the result obtained in a healthcare environment based on prior work [9], is presented in detail. As a future work, we pretend to implement the RB technique in other kind of scenario, the knowledge base is used only to store and recover activities and the knowledge base is static. We wish to apply data mining and artificial intelligence techniques in order to infer new activities based on the behavior of humans.

5 Acknowledgments

This work was founded by PROMEP under the contract PROMEP-103.5-13-6575 (UASLP-PTC-452) provided to the first author.

References

1. Cook, Diane J. and Augusto, Juan Carlos and Jakkula, Vikramaditya R.: Ambient intelligence: Technologies, applications, and opportunities. Pervasive and Mobile

Computing, pages 1–38,(2009)
2. Hu, Jinhui and Boulgouris, Nikolaos V.: Fast human activity recognition based on structure and motion. Pattern Recognition Letters, pages 1814–1821, 32, (2011)
3. Qian, Huimin and Mao, Yaobin and Xiang, Wenbo and Wang, Zhiquan: Recognition of human activities using SVM multi-class classifier. Pattern Recognition Letters, 31, 100–111,(2010)
4. Aggarwal, J.K. and Ryoo, M.S.: Human activity analysis: A Review ACM Computing Surveys, 3, 1–43,(2011)
5. Gu, Tao and Wang, Liang and Wu, Zhanqing and Tao, X.: A pattern mining approach to sensor-based human activity recognition. IEEE Transactions on Knowledge and Data Engineering, 23, 9, 1359–1372 (2011)
6. Kim, Eunju and Helal, Sumi and Cook, Diane: Human activity recognition and pattern discovery. Pervasive Computing, IEEE, 9,1, 48–53, (2010)
7. Bødker, Susanne: Context and consciousness activity theory and human-computer interaction,Applying Activity Theory to Video Analysis: How to Make Sense of Video Data in Human- Computer Interaction. 7, 147–174, (1996)
8. Rincón, J., Santofimia, M. J., and Nebel, J.: Common-Sense Reasoning for Human Action Recognition. Pattern Recognition Letters, (2012)
9. Martínez-Pérez, Francisco E. and González-Fraga, Jose Ángel and Cuevas-Tello, Juan C. and Rodríguez, Marcela D.: Activity Inference for Ambient Intelligence Through Handling Artifacts in a Healthcare Environment. Sensors, 12, 1, 1072–1099, (2012)
10. Lu, Guoliang and Kudo, Mineichi and Toyama, Jun: Temporal segmentation and assignment of successive actions in a long-term video. Pattern Recognition Letters, (2012)
11. Ahuja, N.: Exploiting nonlocal spatiotemporal structure for video segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 741–748. IEEE (2012)
12. Lezama, J., Alahari, K., Sivic, J., Laptev, I.: Track to the future: Spatio-temporal video segmentation with long-range motion cues. In: Cvpr 2011, pp. 3369–3376, IEEE (2011)
13. Trichet, R., Nevatia, R.: Video Segmentation with Spatio - Temporal Tubes. In:International Conference on Advanced Video and Signal Based Surveillance Video, pp. 330–335, IEEE (2013)
14. Cho, S., Kwak, S., Byun, H.: Recognizing humanhuman interaction activities using visual and textual information. Pattern Recognition Letters, 1–9, (2012)
15. Martinez-Perez, F. E., Gonzalez-Fraga, J. A., Tentori, M.: Artifacts' Roaming Beats Recognition for Estimating Care Activities in a Nursing Home. In: 4th International Conference on Pervasive Computing Technologies for Healthcare 2010, (2010)
16. Martinez-Perez, F. E., González-Fraga, J.A., Tentori, M.: Automatic activity estimation based on object behaviour signature. In: Proceedings of SPIE, 1,pp 77980E, (2010)