Automatic activity estimation through recognizing and handling objects in video sequences for image annotation and retrieval

Francisco E. Martínez-Pérez
Facultad de Ingeniería
Universidad Autónoma de Baja California,
Ensenada, México
fmartinezperez@acm.org

J. Angel Gonzalez-Fraga
Facultad de Ciencias
Universidad Autónoma de Baja California,
Ensenada, México
angel_fraga@uabc.edu.mx

Héctor G. Pérez-González
Facultad de Ingeniería
Universidad Autónoma de San Luis Potosí,
SLP, México
hectorgerardo@uaslp.mx

Mónica Tentori
Facultad de Ciencias
Universidad Autónoma de Baja California,
Ensenada, México
mtentori@uabc.mx

Abstract

Automatic estimation of human activities is widely studied topic. However, the process becomes difficult when we want to estimate activities from a video stream, because human activities are dynamic and complex. Our contribution is focused on activity estimation based on object behavior through automatic analysis of video sequences. Another contribution is focused on providing a tool with the aim of monitoring activities in a health-care environment. Our activity estimation process was developed in four phases: The first phase includes the detection of the interactions in the setting by slit-scanning technique; the second phase includes object recognition by composite correlation filters; the third phase follows several criteria for activity estimation. When the behavior of the objects related to the activities is validated, the estimation of an activity is confirmed. Each activity is related to the handling of objects, date and time of the activity, and activity description. All this information is recorded in a database; and after this the last phase includes the activity representation, using indexes for image recovery related to each activity, allowing us to create an activity representation. The activities are estimated at a 92.72 percentage of accuracy including hygiene, feeding and taking of vital signs.

1 Introduction

The tracking of human movement (human tracking [20]) using video sequences and the recognition of the type of human activities (human activity recognition [4]) are important tasks with multiple applications for video surveillance: human computer interaction including teleconferencing and content-based video-retrieval [19] from digital repositories and so forth. These areas have mainly focused on the context retrieval based on data related to different kinds of environments [22]. It has contributed to the development of computational systems capable of interpreting automatically a video sequence and extracting useful information. It is necessary to know which information is really relevant to the automation process related to human activity recognition, in order to capture and include this kind of information in the requirements analysis to develop specific algorithms.

Currently, one of the main used techniques for activity recognition is computer vision. The use of this technique is attributable to the increased computational power that allows huge amounts of video to be

Luis Enrique Sucar and Hugo Jair Escalante (eds.): AIAR2010: Proceedings of the 1st Automatic Image Annotation and Retrieval Workshop 2010. volume 1, issue: 1, pp. 11-23

processed and stored. However, automatic activity recognition is not a simple task due to the particular way humans perform their activities and the different tools or objects used in those activities. Therefore, the nature of human activities poses the following challenges: i) Recognizing concurrent activities; ii) Recognizing interleaved activities; iii) Ambiguity of interpretation and iv) Recognizing multiple persons as shown in [4].

Due to the availability of large and steadily growing amounts of visual and multimedia data, Content Based Image Retrieval (CBIR) has created thematic access methods that offer more than simple text-based queries based on matching exact database fields. For example, in the medical field, digital images used for diagnostics and therapy, are produced in ever increasing quantities. In reference [14] was reported a review that concentrates on image retrieval in the medical domain, that does a systematic overview of techniques used, visual features employed, images indexed and medical departments involved. Although the need for information in a quick and timely decision-making has been increasing in hospital environment, a few research efforts have being reported for automatic activity retrieval. In reference [18] was reported the need to maximize the attention span and decrease the spent time to record health-care of patients where one option is the automation of recording of health-care activities. Many systems propose to use text from the patient record [7] or studies [2] to search by content in distributed data bases. Other researches classify the images to augment text-based search with visual information analysis [8, 21] or a semi-automatic method for image annotation shown in [9] . Basically all systems that give details use color and grey level features, mostly in the form of a histogram [17, 21].

The work cited above, it has assumed that images or video sequences are previously stored in several data bases before the retrieval process; even images or video sequences are acquired or labeled in a manual form. This work shows the whole process since acquisition to activity representation, and automatic labeling of images in video sequences. We use the activity term as annotations in images that are linked to one or several objects used in the health care activities.

One contribution of this paper is the activity estimation based on object behavior through automatic analysis of video sequences. This approach automatically recognizes the human interactions that happen in a specific setting, through handling several objects in a base location, which leads us to infer the activity in an automatic fashion. To check when an activity happens, we recognize the interactions taking into account a whole view of the scene. The information obtained from the recognition and behavior of objects is processed and used to obtain visual representations of activities.
Another contribution is focused on providing a tool with the aim of monitoring activities in a particular health-care environment. Specifically, we provide inferences and representations related to activities that caregivers perform to the elderly. We identified four important activities related to health-care activities: the taking of blood pressure; the measuring of blood glucose; the activity of providing patient hygiene and the feeding activity.

This paper is organized as follows. In section 2, we present the implemented methods. Section 3 discusses the interactions in the scene. In section 4, we present the object recognition using composite correlation filters. In section 5 we show the activity classifications. In section 6 we present how recovering information is related to the activities. In section 7 we present our results. Finally, Section 8 presents the conclusions and directions for future work.

2 Methods

In our approach, activity recognition is focused on what is happening in the scene, where both object and people are involved for their recognition, on tracking such objects from frame to frame, and on analyzing object tracks to recognize their behavior in a scene. So, it is necessary to know and model the context/environment that we are interested in monitoring as mentioned in [20, 22, 5]. Therefore, we developed a case study that provided us functional requirements to create a monitoring system. The case study also considered the user's requirements related to recovery and representation of the information based on performed activities.

The study was conducted in a private nursing home in the city of Ensenada , Mexico. Figure 1 shows the patient's room being monitored and the areas where the interactions flow of humans and objects is realized. We captured almost 400 hours of video sequences of the setting. The observational study was recorded with two cameras as shown in Figures 1b and 1c.

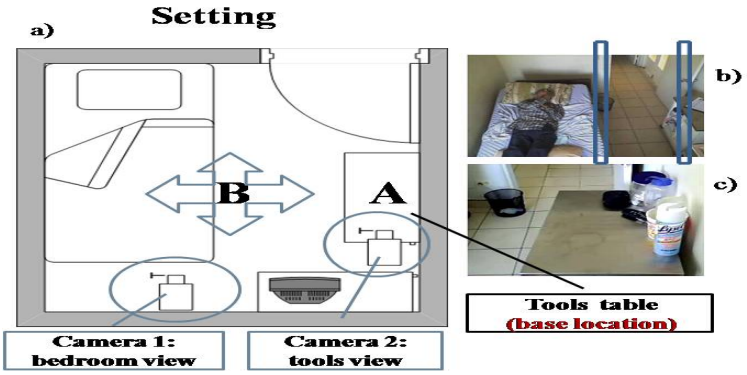

Figure 1: (a)Position of the cameras in patient room; (b) A patient recorded with the camera with room view (zone B); (c) The tools table recorded with the camera with tools view (zone A)

2.1 Methodology

This work is based on the hypothesis that human activities can be inferred by the interactions with objects [16][6].For this reason, the proposed process for activity estimation was developed in four phases: The first phase includes the detection of interactions between humans and objects in the scene through the implementation of the slit-scanning technique proposed by [12, 3]. If an interaction is detected, it means that an activity has started and must be validated.

Validation of an activity involves the recognition of objects that are handled in the setting, their behavior and the duration of the activity. The second phase takes care of object recognition, which was implemented with composite correlation filters. In the case study we identified the kind of objects that were related to a particular activity and the duration of each activity; so, as long as we have these features we will be able to create an activity description. All this information is recorded in a database.

The third phase identifies objects behavior and it automatically links them to a specific activity. In an activity, caregivers can use one, two or more objects, depending on the activity, the objects show a different handling behavior, and so we need a recognition filter for each current object in the setting.

Confirmation of activities allows us to create both indexing and linking of the video sequenced images with each activity in the database, as well as recovering more information from the scene for contextual representation of the activity. This kind of representation can be used as a tool for quick reference video related to monitoring. The last phase explains how the indexes are used for information retrieval related to the activities that happened in the setting and it allows us to create an activity representation and description. These four stages were adapted to the general framework of visual surveillance pro-

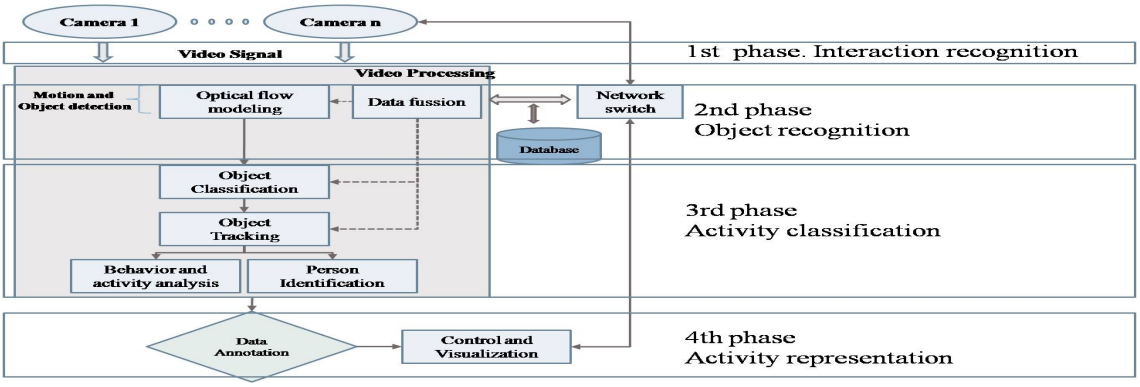

Figure 2: General framework to visual surveillance and our four stages implemented

posed by [11]. We adapted the object detection model for an optical flow model. The second adaptation is related to the elimination of object segmentation module of the original framework. This adaptation was necessary because the composite correlation filter method it does not need image segmentation to recognize and track objects. Figure 2 shows the general framework that includes these adaptations in the second stage (mentioned above) and stages related to this work that will be discussed later .

3 Detecting interactions

In healthcare environments for elders it is very important to know what kinds of interactions are performed by caregiver towards them. For example, it is important to check when healthcare activities are performed; therefore, it is necessary to check that the activities are being performed in a correct form. We will use the term interaction to refer to the action in which a caregiver is handling objects and the actions when the caregiver is moving from base location to the patient's bed or vice versa . This is with the purpose to know interactions in the scene. According to figure 1, interactions can be developed in zones A and B. Interactions in zone A is when caregivers are handling objects in a base location. An interaction in zone B is when someone performs an activity from base location to the patient's bed or vice versa. Interactions detection in both zones A and B are obtained using slit-scanning technique. This

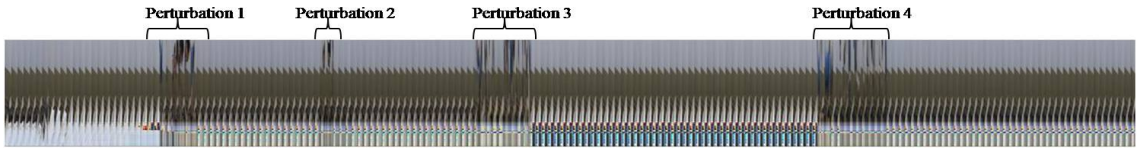

Figure 3: Composite image over time, showing changes and perturbations in the scene

technique creates a composite image of video activities over time. Slit scanning, originally developed in photography, exposes film to only a narrow slit from a scene; while panning the camera smoothly

captures a normal scene, interesting images are created by irregular panning (spatially distorted scenes), or when objects moving in the scene are seen as motion over space. The same approach is realized in video by video slicing. Video slicing first extracts a scan line (vector) from a video frame, and then adds that line to a composite image over time as shown in figure 3. The contiguous video slices in each row give observers a sense of the video history, where changes and perturbations are easily seen as presented in [15]

3.1 Interactions as cues

To get the interactions , we obtained two pixel vectors of each image in a video sequence as shown in figure 4. Vector **a** belongs to the patient bed edge and the width size is a pixel (figure 4-(a)), and vector **b**, belongs to the base location as shown in figure 4-(b). Position related to the two vectors is due to the behavior that the caregivers showed in their activities. The main idea of handling of these two vectors is to find drastic differences which are linked to interactions in the scene, being a sign that something happens.

The process to obtain the interactions is as follows: Each vector is normalized using its highest value, because highest values show a significant change related to the person or object in the scene. From vector 2 to 30, we obtain the sum of the absolute difference by the incoming vector with its predecessor. If there is a difference greater than 2, then the first vector is stored in a temporary variable, and also it is considered that there is an interaction in the scene. In this way, the system starts to record the behavior and take the significant changes as an interaction cue that is performed. We take into account the previous 15 minutes before the interactions happen in the scene, as a threshold reference to compare the behavior in the scene versus lighting variations that happen in a day. When highest values exceed the threshold, the interaction is checked in the room. Figure 4 shows an example related to the execution of blood

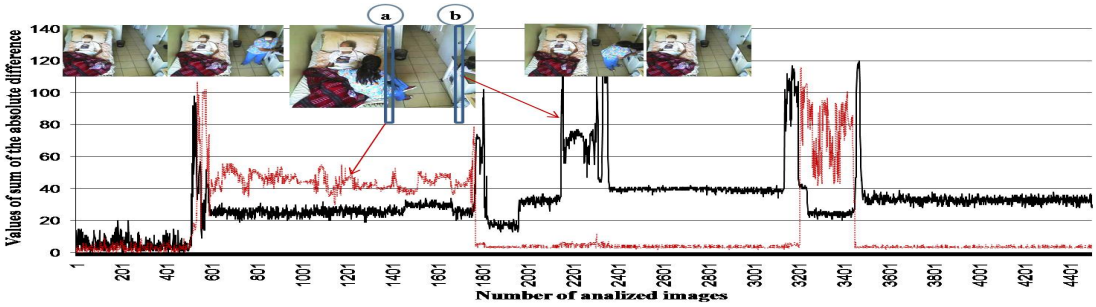

Figure 4: Analysis of blood pressure activity, a) Vector in zone B, b) Vector in zone A

pressure activity. The black line in figure 4b shows the result of the interactions performed in the zone A. The red line in figure 4a shows the result of the interactions performed in the zone B. The process starts without interactions in the room from frame 1 to 500 in figure 4. After that, a caregiver enters in the room and moves to the base location where he puts an object on the table (from frame 501 to 550). In this zone, the value of the line changes and remains high, and this is due to the object that was put there as shown from frame 601 to 1750. In the range from frame 501 to 1801 it can be seen how the lines are crossing, meaning the transition between zones from A to B. From frame 2101 to 2401 it can be seen another example performed in the zone A. Finally, from frame 3150 to 3450 it can be seen two interactions performed in the zone A but between them there is an interaction performed in the zone B.

The analysis showed that when there is not an activity performed in the room, the lines behavior is almost constant and without exceeding a difference more than 25 points as shown in figure 4 from frame 1 to frame 500. Several times peaks are visualized that belong to adjustments of the intensities of light that

affected the cameras as shown in figure 4 from frame 1950 to 1980 in black line (figure 4b). When a caregiver enters and leaves at an instant in the room, the lines in the graph shows a significant variation as can be seen with the highest peaks in figure 4 from 3101 to 3201 among other times. However, when a caregiver remains immobile, his variation leads to stabilization within a range of at least 15 as shown from frame 501 to 1801 of the line in figure 4a.
Using slit-scanning technique it was possible to check the interactions performed in the setting. The highest values are related to the activity performed when it starts or ends. Therefore, these highest values represent a cue of interaction, namely something is happening in the scene. Based on these highest values, the indexes related to these frames can be saved in a database as a cue of when the interaction starts or ends. However, these data are just to begin the activity estimation, so in the following sections are discussed the steps to consolidate the activity estimation.

4 Object recognition by correlation filters

Based on our study, we decided to create tags for the recognition of six objects related to four impo rtant health activities: the baumanometer object is related to the activity of taking blood pressure; the dextrose kit with the sample is related to the activity of measuring blood glucose; the tray is related to feeding and there are three objects involved with the activity of patient hygiene such as the toilet paper, saline solution and a lotion for the body. These objects are manipulated in a base location so that we confirm the hypothesis for the inference of activities from the handling of objects. We used the optical flow advantage by composite correlation filters implementation as objects recognition and tracking method in dynamic settings. To this end, we use MatLab as a programming language.

4.1 Non-linear composite correlation filters

Object recognition based on correlation filters computes a level of similarity between two images: i) the reference images and ii) the test image or the captured setting frame as figure 5 shows. The image of the workplace scene is used to test the filter in real time and it is matched with a reference image previously recorded and used to train the filter. One of the advantages in using correlation filters is that we can locate multiple objects without segmentation in a scene, reducing the processing time. We

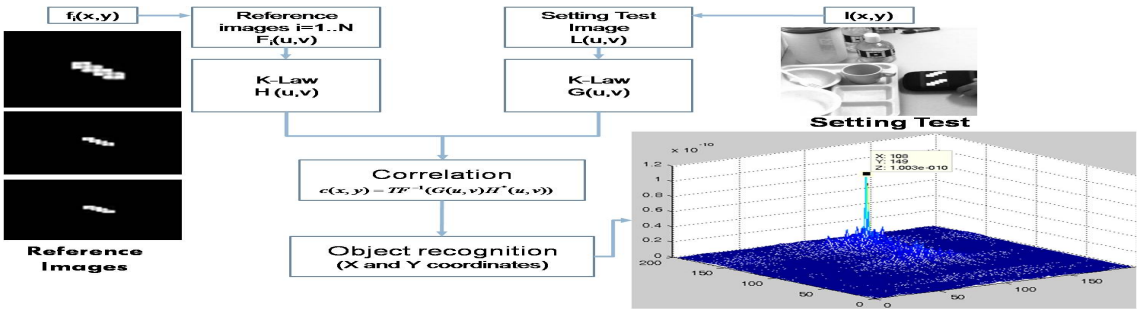

Figure 5: Object recognition by composite correlation filter

implemented the kth-law synthetic-discriminant function because it has been shown that nonlinear filters have tolerance to some object deformations and good performance in the presence of different types of noise [1].

The kth-law composite filter in the frequency domain is:[1]

$$\hat{h}^k = \hat{S}^k((\hat{S}^k)^+\hat{S}^k)^{-1}c^* \tag{1}$$

Where the **S** matrix of size PxN (number of pixels in each images and number of training images), has the vector form of the Fourier transform of each training image as its i-th column, S^+ is the complex conjugate transpose of S and $()^{-1}$ denotes the inverse matrix. Vector c contains the desired cross-correlation peak value for each training image and factor k is a nonlinear operator affecting the module of each Fourier transform in **S**, k is a real value between 0 and 1. For doing the correlation operation we re-ordered the vector **h** into his 2D form, and finally we get the filter $H(u,v)$. The correlation operation showed in figure 5 can be defined in terms of Fourier transform by

$$c(x,y) = TF^{-1}(G(u,v)H^*(u,v)), \tag{2}$$

where $H^*(u,v)$ is the complex conjugate of the k-law filter, $G(u,v)$ is the test scene preprocessed with the nonlinear k-law factor, TF^{-1} is the inverse Fourier transform and $c(x,y)$ is the correlation output.

5 Activity classification

The correlation values, obtained in the previous section, are converted to ones and zeros where1 stand for an object that has been recognized and 0 when there is not any object in the scene. In this way we converted the whole stream of video sequences in a train of pulses as can be seen in figure 6a and 6b. Each activity can be composed by one, two or several objects. Each object has a different behavior related and according to the activity it performs. An objects can appear in the setting one time, two times or several times, producing a signature related to its behavior when entering and leaving the setting as shown in figure 6; and is processed in a thread in which its behavior is analyzed. In order to relate each

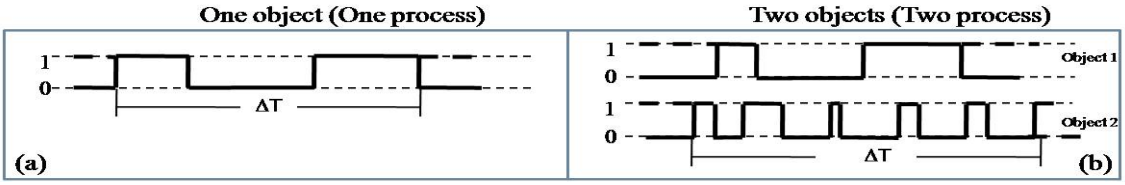

Figure 6: Grouping Objects' behavior in a specific activity, a) Activity using a simple object b) Activity using two objects

object to an activity, it is necessary to validate three states; and every observed object must perform one of them [13]:
i) Initialized, when an object is being set up and placed in its base location (e.g., the tools table).
ii) Activated, when object's beat changes (e.g., from a motionless to a mobile state) or remains in the same state (e.g., mobile) and;
iii) Suspended, when an artifact remains motionless and it is suspended.
These three states are checked all the time. Namely, the initialization phase starts when the first object appears in the base location. This occurrence is an action and is recorded in a temporal database. When the object is recognized, the time and date in which the event started are stored; then, the activated state starts and checks the objects' behavior in the base location. This means that an object's behavior is happening again. The time and date in which the recognized objects enters and leaves are recorded again. Finally, once the period of time after the object disappears or remains motionless longer than a

threshold, the activity is suspended. At this point, if the object's behavior is similar to the programmed behavior then the activity is inferred and data related to that activity (activity name, number of objects recognized, date, activity time in which starts and ends, and the duration of the activity) are recorded in the database. Criteria timeouts for mobility-immobility were proposed according to the average times observed in the case study and is equal to the average total time of activity divided by the number of appearances on the scene.
In order to implement this behavior signature in the inference of the four activities, it was necessary to use concurrent processes. For each process, we classified the activities independently which allowed us to infer several activities simultaneously. Each process is responsible for retrieving information relevant to the object, and the proper analysis of that behavior.
Grouping objects and linking these groups to specific activities; and three states related to the object's behavior, allowed us to accomplish the early steps of the challenges related to activity recognition. Such activities include concurrent activities, interleaved activities and ambiguity of interpretation, using vision techniques. Each time an object enters or leaves the scene it will be recognized the index related to the video sequence will be linked and will be recorded in the temporal database. This process will be executed including the camera installed in the base location and the other cameras installed in the scene. All this, allowed us to obtain a better representation related to the activity that is discussed in the next section as an example of recovering a representation.

6 Recovering a representation

Building a representation based on activity interactions is a hard work, because it is necessary to know what information is relevant related to specific points. Such points in this approach are obtained through indexes created in the inference of the activity. The activity includes where the starting and ending indexes are and also the indexes related to the handling object. Therefore, indexes are used for image recovery related to each activity, allowing us to create an activity representation. In the same way, slit-scanning technique gave us an interaction representation related to the scene as mentioned above, however there are a few elements of the event to highlight.
We relied on our inference results where each activity is related to indexes and has a duration of execution. We selected timeline technique to create a representation based on a span of time (duration of activity) in an easy way. The timeline allows us to a very rapid and detailed exploration of the video history where each slice can highlight details related to an activity. We extend the timeline representation using slit-scanning technique by obtaining several vectors to complete a window size of 250, which starts from 100 to 350 lines obtained of 250 frames of video sequence as shown in figure 7d. This kind of view extends the exploration area, and in this case our representation allows us to show a better view of the objects that are being used in a clear way as shown in figure 7b(4). Figure 7 shows an example in which one hour was obtained (figure 7a) and four performed activities can be seen. Figure 7b shows a specific activity performed in which can be seen the specific object used in the activity. After that, figure 7c shows a representation obtained in one minute. Finally, figure 7d shows the x position in which starts getting the vectors and this position goes increasing and moves according to the frame in the video sequence until the complete window size of 250 and restarts again at starting position of x and repeats again the process.
Another representation is based on windows that are created by each activity. Where images are recovered that is based on indexes related to the activity of video sequence. Namely, for each window created, we obtain two images that are located in the middle zone of each one of them; also two other images are located in the start and end of each window. For example figure 7b shows seven windows created in a whole activity, so we obtained 28 images to represent the activity. We know that with this representation

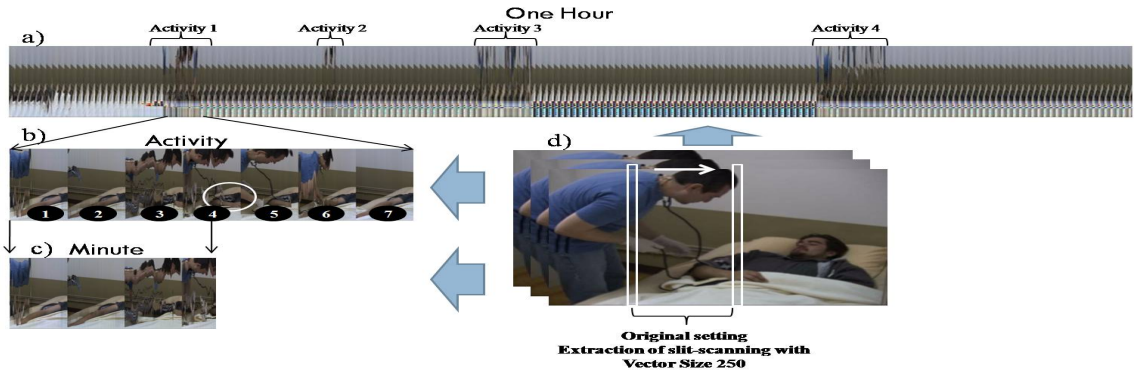

Figure 7: a)Slit-scanning's window per hour, b)Per activity c)Per minute d)Getting a window with size 250

we can omit some details related to activity performance; however, our intention is to reduce the query time in videos so that this does not become too tedious. Furthermore, this information can be configured to send alarms related to risk events.

6.1 Gathering phases

In this subsection, we present our approach gathering the phases as a system. First, we introduce the interactions in the setting. This phase is strongly linked to the object recognition (phase 2). Namely, the system checks any interactions that happen in the scene. Once detected, the system is waiting that the object recognition module gets the first result with the aim to record it in the database. After the system continues awaiting that next interaction will happen in the adjacent area. Once time the system identified several movements in both areas, the activity is confirmed and something happens in the setting.
On the other hand, for each time object recognition is done, these results are sent to the phase 3. In which, the results are converted to create the activity estimation based on object behavior and their manipulation. Finally, as result of this phase, we obtained the indexes related to the video sequences and it is possible to construct an activity representation and determines whether it is necessary to send an alarm or simply save the representation as a reference tool.

7 Results and discussion

The system was evaluated in the usability lab at UABC campus Ensenada, Mexico, over a period of 5 days and where we replicate the handling of the objects used within the activities inside the nursing home. In this evaluation, we performed 11 activities per day: 4 hygiene activities, 3 feedings, 2 blood pressure measure and 2 samples of glucose (dextrose). In total, we carried out 55 activities that were recorded and processed. Our system was able to recognize 51 activities, and the rate of effectiveness for activities inference was 92.72 percent. The missing activities that were not identified, were because, these did not exceed the established threshold. This was due to occlusions, shadows and light changes, affecting the outputs of correlation filters and producing results below threshold. Another cause was the speed how caregivers perform the activity because the time for object recognition is not enough. On the other hand, the interaction was used as cue to detect whether an activity is performed, but the activity is confirmed only if the objects have a known behavior as described in Activity Classification Section.

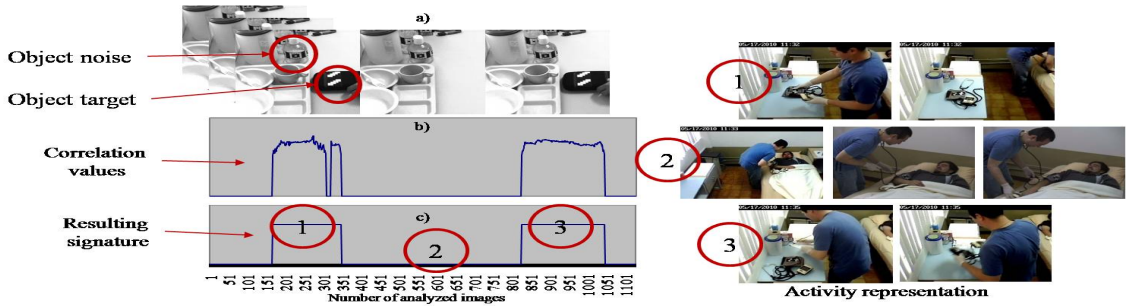

Figure 8: Object behavior to infer activities

7.1 Estimation and representation of activities

The execution of the activities was recorded by three cameras wvc53gca Linksys in the usability lab. The video was captured using MPGe-4 format with a resolution of 320x240 pixels. One camera was located on the top of the room; another one was located in proximity to the preparation table and the third was located near to the left side of the patient. Figure 8 shows an activity performed related to the blood pressure activity. Also figure 8a shows the video sequence captured in which it can be observed that there are several objects however just one object is the target (dextrose kit object) and other ones represent a kind of noise. The target object appears in the setting two times as shown in figure 8b. In this process the three states of initialization, activation and suspension that are related to the object are recognized. After that, the correlation results are converted in a train of pulses as shown in figure 8c.
Obtaining this conversion allowed us to create a meaning related to the activity and the train of pulses. The information obtained of the conversion, belong to object recognition that in this case the object is the baumanometer; number of occurrences; time and date in which the activity was performed; duration of the activity; and a meaning that belongs to the behavior related to: 1) Object appears in the base location, meaning that the user is interacting in the room; 2) the user is interacting with the patient because the object is being used and 3) the object appears again and the user is recording the activity as shown in figure 8 (1,2,3). Also, we obtained a representation that is based on the indexes of each frame, which was obtained in the activity performed where we can highlight that: 1) the image in which first time the object appears in the setting and the user is changing his position from zone A to zone B; 2) the image when the object disappears of the base location; the image of the activity performed and the image obtained 2 seconds before that the object was removed of the patient; 3) the image when the user is recording the activity; and the image where the object is removed to the base location.
Once the activity has been inferred, the information related to the activity is stored and available in the database. The activity is labeled and is possible to identify it as ”Blood pressure activity”. Also, this activity is linked to an abstract level that belongs to the vital signs activity.
These results allowed us to identify the interactions in the scene. Object recognition and classification based on the states that arose according to objects’ behavior was accomplished. Finally, we presented an activity representation whose aim is to provide a tool for checking the execution of the activities.

8 Conclusions and future work

We presented an approach for automatic activity estimation. Our approach was developed based on a modified version of the general framework proposed by [10]. We divided our development in four phases that includes recognizing of the interactions in the scene; object recognition through using com-

posite correlation filter; activity classification; and two activity representations related to the healthcare activities performed by caregivers in a visual fashion.

Using slit-scanning technique we checked the interactions that happen in a scene. The highest values are related to the activity performed when it starts or ends. Therefore, these values represent a cue of interaction, namely something is happening in the scene. Based on these highest values, the indexes related to the activity are saved in a database as a cue of when the interaction starts or ends.

Grouping objects and linking these groups to specific activities allowed us to accomplish the early steps of the challenges related to activity recognition. Such activities include concurrent activities, interleaved activities and ambiguity of interpretation, using vision techniques.

Our algorithms were able to recognize 51 activities from 55 performed, thus the rate of effectiveness for activities inference was 92.72 percent.
Indexes generation of a video sequences allowed us to obtain data related to activities such as time, date, index of each frame and objects involved in a specific activity. Using slit-scanning technique, we obtained a better and wider interpretation related to activity estimation through of a visual representation. We know that with this representation we can omit some details related to activity performance; however, our intention is to reduce the query time in videos so that this does not become too tedious. Furthermore, this information can be configured to send alarms related to risk events.
On the other hand, for each activity estimation allows us to create annotation in a data base linked to the indexes of the video sequences in real time. Using these indexes, we provide an abstract meaning that it is possible to extend showing their components involved. At the same form, using these indexes through linking to other video sequences acquired by other cameras, which are installed in the setting. Therefore, to give another meaning to the activities, where it is easy to see and identify actions. Additionally, it is possible to see specific situations like activities without to check all video sequence. Allowing that other applications as shown [7, 2] will be focused on images retrieval specifically on these video segments.

This approach was implemented in Matlab, it is due to, allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including java, C, C++, and Fortran. However our intention is develop this system in another language as java, because it provides several advantages like portability, general-purpose, concurrent, and so forth. In addition, we must be aware of the amount of data that we can obtain from handling video stream. It involves to the elderly, staff and relatives, but their perceptions and implications have not been considered in this research. Other topics for us are the privacy and security of data, but these are beyond our reach.
The evaluation of this work was developed in our usability lab, however our intention is to implement this approach in real scenarios and this implies having a better performance in our filters. Also, to extend this approach we have planned to recognize the people that are performing the activities in the room. Another ongoing work is related to knowing and monitoring the postures of immobile patients. Finally, we intend to improve our activity representation related to the activities by responding user requirements.

9 Acknowledgment

We thank the personnel at Lourdes Nursing Home in Ensenada, México, especially to Argel Grisolle. This work was funded through the scholarship provided to the first author (CONACYT 243422/217747). This work has been developed within the program Maestría y Doctorado en Ciencias e Ingeniería (My-DCI) at UABC.

References

[1] Javidi Bahram, Wang Wenlu, and Zhang Guanshen. Composite fourier-plane nonlinear filter for distortion-invariant pattern recognition. *Optical Engineering*, 36(10):2690–2696, 1997.

[2] Amalia Charisi and V. Megalooikonomou. Content-based medical image retrieval in peer-to-peer systems. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 724–733. ACM, 2010.

[3] Andrew Davidhazy. Slit-scan photography. In *School of Photographic Arts and Sciences Rochester Institute of Technology*, 2007. `http://people.rit.edu/andpph/text-slit-scan.html`, last viewed Nov 2010.

[4] Kim Eunju, Helal Sumi, and Cook Diane. Human activity recognition and pattern discovery. *Pervasive Computing IEEE*, 9(1):48–53, 2009.

[5] Bremond F. and Nevatia R. Representation and optimal recognition of human activities. *Proc IEEE Conf. on Computer Vision and Pattern Recognition CVPR 2000*, pages 818–825, 2000.

[6] J. Favela, M. Tentori, Luis A. Castro, Victor M. Gonzalez, and Elisa B. Moran. Hospital Workers' Activities and its use in Context-Aware Hospital Applications . In *Pervasive Healthcare*, Austria, 2006.

[7] Sergio S. Furuie, Marina S. Rebelo, Ramon a. Moreno, Marcelo Santos, Nivaldo Bertozzo, Gustavo H. M. B. Motta, Fabio a. Pires, and Marco a. Gutierrez. Managing Medical Images and Clinical Information: InCor's Experience. *IEEE Transactions on Information Technology in Biomedicine*, 11(1):17–24, January 2007.

[8] Hayit Greenspan and Adi T Pinhas. Medical image categorization and retrieval for PACS using the GMM-KL framework. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 11(2):190–202, March 2007.

[9] L. Hollink, S. Little, and J. Hunter. Evaluating the application of semantic inferencing rules to image annotation. In *Proceedings of the 3rd international conference on Knowledge capture*, pages 91–98, New York, New York, USA, 2005. ACM.

[10] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transact on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(3):334–352, 2004.

[11] Teddy Ko. A survey on behavior analysis in video surveillance for homeland security applications. In *Applied Imagery Pattern Recognition Workshop 2008 37th IEEE*, pages 1–8. IEEE, 2009.

[12] Golan Levin. An informal catalogue of slit-scan video artworks. 2006. `http://www.flong.com/writings/lists/list_slit_scan.html`, last viewed Nov 2010.

[13] F. E. Martinez-Perez, J. A. Gonzalez-Fraga, and M. Tentori. Artifacts' Roaming Beats Recognition for Estimating Care Activities in a Nursing Home. In *4th Internat. Conf. on Pervasive Computing Technologies for Healthcare*, Munchen Germany, 2010.

[14] H. Muller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content-based image retrieval systems in medical applications clinical benefits and future directions. *International journal of medical informatics*, 73(1):1–23, February 2004.

[15] M. Nunes, S. Greenberg, S. Carpendale, and C. Gutwin. What Did I Miss? Visualizing the Past through Video Traces. In *ECSCW 2007 Proceedings of the 10th European Conference on Computer Cooperative Work Limerick Ireland 24-28 September 2007*, number 2007, page 1. Springer Verlag, 2007.

[16] Matthai Philipose, Kenneth P, Mike Perkowitz, Donald J. Patterson, Dieter Fox, and Henry Kautz. Inferring Activities from Interactions with Objects. *Portal*, 2004.

[17] Marcelo Ponciano-Silva, A.J.M. Traina, P.M. Azevedo-Marques, J.C. Felipe, and Caetano Traina. Including the perceptual parameter to tune the retrieval ability of pulmonary CBIR systems. In *Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on*, pages 1–8. IEEE, August 2009.

[18] Monica Tentori and Jesus Favela. Activity-aware computing for healthcare. *IEEE Pervasive Computing*, 7(2):51–57, April 2008.

[19] Hu Weiming, Xie Dan, Fu Zhouyu, Zeng Wenrong, and Maybank Steve. Semantic-based surveillance video retrieval. *IEEE transactions on image processing*, 16(4):1168–81, April 2007.

[20] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking A survey. *ACM Computing Surveys*, 38(4):13–es, December 2006.

[21] Lei Zheng, A.W. Wetzel, John Gilbertson, and M.J. Becich. Design and analysis of a content-based pathol-

ogy image retrieval system. *Information Technology in Biomedicine, IEEE Transactions on*, 7(4):249–255, December 2004.

[22] J. Zhu, Y. Lao, and Y. Zheng. Object Tracking in Structured Environments for Video Surveillance Applications. *IEEE Transact on Circuits and Systems for Video Technology*, 20(2):223–235, February 2010.