International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:12, 2009

Real time detection, tracking and recognition of medication intake

H. H. Huynh, J. Meunier, J.Sequeira, and M.Daniel

Abstract — In this paper, the detection and tracking of face, mouth, hands and medication bottles in the context of medication intake monitoring with a camera is presented. This is aimed at recognizing medication intake for elderly in their home setting to avoid an inappropriate use. Background subtraction is used to isolate moving objects, and then, skin and bottle segmentations are done in the RGB normalized color space. We use a minimum displacement distance criterion to track skin color regions and the R/G ratio to detect the mouth. The color-labeled medication bottles are simply tracked based on the color space distance to their mean color vector. For the recognition of medication intake, we propose a three-level hierarchal approach, which uses activity-patterns to recognize the normal medication intake activity. The proposed method was tested with three persons, with different medication intake scenarios, and gave an overall precision of over 98%.

***Keywords*—**Activity recognition, background subtraction, tracking, medication intake, video surveillance

I. Introduction

As the aging population increases rapidly in developed countries, methods are required to effectively and automatically take care of the aged persons in their home with the lowest possible cost. For this purpose, many healthcare tools, integrating information technology, might be proposed such as mobile-care, home-care, and telemedicine. They all focus on reporting the health status of the aged persons to their family or caregivers.

In this context, efforts have recently been made on developing computer vision systems for monitoring medication intake [1], [2], [3]. In [1], constraints are used to locate the face region and template matching with a few assumptions allows detecting and tracking the face. In the same paper, an approach based on grayscale sharpness is developed to localize and track the hands. In addition, the orientation of the fingers is used to recognize the opening and closing of medication bottles (named bottles in the following for conciseness). In their study, the fingers must always be visible, which is not always the case in a real situation. Also, the medication intake is simply detected if a predetermined sequence of actions occurs in a sequence of frames without analyzing the duration of these actions or other possible sequences of actions, which may cause an increase rate of false or missed detections.

In [2], the authors used the algorithm described in [4] to track the head. They modeled the head with an ellipse whose size can vary from frame to frame. For each image, a local search determines the best fitting ellipse, based on the gradient magnitude around its perimeter and the likelihood of skin color inside it; the hands are localized based on regions with high skin color likelihood. These regions are extracted from the skin color likelihood map created during the head tracking process. Several simplifying assumptions of the problem were used such as supposing that the patient is wearing a long-sleeved shirt. This avoids identifying the hand in the arm region which makes the system less flexible. The medication intake activities are represented in a hierarchal way, from low to high level events and are recognized using two simple automata representing two different ways of taking the medication.

In [3], the detection and tracking of skin regions are done using color information; bottles are described by color and shape features. As in [1], many constrains are used to detect and track the face. Hand tracking is done by exploiting the edges and centroid properties of the regions. For modeling medication intake activities, the authors used a Petri network. The token transition from a place to the other (during medication intake) occurs only if the events go on during a specific number of frames.

In all these papers, the experimental settings were very simple and do not include, for instance, the use of a glass of water during the medication intake activity. Furthermore, the real time aspect of the algorithms was not of prime importance. Finally, no parameter learning was used to better model the specific behavior of the elderly.

In this paper, we present a real time system for detection, tracking and recognition of normal/abnormal medication intake activities. We use color information to detect and track moving object and we propose a hierarchal approach, based on activity-patterns to recognize medication intake. Our work proposes improvements in the detection and tracking of moving objects in comparison with the works presented in [1],

H. H. Huynh is with the Department of Computer Science, Université de la Méditerranée, Marseille, France (e-mail: Hung.Huynh-Huu@univmed.fr) and the Danang University of Technology, Danang, Vietnam (e-mail: hhhung@ud.edu.vn).

J. Meunier, is with the Department of Computer Science and Operations Research, Université de Montréal, Montreal, Canada (e-mail: meunier@iro.umontreal.ca).

J. Sequeira and M. Daniel are with the Department of Computer Science, Université de la Méditerranée, Marseille, France (e-mails: {Jean.Sequeira, Marc.Daniel}@univmed.fr).

[2], [3]. For example, we also consider other events such as taking a glass of water and we describe how to use a simple learning strategy for the particular behavior of an elderly. Moreover, the proposed approach can recognize different ways of medication intake, with a high success rate.

This document is structured as follows: in section II, we present the general algorithm of the system. The detection and segmentation of moving objects is detailed in section III. The segmentation of skin color regions in RGB normalized color space (rgb) is also presented in this section. Tracking of moving objects and occlusion handling are presented in section IV. In section V, we introduce a new approach for recognition of medication intake activity, an activity-pattern based hierarchy combined with real time event compression. We present the results and the performances of our implementation in section VI and a discussion with conclusion in section VII.

II. SYSTEM OVERVIEW

The medication intake activity involves many objects and takes place in many different ways. The recognition of medication intake is consequently composed of several steps: from low level to high level ones. In our system, processing at a lower level consists of: (a) modeling the background, (b) identifying moving objects, (c) segmenting skin regions, (d) detecting the face and mouth, and (e) tracking moving objects. At a higher level we find (f), the recognition per se of the medication intake activity.

For the low level processing (a-b), background subtraction in the YCrCb color space is used to detect moving objects while the segmentation of skin regions (c) and detection/tracking of the mouth (d) is done in another color space: rgb. The tracking (e) of skin regions is based on a minimum displacement distance criterion while the (colored) bottle tracking uses the color space distance to their mean color vector.

At the higher level (f), a hierarchy-based recognition approach combined with real time event "compression" and activity duration constraints are used to recognize normal/abnormal medication intake activities.

III. DETECTION ANDD SEGMENTATION

In our system, we have to track several types of mobile objects: the face, the mouth region, the hands, a glass of water and the bottles. Fig. 1 shows an indoor scene with the objects to be tracked.

A. Moving object detection

Moving object detection provides a classification of the pixels in the video sequence into either foreground (moving objects) or background. Background subtraction is a method typically used to segment moving regions in image sequences taken from a static camera by comparing each new frame to a model of the scene background. This approach is widely used for various applications such as video surveillance systems and traffic monitoring. It maintains a background image or model, to classify new observations as background or foreground regions. The difference between the current frame and the background model should be well-measured; the pixels where the difference is above a threshold are classified as foreground or objects of interest. Most of the methods use only pixel color or intensity information for this processing.

Fig. 1 Example of the detection and tracking of moving objects.

In our work, medication intake activities take place in an indoor environment and the background subtraction in the YCrCb color space is used to segment the moving objects. The YCrCb color space separates the intensity from the chrominance, so we can more easily eliminate shadows (that imply an intensity reduction with the same chrominance) from the moving object. The background is modeled by means (μ_{Cr}, μ_{Cb}) and standard derivation (δ_{Cr}, δ_{Cb}), trained in a specified period of time, without the presence of moving objects. In our study, a pixel is considered as foreground if its value is significantly different than the mean (absolute difference > 3δ) for one of the chrominance component.

B. Segmentation of skin regions

Skin region segmentation is the process of finding skin-colored pixels and regions in an image or a video sequence. This procedure is typically needed as a preprocessing step to find regions that potentially have human faces and limbs in images. A skin detector typically transforms a given pixel into an appropriate color space and then uses a skin classifier to label the pixel as a skin or non-skin pixel. A skin classifier defines a decision boundary for the skin color class in the color space based on a training database of skin-colored pixels.

Skin color detection methods can be classified into three different classes: methods using explicitly defined skin cluster boundaries, non-parametric methods and parametric methods [5]. The main advantages of the methods using explicitly defined skin cluster boundaries are the computing time, the simplicity and intuitiveness of the classification rules. However, the difficulty is the need to find both good color space and adequate decision rules empirically.

The non-parametric methods are simple for both training and classification, independent of distribution shape and therefore, to some extent, to the color space selection; however, they require much more storage space, a

representative training dataset and more computing time [5].

The parametric methods can be fast, have a useful ability to interpolate and generalize incomplete training datasets, are expressed by a small number of parameters and need very little storage space. However, they can be slow in training and their performance depends strongly on the skin distribution shape. Besides, most parametric skin modeling methods ignore the non-skin color statistics. This, together with dependence on skin cluster shape, results in higher error rate, compared to non-parametric methods [5].

In our work, we need a fast and real time execution so the segmentation of skin color using explicitly defined skin cluster boundaries is more convenient. It has been observed that under certain assumptions, the differences in skin color due to lighting conditions and ethnicity can be greatly reduced in the rgb (normalized) color space. Also, the skin-color cluster in the rgb color space has a relatively lower variance than the corresponding clusters in RGB space and hence the rgb color space was shown to be good for skin-color modeling and detection [6]. Notice that other color spaces could give comparable skin detection performance [15] but we chose the rgb color space because of its simplicity and rapidity of computation.

Fig. 2 Example of segmentation of skin regions.

1) Face detection

After thresholding and segmenting the skin regions, we can detect the face and hands in the source image. We detect the face in the first image and track it in successive images.

There are various solutions to the problem of face detection such as support vector machine [7], AdaBoost with Haar-Like features [8], neural network [9] etc. Most of these methods are concerned with frontal face detection. Among these methods, the one proposed by Viola and Jones [8] widely improved the speed and accuracy of face detection, which turns face detection into real practical applications.

In the context of medication intake, the detection of face in [1], [3] was done using geometric constrains such as the ratio of the major and minor axis length of the skin region. These constraints could lead to erroneous detection in some cases such as when a hand approaches the camera (and becomes larger in the image).

In our system, we detect the face in order to differentiate it from the hand regions by checking for eyes in skin regions. Our experiences have shown that the detection of eyes using AdaBoost is faster than the detection of frontal face using the same method [8]; this will be useful for our real time application.

2) Mouth region detection

The detection of the mouth region is necessary for the accurate recognition of pill intake, a step that was not tackled in [2] and [3]. The current approaches for lip segmentation can be categorized into four groups: color based, contour based, model based and learning based. In the surveillance of medication intake, the face is quite far from the camera, so the facial features are not very fine. Moreover, the real time processing is mandatory for the detection, tracking and recognition of medication intake activity. For these reasons, we used AdaBoost [10] for mouth detection in the first frame and a simple comparison of the R/G ratio for its tracking (see below).

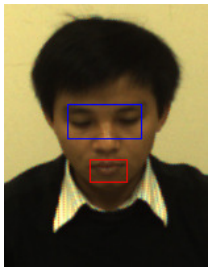


Fig. 3 Detected eyes and mouth using AdaBoost.

3) Hand detection

After localizing the skin region which represents the face, the remaining skin regions are assumed to be the hands. However, detecting and tracking hands becomes difficult when the person wears a short-sleeved shirt. In this case, it is necessary to identify the hand in the arm region.

In [3], the hand region is located by detecting the region with the highest density of contours. Each hand region is represented by a model, with information such as bounding box and density of contours. A constraint for hand width is applied in relation to the face width; unfortunately this constraint is not always true when the hand approaches the camera for instance.

In our previous work [11], we detected the hand region by calculating the local gray level variation, using the Sum-Modulus-Difference operator proposed in [12], also used in [1], and searched for regions having the maximum variation. This approach was computationally intensive and not convenient for real-time application. In this paper, we detect the hand region base on the contours, using the Sobel operator similarly to [3]; the square regions having the greatest density of contours are hand regions. To increase the performance of the system, the search for regions with greatest density of contours is done along the major axis of an ellipse fitted to the arm regions. The size of the (square) hand region is approximated by the size of minor axis of the ellipse.

C. Segmentation of medication bottles

The detection of bottles is already presented in [1], [2] and

[3]. In [1], a library of bottle images was used for detection and tracking by searching the Canny edge map at the start of each sequence for objects with rectangular shapes. The bottles were constrained of having a rectangular shape with height/width ratios of approximately 2:1.

In [3], the authors used the color histograms and the Hu moments [18] for detecting and tracking bottles. Color thresholds of the table on which the bottles will be put are also learned to avoid multi-scale searches and false detection. This approach requires that the bottles are always on a table background during the medication intake.

The detection of bottles in our work is inspired by the work presented in [2], and uses color information, except that simple thresholds are here used to speed up the detection. Each bottle is covered by a uniform color strip to facilitate its detection and also its identification by the elderly. The bottles can be anywhere in the scene. For each bottle, color examples are collected, under different condition of luminance, then the mean and standard derivation (s.d.) are calculated. Finally, the bottle detection (and tracking) is based on the color space normalized (with s.d.) distance to its mean color vector.

D. Objects representation

After segmenting skin regions and bottles, we need to represent these objects with different information. In the context of medication intake, each skin region is represented by: the skin region bounding box, the hand region bounding box, a Kalman filter for the bounding box centre of each hand region and the mouth region. For bottles, they are simply represented with their bounding boxes, and average color vectors and standard derivations.

IV. Tracking

Object tracking is an important task in many computer vision applications including surveillance, gesture recognition, smart rooms, vehicle tracking, augmented reality, video compression, medical imaging, etc. In surveillance system of medication intake, it is necessary to track skin regions, hand regions, mouth region, and bottles. We also need to handle object occlusions during the tracking process. In the following sub-sections, we present the tracking method for each object of interest and the robust occlusion handling.

A. Tracking of skin regions

In [1], feature and template matching are used to detect the face in the successive frames; unfortunately this is a costly operation, not adequate for real time applications. The hand regions are located after detecting the face. The Sum-Modulus-Difference sharpness method [12] is used to detect the hand regions.

The algorithm presented in [4] is used to track the head in [2]. The head is modeled with an ellipse whose size can vary from one frame to the other. For each image, a local search determines the best fitting ellipse, based on the gradient magnitude around its perimeter and the likelihood of skin color inside it. In addition, authors used the Continuously Adaptive Mean Shift Algorithm (Camshift) algorithm to track the hand regions. In [3], after locating the face in the first frame, the Hu moments are calculated and this information is used for tracking the face in the successive frames. The left hand is supposed to be always on the left and vice versa for the right hand.

In fact, the face tracking can be simpler. Here, we track skin regions for two cases: occlusion and non-occlusion. Tracking skin region with occlusion is presented in the section Occlusion handling below. For non-occlusion, considering that the frame-to-frame motion of skin regions is always relatively small, we simply use their displacement distances for tracking from one frame to the next. For hand regions, gray level histograms are also calculated and updated for occlusion handling (see below).

B. Mouth tracking

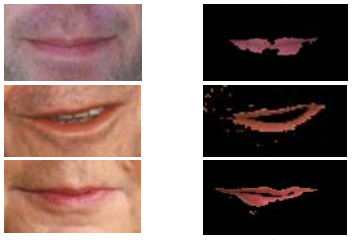
After detecting the mouth in the first image, it is tracked in consecutive images based on lip color. Hsu et al. [13] proposed a method to build a lip map, in the YCrCb color space. The main idea of this method is that, lip regions have high Cr and low Cb values. However, Nasiri [14] has shown that this map does not work properly for various kinds of images under various lightening conditions; so a Particle Swarm Optimization (PSO) was proposed to find out the optimal parameters for the mouth map construction proposed in [13].

Unfortunately the approach presented in [13] and [14] are time-consuming. We introduce a simpler method for segmenting the mouth region, using the ratio R/G in the RGB color space.

We observed that, in skin regions, pixels belonging to the lip region have their red component higher than non-lip pixels and experiments showed that it corresponds to a higher R/G ratio: this information is used to track the mouth region. The algorithm for mouth region tracking is described as follows:

- It defines a search region around the mouth thanks to a Kalman filter and the current bounding box of mouth.
- It calculates the mean $\mu_{R/G}$ and standard derivation $\delta_{R/G}$ for this search region.
- It classifies as lips, pixels with high *R/G* ratio with respect to $\mu_{R/G}$ and $\delta_{R/G}$.

Experiments have shown that this approach works well, even if the mouth is partially occluded by the hand. We have tested this method with several images taken from Internet and some results are presented in Fig 4.



(a) (b)

Fig. 4 The segmentation of mouth images.

The tracking of the mouth stops when the mouth region is occluded by the bounding box of a hand region. In this case, the center of the mouth bounding box is predicted using the Kalman filter and the mouth region is calculated using this prediction. This strategy is also used for the occlusion between the mouth and a glass of water.

C. Tracking medication bottles

The color of each bottle is represented by a mean color vector, with two components g and r in the rgb color space (recall that the b component is redundant since b=1-r-g). After segmenting the bottles in the first image, the tracking is done by comparing the distance to the mean color vector of non-skin moving object of interest. This approach speeds up the tracking in comparison with the accurate tracking of each bottle separately.

D. Occlusion handling

In the context of medication intake, occlusions between skin regions or between a skin region and a bottle are possible. The occlusion handling is divided in two cases: occlusion between skin regions and occlusion between bottles and other moving objects (skin regions, bottles).

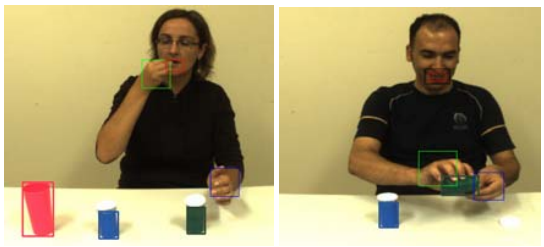


Fig. 5 Example of occlusion handling.

For the first case, a possible occlusion takes place when the number of skin regions, three in the previous frame, becomes less than three in the current frame. This means that two (or more) skin regions are overlapping in the current image if there are at least two centers of skin regions (bounding boxes) in the previous frame that are within the skin region bounding box in the current image. Otherwise, a skin region has simply left or disappeared from the scene. A skin region is not occluded if only its bounding box overlaps its previous image center.

In the case of hand-face occlusion, position of the hand in the occluded skin region is determined as follows:

- Define a search region, using a Kalman filter.
- Slide the hand bounding box window (search window) over the search region and calculate the gray level histogram in the search window.
- Search for the position where the histogram intersection between hand region and search region is maximal.

Hand-hand occlusions are treated similarly. The handling of occlusions involving bottles is done afterward.

When a medication bottle disappears, it is either hidden behind another bottle or a hand region (when the hand takes a bottle). The decision is simply made according to a minimum distance criterion. Finally, the bottle tracking (based on the color distance) resumes when it reappears (section III.C).

V. Recognition of Medication Intake

The recognition of medication intake based on the concept of scenarios was presented in [2]. Three levels of scenarios were proposed: single-state scenarios, multi-state scenarios and complex scenarios, which correspond to more or less complex activities. The recognition was accomplished by two automata corresponding to two possible ways of taking the medication. In [3], a Petri network was used in a similar manner for modeling the medication intake activities in a sequential way. The token transitions from a place to the other occur only if the corresponding event goes on during a specific minimum number of frames. Both methods were limited to ideal medication intake progression since the transitions between states were strongly defined for the different scenarios.

In this section, we present our hierarchal approach based on activity-patterns to recognize more realistic medication intake activities for different situations.

A. Medication intake activity hierarchy

We define the medication intake activity in a hierarchal way consisting of three levels: primitive activity, simple activity, complex activity, as described in Fig. 6.

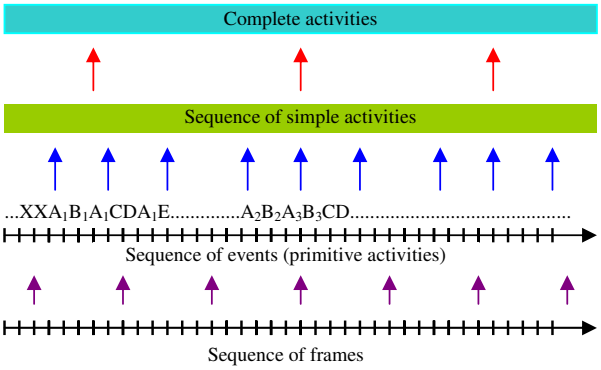


Fig. 6 Hierarchal approach for recognizing medication intake.

A primitive activity is the simplest activity (or event for brevity) and is one of the following for medication intake surveillance:

- A_i: one hand takes bottle i.
- B_i: two hands take bottle i, corresponding to the activity of opening/closing the bottle.
- C: one hand approaches the face.
- D: one hand touches the mouth region.
- E: one hand takes a glass of water that occludes the face region.
- X: non-interesting event, such as: a hand moves freely, a hand takes a glass of water, a hand occludes the other hand.

At the second level, we define simple activity of medication

intake as:

- A_iB_i: activity involving the manipulation of the bottle which corresponds to taking and opening the bottle; taking pill(s) from bottle; and closing the bottle.
- *CD*: activity related to the medication intake as such, with one hand bringing the pills to the mouth followed by (implicit) pill swallowing.

At the third and last level, we define the complete activity – the full medication intake process, which consists in opening and taking the medicine (A_iB_i) followed by swallowing pills ($A_iB_i \Rightarrow CD$).

B. Recognition of medication intake

As shown in Fig. 6, medication intake is represented in a hierarchal way. The recognition must be done in the same order, from low to high level. The resulting algorithm for real time recognition of medication intake activity is presented as below:

For each input frame

- *Detect an event in the input frame and add it to a sequence of current events.*
- *Compress the sequence of current events (see below).*
- *Recognize simple activity from the sequence of current events.*
- *Recognize a complete activity from the sequence of simple activities.*

End for

The following sub-sections present the recognition of medication intake in details from low to high level.

1) Recognition of primitive activity

At the low level, we have to detect events from the input frames; each frame represents a state of medication intake and has a corresponding event label. The event "one hand takes bottle *i*" is detected when there is an overlap between the hand bounding box and the bottle *i* bounding box. Similarly, when the bottle bounding box occludes the two bounding boxes of both hands the event "opening/closing bottle *i*" is detected. In the same way, the event "hand approaches the face" corresponds to the occlusion between the hand bounding box and face bounding box, "hand touches the mouth region", the occlusion between hand and mouth bounding boxes and "water drinking" the overlapping of the glass of water and face bounding boxes.

Notice that the activity of water drinking is an independent activity that is recognized directly as it occurs without further consideration in the next steps. Indeed, water drinking is not mandatory for medication intake detection.

Moreover, a primitive activity can be repeated in successive frames, so it is more efficient for the next step of recognition to "compress" repetition into only one event. In our work, the same consecutive events are therefore compressed in real time as illustrated in Table I.

TABLE I
AN EXAMPLE OF CONSECUTIVE EVENTS COMPRESSION

Frame order	Detected event	Sequence of current events	Sequence of compressed events
1	X	X	X
2	X	XX	X
3	A_1	XXA_1	XA_1
4	A_1	XXA_1A_1	XA_1
5	B_1	$XXA_1A_1B_1$	XA_1B_1
6	B_1	$XXA_1A_1B_1B_1$	XA_1B_1
7	C	$XXA_1A_1B_1B_1C$	XA_1B_1C
8	C	$XXA_1A_1B_1B_1CC$	XA_1B_1C
9	D	$XXA_1A_1B_1B_1CCD$	XA_1B_1CD
...	...	...	...

In a real situation, the medication intake activity may take different paths, with state repetition or not. The compression of repetitive events gathers the related events into smaller groups, which helps the recognition of simple and complete activities in the following steps. Moreover, this compression helps us identify activities that are not tackled in [1], [2], [3].

2) Recognition of simple activity

The activity-pattern A_iB_i represents the manipulation of a bottle. In our experimental setup, there are two bottles so A_iB_i can be: A_1B_1, A_2B_2. The other simple activity is *CD* representing the activity of taking pills to the mouth and swallowing them. Only the recognized simple activities are kept for the next steps and all others non-interesting events are bypassed.

3) Recognition of complete activity

The recognized simple activities of the above step are used here to check if the complete activity-pattern $A_iB_i \Rightarrow CD$ happens. Notice that the medication intake activity can take place in different ways, one pill at once or many pills together and in any order (see Fig. 7).

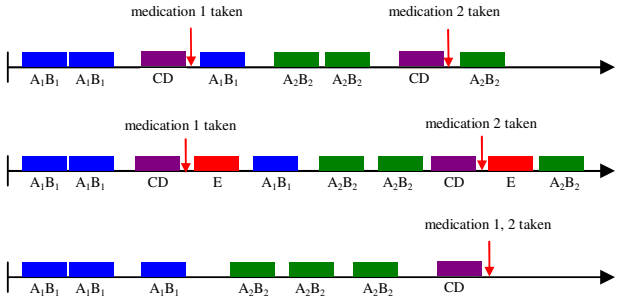


Fig. 7 Examples of the representation of medication intake (arrows) by mean of video sequence analysis

In our system, at any time the pill swallowing activity is recognized (*CD*), it is assumed that all the pills taken in previous steps (A_iB_i) are swallowed at this moment; therefore the recognition of activity-pattern $A_iB_i \Rightarrow CD$ is called recursively for this situation.

For each input simple activity of the same type A_iB_i, the first A_iB_i is supposed to be the bottle opening activity; the second A_iB_i, the pill taking activity and the third A_iB_i, the bottle closing activity.

4) Abnormal activities

To improve the robustness of our algorithm, abnormal situations are discarded based on the duration of activities in the following cases:

- A primitive activity (A_i, B_i, C, D, ...) lasts for a too long or a too short period of time (number of frames).
- The duration of a simple activity (A_iB_i or CD) is too long or too short.
- The relative time length of a simple activity is too long or too short with respect to the complete activity duration.

For instance there are situations where a hand (or two hands) occludes a bottle by chance for a short time (small number of frames) during the video acquisition, which does not correspond to neither bottle opening, nor bottle closing, nor pill taking. On the contrary, a hand (or two hands) can be placed in front of a bottle in such a way that from the camera point of view there is an too long occlusion to be considered as a valid event (or activity). Other situations that will lead to abnormal activities are: an elderly person who cannot open a bottle (A_iB_i without the corresponding CD), a person who touches his/her mouth for whatever reason such as scratching or eating (CD without the corresponding A_iB_i), etc. The abnormal activity parameters (duration thresholds) are computed after testing the system for the particular behavior of each elderly. This simple learning strategy allows adapting the system to each individual for better performance.

VI. Results

During a preliminary training step, the ranges of skin color (and appropriate thresholds) are experimentally determined for different types of skin color and indoor environment. A similar procedure is done for the medication bottles. The activity duration thresholds are also computed during this training (learning) phase. In practice, this step would be repeated for each elderly in his/her home setting for more accurate detection.

We captured the video sequence with a Prosilica camera [16], at a frame rate of 20 fps, with an image resolution of 659x493 pixels without video compression (for better image quality). The positions of the head, hands and bottles are automatically initialized as the person completely enters the camera field of view.

The detection, tracking and recognition of medication intake were tested with sequences taken with three different persons. Table II shows the results of detection and tracking.

As we can see in Table II, the detection and tracking of bottles was absolutely correct. Errors occurred with the mouth region at the moment of swallowing the medication, when the face suddenly moved upward while the mouth region was overlapped by the hand region, preventing the correct tracking of the mouth region. Fortunately, this does not influence the recognition of medication intake because we had already recognized it as soon as an overlap between the hand region and the mouth region took place. For hand tracking, errors happened when one hand region was hidden behind the (large enough) glass of water. Some inaccuracies were also observed when the hand overlapped the face.

TABLE II
SUCCESS RATE OF DETECTION AND TRACKING

Video Sequences	Number of frames	Hands region tracking	Bottles tracking	Mouth region tracking
Caroline: 17 sequences	2691	99%	100%	98%
Hung: 14 sequences	1937	100%	100%	97%
Mohamed: 11 sequences	2439	99%	100%	98%

TABLE III
SUCCESS RATE OF ACTIVITY RECOGNITION

Video Sequences	Open bottle and take pill	Swallow pill	Close bottle	Swallow water	Complete activity
Caroline sequence	98%	100%	100%	100%	98%
Hung sequence	100%	100%	100%	100%	100%
Mohamed sequence	98%	100%	100%	100%	98%

Table III shows the results of medication intake activity recognition. The activity of bottle opening/closing, and pill taking was recognized if there was an overlap between two hand regions and the bottle. In our experiences, the cap of each bottle was already opened and simply put on the bottle, so there was some rare situations where there was no overlap between these three regions because the opening/closing was simply too easy and too fast. In these cases the opening activity was not detected and therefore we could not recognize the medication intake activity.

TABLE IV
ACTIVITY STATISTICS (average # of frames and % of the complete activity)

Video Sequences	Open bottle & take pill	Swallow pill	Close bottle	Swallow water	Complete activity
Caroline sequences	10 (33%, 21% with water)	11 (36%, 23% with water)	8 (31%, 16% with water)	19 (40%)	29 (48 with water)
Hung sequences	6 (30%, 19% with water)	8 (46%, 26% with water)	4 (24%, 14% with water)	11 (18%)	18 (29 with water)
Mohamed sequences	10 (34%, 17% with water)	13 (37%, 24% with water)	10 (29%, 18% with water)	20 (41%)	33 (53 with water)

The duration statistics, estimated absolutely and relatively (with respect to the total duration of the complete activity) for each activity, is presented in Table IV. The results show that, different persons can take their medications at different speed (difference in number of frames) though similar values were obtained for the two first subjects in this study. Similar statistics (with their normal duration variations) were computed in the training phase for recognizing abnormal medication intake activities.

VII. Discussion and Conclusion

As much as 50% of all medication prescribed to seniors is used inappropriately and between nearly 30% of hospital admissions for patients over 50 years of age occur as a result of medication problems. Moreover, approximately 125,000 people with treatable illnesses die each year in the USA because they do not take their medication properly [17]. Several mechanical and automatic medication dispensers incorporating alarm and automatic opening and closing mechanisms have been developed to reduce medication errors. Despite their advantages, none of these devices is perfect. For instance, they do not check if the patient is actually taking the medication. In addition the elderly person is often annoyed by the alarm that alerts him regularly to take the medication according to the time set. This causes frustration and sometime abandon of the system.

Our computer vision system for monitoring medication intake solves these problems being passive most of the time except in case of medication error which would trigger an alarm. We have also proposed a three-level hierarchical approach that takes into account action durations and time laps between actions and the fact that medication intake can be done in many different ways.

The proposed solution for the detection of skin color region using color information, tracking skin region using minimum displacement distance, and detecting/tracking bottles using color information works very well in real time.

The recognition of medication intake based on activity-pattern hierarchy combined with real time compression of event repetition is effective and gives impressive results. Moreover, we can detect and recognize normal vs. abnormal medication intake activity from activity duration statistics in a training phase.

Notice that in this setup individual pills are not detected per se, since the camera resolution does not allow it. The detection of a sequence of gestures including the displacement of the hand up to the mouth confirms the medication intake. We also assumed that the camera is monitoring a medication area (e.g. table top) containing a number of medication bottles already in view, a realistic situation for elderly people. Finally, we expect that the elderly person will be collaborating (not trying to fool) with the system.

For future work, we plan to also investigate a two-camera system to have a better assessment of depth (with stereovision techniques) of the moving objects to better handle occlusions. For instance, during some particular gestures, if one hand appears in front of the mouth (or a bottle) without touching it, this could cause a false detection of medication intake (or bottle grabbing) with a single view. With a depth map, if the hand is at some distance (in front) of the mouth (or bottle), there will be no false detection.

References

[1] D. Batz, M. Batz, N. d. V. Lobo and M. Shah, “A computer vision system for monitoring medication intake,” The 2nd Canadian Conference on Computer and Robot Vision, pp. 362-369, 2005.

[2] M. Valin, J. Meunier, A. St-Arnaud and J. Rousseau, “Video surveillance of medication intake,” IEEE Engineering in Medicine and Biology Society, vol. 1, pp. 6396-6399, 2006.

[3] S. Ammouri and G.-A. Bilodeau, “Face and hands detection and tracking applied to the monitoring of medication intake,” Canadian Conference on Computer and Robot Vision, pp. 147-154, 2008.

[4] S. Birchfield, “Elliptical head tracking using intensity gradients and color histograms,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 232-237, 1998.

[5] V. Vezhnevets, V. Sazonov and A. Andreeva, “A survey on pixel-based skin color detection techniques,” GRAPHICON’03, pp. 85-92, 2003.

[6] J. Yang, W. Lu and A. Waibel, “Skin-color modeling and adaptation,” ACCV'98, pp. 687-694, 1998.

[7] E. Osuna, R. Freund and F. Girosi, “Training support vector machines: an application to face detection,” IEEE Conf. Computer Vision and Pattern Recognition, pp. 130-136, 1997.

[8] P. Viola and M. Jones, “Robust real-time object detection,” Int. J. Computer Vision, vol. 1, pp. 511-518, 2001.

[9] H. A. Rowley, S. Baluja and T. Kanade, “Neural network-based face detection,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, pp. 23-38, 1998.

[10] M. Castrillon-Santana, O. Déeniz-Suarez, L. Anton-Canalis and J.Lorenzo-Navarro, “Face and facial feature detection evaluation: performance evaluation of public domain Haar detectors for face and facial,” VISAPP’08, 2008.

[11] H.-H. Huynh, J. Meunier, J. Sequeira and M. Daniel, “Detection and tracking of interest regions for the surveillance of medication intake,” GRETSI'09, 2009.

[12] N. Ng Kuang Chern, P. A. Neow and M. H. A. Jr, “Practical issues in pixel-based autofocusing for machine vision,” IEEE Int. Conf. Robotics and Automation, vol. 3, pp. 2791-2796, 2001.

[13] R.-L. Hsu, M. Abdel-Mottaleb and A. K.Jain, “Face detection in color images,” IEEE Trans Pattern Analysis and Machine Intelligence, vol. 24, pp. 696-706, 2002.

[14] J. A. Nasiri, “A PSO tuning approach for lip detection on color images,” 2nd UKSIM European Symposium on Computer Modeling and Simulation EMS'08, pp. 278-282, 2008.

[15] B.D. Zarit, B. J. Super, F. K. H. Quek, “Comparison of five color models in skin pixel classification” Proceedings of the International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, pp. 58-63, 1999.

[16] www.prosilica.com

[17] C. Nugent et al., “Can technology improve compliance to medication?”, Int. Conference on Smart Homes and Health Telematic, Sherbrooke, QC, Canada, pp. 65-72, 2005.

[18] M. K. Hu, "Visual Pattern Recognition by Moment Invariants", IRE Trans. Info. Theory, vol. 8, pp. 179-187, 1962.