Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Video Surveillance of Medication Intake

Myriam Valin, Jean Meunier, Alain St-Arnaud and Jacqueline Rousseau

Abstract— **In the context of the growing proportion of seniors in the western world population and the efforts provided in home care services, we have developed a computer vision system for monitoring medication intake. The system detects automatically medication intake using a single low-cost webcam. Person detection and tracking over the video sequence is done using color-based techniques while the recognition of the medication intake activity is performed using our main contribution, a three-level scenario model. Experimental results in controlled conditions are shown and we discuss improvements to our system.**

I. Introduction

In Canada, more than 30% of all people aged 65 and over live alone [1], with the associated risks for their health. According to the *Public Health Agency of Canada*, 18% to 50% of all medication used by seniors is used inappropriately and between 19% and 28% of hospital admissions for patients over 50 years of age occur as a result of medication problems [2]. Moreover, as mentioned in [3], approximately 125 000 people with treatable ailments die each year in the USA because they do not take their medication properly. A video surveillance system for monitoring seniors medication intakes could help reducing the risks as well as family member concerns.

The system presented here is part of a vaster project of video surveillance for home care services for seniors. It has been developed for monitoring medication intake in the context of a person living alone and whose medications are in bottles. In order to be as accessible as possible, a single low-cost webcam is used. The system has therefore to deal with low-quality digital images.

Recently, efforts have been made on developing computer vision systems for monitoring medication intake [4]. One problem with this approach is that the skin segmentation needs to be very precise to detect effectively occlusions, hand positions and orientations and face regions, which cannot be reached with low-quality images. Moreover, the authors consider that the medication intake activity occurs if a sequence of actions, verified every frame, is observed, disregarding the action durations and time laps between actions. Thus, the number of false alarms might become very high in real home situations. Because of the expected lack of tracking precision and the complexity of the activity to recognize, we developed a more complex algorithm to improve medication intake activity recognition, which forms our main contribution.

II. Overview

In order to perform medication intake monitoring, the system must first detect if the person is taking his medication. The proposed method is inspired by that in [5]. In our method, processing is divided in two parts:

- Low level processing: moving objects (regions) are detected and tracked at every frame (section III).
- High level processing: activity recognition is performed based on moving object characteristics, using our three-level scenario model (section IV).

When medication intake is detected, the used bottle and the time of detection are recorded. Results are presented in section V and system limitations and future work are discussed in section VI.

III. Detection and Tracking

In our approach, three types of mobile objects are tracked: the person's head and hands and the medication bottles. Fig. 1a shows these objects. In this experimental setting, three medication bottles are present.

The head tracking algorithm is described in [6]. The head is modeled as 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. The gradient magnitude of a pixel corresponds to the rate-of-change of intensities in the gray-scale image over a small local neighborhood. The color histogram used to compute skin color likelihood includes the person's hair color since the head might be turned or leaned.

Hands are positioned based on regions with high skin color likelihood. These regions are extracted from the skin color likelihood map (Fig. 1b) created during the head tracking process. Possible occlusions between the head and the hands or between both hands are also dealt with based on previous positions and a few assumptions.

Since we want to detect which medication is being taken and labels cannot be identified automatically in low resolution video, color bands are affixed on the bottles to better differentiate them.

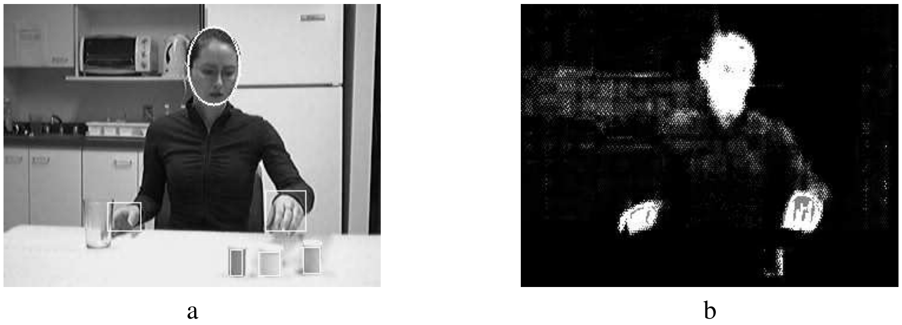
The detection of medication bottles is done using color models as described in [7]. Pixels classification is performed based on their Mahalanobis distance to the color models,

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

M. Valin and J. Meunier are with the Biomedical Engineering Institute and the Department of Computer Science and Operations Research, Université de Montréal, Montreal, Canada (e-mail: {valinmyr,meunier}@iro.umontreal.ca).

A. St-Arnaud and J. Rousseau are respectively with the Centre de santé et de services sociaux Lucille-Teasdale and the Research Center of the Institut Universitaire de Gériatrie de Montréal, Montreal, Canada (e-mail: astarnaud@ssss.gouv.qc.ca, jacqueline.rousseau@umontreal.ca).

1-4244-0033-3/06/$20.00 ©2006 IEEE.



a

b

Fig. 1. a) Mobile objects detection and tracking and b) corresponding skin color likelihood map.

which are created from training samples. Possible bottle regions (objects) are extracted using connected component analysis. Bottle positioning is done considering the size of these objects and bottle previous positions, since bottles can be completely occluded by the hands.

IV. Activity Recognition

Activity recognition is based on the concept of scenarios, which correspond to long-term activities of mobile objects. There are three levels of scenarios: single-state scenarios, multi-state scenarios and complex scenarios. Fig. 2 shows how these scenario levels are related. The concepts of single-state and multi-state scenarios have been developed in [5] and are briefly described here.

A. Single-state Scenarios

Single-state scenarios are defined by a set of mobile object properties and verified at every frame. For example, the scenario *"the person touches his head"* depends on the distances between the hands and the head. For each scenario, an all-or-none probability of occurrence is determined heuristically using logical functions. The single-state scenarios in our system are defined as follows:

- S_1: Exactly one hand manipulates a bottle,
- S_2: Both hands manipulate a bottle,
- S_3: One hand touches the head,
- S_4: One hand approaches the head,
- S_5: One hand moves away from the head,

For each scenario, the implicated mobile objects are kept.

In [5], the authors present another way of computing probabilities by using Bayesian classifiers and learned conditional probabilities of the mobile object properties, but this did not prove to be necessary for this application.

B. Multi-state Scenarios

A multi-state scenario corresponds to a sequence of single-state scenarios and is verified over a longer sequence of frames. A multi-state scenario (*MS*) is recognized if all its single-state scenarios are recognized consecutively.

The main idea is to compute the maximum probability that, at time t, the sequence of N single-state scenarios occurs given a sequence of observations O. This can be expressed as follows:

$$P(MS^*|O) = \max_{(t_1,t_2,\ldots,t)} P(S_{1_{(t_1,t_2-1)}} S_{2_{(t_2,t_3-1)}} \ldots S_{N_{(t_N,t)}}|O), \quad (1)$$

where $S_{i_{(t_i,t_{i+1}-1)}}$ means single-state scenario i occurs between frames t_i and $t_{i+1}-1$. t_1 and t are respectively the first and last frames of the multi-state scenario in the video sequence. The observations O correspond to the mobile object properties. The algorithm then consists of finding the best transition times t_i from one state to the next so that the whole sequence is completed with the maximum probability.

In the algorithm, $P(S_{i_{(t_i,t_{i+1}-1)}})$ is approximated by the expected recognition value of $S_{i_{(t_i,t_{i+1}-1)}}$ which is the mean of $P(S_{i_t})$ for $t_i \leq t \leq t_{i+1}-1$ and whose value is between 0 and 1.

The system developed for medication intake is divided in three multi-state scenarios:

- MS_1: The person opens a medication bottle and takes the pill(s),
- MS_2: The person swallows the pill(s),
- MS_3: The person closes the medication bottle,

each scenario having a probability of occurrence and starting and ending frames.

C. Complex Scenarios

Multi-state scenarios can represent simple activities, but do not work with more complex situations. For example, the activity of medication intake can be done in many different ways: the person can take the pills out of the medication bottle, swallow the pills and then close the medication bottle or take the pills out of the medication bottle, close the medication bottle and then swallow the pills. Moreover, actions not related might happen during the sequence, such as drinking water or putting the pills on the table. Finally, if more than one medication are being taken, the simple states which form the entire sequence might not occur consecutively. Thus, activity recognition cannot be done by searching for a fixed sequence.

We solve this problem by splitting the complex scenario into a sequence of multi-state scenarios (states) and then use a *"state-transition"* system to recognize the complex activity.

Once the multi-state scenarios are detected, before pushing the states into the state-transition system, a *"likelihood"*

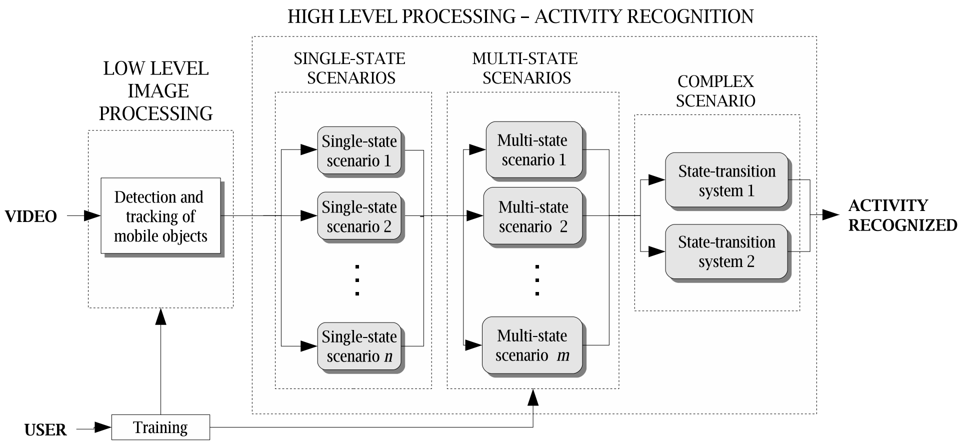


Fig. 2. Overview of the system.

factor is used to penalize the scenarios that are too long (or too short) compared to their mean durations, or lengths, by decreasing their probability. This factor is defined by a Butterworth filter transfer function and is expressed as follows:

$$f(\ell) = \frac{1}{1+\left(\frac{\ell-\bar{\ell}}{w_c}\right)^{2n}} \tag{2}$$

where $\bar{\ell}$ is the mean scenario length, which varies from one multi-state scenario to the other and is manually defined. w_c and n (order of the filter) are fixed. As shown in Fig. 3, this function does not penalize recognition for small variations of duration around the mean. The resulting multi-state scenarios are then sorted according to their starting frame.

The two possible representations of the complex scenario used are shown in Fig. 4. The circles represent the multi-state scenarios and the arrows represent the transitions. Loops over the same state are possible because more than one medication can be taken. Since pills are not tracked specifically and more than one might be swallowed at the same time, the same event MS_2 might be used to recognize the medication intake for all pills.

Transitions are used to penalize the recognition if the number of frame increases too much between two states. Penalization is done only between states so the recognition of the intake of medication 1 does not fail if the person takes pills out of another medication bottle before swallowing the pills from bottle 1. The *"likelihood"* factor used here is also defined by another Butterworth filter transfer function. However, instead of depending on the state lengths, it depends on the transition lengths. The parameters have been set by hand after some experimentation, but could easily be learned automatically.

V. Results

The skin and bottle color models used for tracking have been computed in a first training step from hand labeled examples. The video acquisition is done at 15 frames per second with an image resolution of 320x240 pixels. The positions of the head, hands and bottles are automatically initialized as the person completely enters the camera field of view. We tailored the initialization procedure to the experimental conditions. Since the efforts were directed toward the recognition part, we do not extend on the performances of the tracking system. We just report here that it is good enough to support high-level activity recognition.

The detection of medication intake has been tested with 41 sequences from 26 videos taken with three different persons. 31 sequences show real medication intake and the 10 others show other activities (lures), like eating or manipulating the bottles.

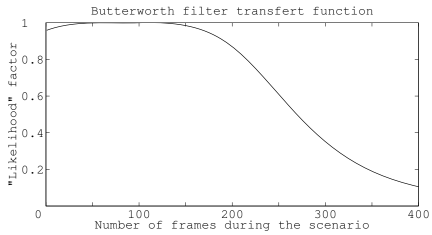
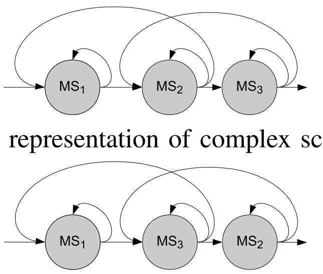


Fig. 3. Butterworth filter transfer function for $\bar{\ell} = 85$, $w_c = 184$ and $n = 2$ used for MS_1.



a) First representation of complex scenario.

b) Second representation of complex scenario.

Fig. 4. State-transition systems representing complex scenario formed by multi-state scenario sequences a) $\{MS_1, MS_2, MS_3\}$ and b) $\{MS_1, MS_3, MS_2\}$.

Typical results are presented in Fig. 5. Fig. 5a shows a sequence of multi-state scenarios obtained from a video with medication intake. The results are 99.6% detection for both bottles. The complex scenario is defined by its second representation, shown in Fig. 4b.

Fig. 5b shows a sequence of complex scenarios involving three bottles. We can see that a detection of medication intake with the bottle 3 overlaps all the other scenarios. This has happened because the bottle was manipulated at the beginning of the sequence. The low probability shows the effectiveness of the penalization, which allows to reject the false detection.

Considering that a scenario is recognized if its probability is over 80%, results for all video sequences are summarized in Table I. In fact, most of the scenarios recognized had a probability over 90%.

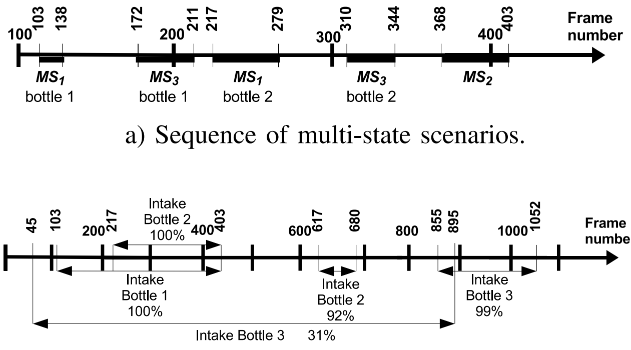
A major source of errors is occlusions between mobile objects. Indeed, if a mobile object passes in front of an other one, from the camera 2D point of view, the two objects are in (false) contact. Since contacts between hands and medication bottles form the core of states MS_1 and MS_3, errors can affect the whole recognition process. To limit this problem, the camera has been placed a little higher than the person's head, decreasing, this way, the number of occlusions.

In the actual system, the tracking part does not work in real time but it has not yet been optimized. Since the recognition part can be done after the person left the table, computation time is fast enough to be used in a continuous application.

VI. Conclusion

Having described our experiments, we now discuss possible issues in less controlled environments. First, detection and tracking errors could be due to:

- low-quality of the webcam images resulting in high noise level and difficulties with pixel color classification,
- presence of objects similar to the medication bottles (same color and size),
- presence of occluding objects between the camera and the person.



a) Sequence of multi-state scenarios.

b) Sequence of complex scenarios.

Fig. 5. Typical results : a) sequence of multi-state scenarios forming the complex scenario $\{MS_{1_1}, MS_{3_1}, MS_{1_2}, MS_{3_2}, MS_2, MS_2\}$ and b) sequence of complex scenarios involving multiple bottles with corresponding probabilities.

TABLE I
Recognition results for 41 sequences.

Activity	Detected	Not Detected
Medication Intake	28	3
Other (lures)	1	9

Because activity recognition depends heavily on tracking of mobile objects, any enhancement of the tracking algorithm would be beneficial. For example, work could be done on finding a better skin color model, performing 3D positioning (to limit false contacts) and using a mouth-to-hand distance (instead of a head-to-hand distance) to test the S_3 scenario.

Second, changes in a person's normal behavior of medication intake could also impair recognition. For example, errors could happen if :

- the person touches his forehead instead of his mouth in the medication intake sequence (the mouth is not specifically tracked in our actual system),
- the person moves away and comes back between opening the medication bottle and swallowing his pills.

We have to mention here that, in this application, we consider that the person agrees to take his medication and that noncompliance is due to memory problems rather than nonacceptance.

Considering these possible sources of errors, we expect that adaptations should be needed for the medication intake monitoring system to work within real home environments. However, the presented algorithms have shown to be effective and promising.

We have tested the activity recognition process for medication intake, but it should be applicable to other everyday home activities.

References

[1] C. Lindsay. "A Portrait of Seniors in Canada," Third Edition. Ottawa: *Statistics Canada*, 1999.

[2] Committee of Officials (Seniors) for the Ministers Responsible for Seniors, "Working together on seniors medication use: a federal/ provincial/ territorial strategy for action," *Public Health Agency of Canada*, Ottawa, June 1996.

[3] C. Nugent et al., "Can technology improve compliance to medication?," in *Proc. 3rd Int. Conf. On Smart homes and health Telematic*, Sherbrooke, QC, Canada, 2005, pp. 65–72.

[4] D. Batz, M. Batz, N. da Vitoria Lobo and M. Shah, "A computer vision system for monitoring medication intake," in *Proc. IEEE 2nd Canadian Conf. on Computer and Robot Vision*, Victoria, BC, Canada, 2005, pp. 362–369.

[5] S. Hongeng, R. Nevatia and F. Bremond, "Video-based event recognition: activity representation and probabilistic recognition methods," *Computer Vision and Image Understanding*, vol.96, no. 2, Nov., pp. 129–162, 2004.

[6] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Proc. IEEE Computer Vision and Pattern Recognition*, Santa Barbara, CA, USA, 1998, pp. 232–237.

[7] N. Habili, C. Lim and A. Moini, "Hand and face segmentation using motion and color cues in digital image sequences," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, Tokyo, Japan, 2001, pp. 377–380.