An Intelligent Surveillance System for Dangerous Situation Detection in Home Environments

Rita Cucchiara
DII – Università di Modena e
Reggio Emilia
Via Vignolese, 905 - 41100
Modena – Italy
cucchiara.rita@unimo.it

Andrea Prati
DII – Università di Modena e
Reggio Emilia
Via Vignolese, 905 - 41100
Modena – Italy
prati.andrea@unimo.it

Roberto Vezzani
DII – Università di Modena e
Reggio Emilia
Via Vignolese, 905 - 41100
Modena – Italy
vezzani.roberto@unimo.it

ABSTRACT

In this paper we address the problem of human posture classification, in particular focusing to an indoor surveillance application. The approach was initially inspired to a previous works of Haritaoglou et al. [5] that uses histogram projections to classify people's posture. Projection histograms are here exploited as the main feature for the posture classification, but, differently from [5], we propose a supervised statistical learning phase to create probability maps adopted as posture templates. Moreover, camera calibration and homography are included to solve perspective problems and to improve the precision of the classification. Furthermore, we make use of a finite state machine to detect dangerous situations as falls and to activate a suitable alarm generator. The system works on-line on standard workstations with network cameras.

Keywords

Home surveillance, posture detection, projection histograms, learning.

1. INTRODUCTION AND RELATED WORKS

Home automation, environmental control and tele-assistance are three hot keywords in current research activities of the computer engineering community. The emerging technologies can offer a very interesting contribution in improving the quality of the life of people in the house, and especially of people with some form of disability.

The research in computer vision on people surveillance joint with the research in efficient remote multimedia access makes feasible a complex framework where the people in the house could be monitored in their daily activities in a fully automatic way, therefore in total agreement with privacy policies. A well formalized set of alarm situations can be defined and can be used as the trigger of some actions such as the communication to remote users, control center or private person. Finally, only in such a situation, remote users can connect also with low-cost devices, such as GPRS phones or PDAs.

In this context, we are developing an integrated framework that aims at exploiting the potentiality of automatic video annotation for detecting dangerous events and, therefore, reacting by sending an alarm to a control center. A set of computer vision and motion analysis techniques are used to extract objects and events from the scene, according with a previously defined ontology. In this specific case, moving people are the main objects of interest, while, through body modeling and posture recognition, a state transition-based high-level reasoning module is able to extract interesting events, such as the fall of the monitored person.

Recently, an increasing number of computer vision projects dealing with detection and tracking of human posture have been proposed and developed. An exhaustive review of proposals addressing this field was written by Moeslund and Granum in [7], where about 130 papers are summarized and classified according with several taxonomies. In particular, they consider three different application fields: video surveillance, control and pure analysis. Our proposal can be included in the first class.

According with [7], we can classify most of them into two basic approaches to the problem. From one side, some systems (like Pfinder [9] or W^4 [6]) use a direct approach and base the analysis on a detailed human body model. In many of these cases, an incremental predict-update method is used, retrieving information from every body parts. Nevertheless, these systems are generally too sensitive, when losing information about features, needing often a reboot phase. For this reason, the segmentation process has to be as much accurate as possible using specific human cues (e.g., detecting the skin regions). Unfortunately, this can often make the system unreliable, because some of these features could not be found in every frame (due to overlapping, for example).

In order to bypass these drawbacks, where no body part control is necessary, many researchers deal with the problem in a indirect way using less, but more robust, information about the body. Many of these approaches are based on human body silhouette analysis. The work of Fujiyoshi et. al. [4] uses a synthetic representation (Star Skeleton) composed by outmost boundary points. In [5] Haritaoglu et al. add to the W^4 framework [6] some techniques for human body analysis using only information about silhouette and its boundary. They first use hierarchical classification in main and secondary postures, processing vertical and horizontal histogram profiles (or *projections*) from the body silhouette. Then, they locate body parts by analyzing the silhouette boundary's corners.

Our approach is similar to [5], as well as concerning projections features, but, differently from it, it is not based on a priori defined model. In fact, the main strength of our approach is a machine learning phase, exploited to create feature models, further used in the classifier.

In this paper, the proposed framework is presented in Section 2. Section 3 gives details of our proposal of probabilistic templates of projection histograms and Section 4 shows some tests on real videos. Conclusions end the paper in Section 5.

2. THE PROPOSED FRAMEWORK

In order to be as flexible as possible, we constraint the problem's dimensionality to the 2D case, reducing the analysis to single camera images. Therefore, we suppose to have a point of view where the human posture can be easily perceived without ambiguities also in the image plane. Indeed, we exploit a 2D ½ space, by computing homography after camera calibration in order to have an approximate position of the person in the 3D space.

We consider video from indoor environment with a calibrated fixed camera. In this paper we do not address problems of people overlapping, neither problems of occlusion of human parts, and we consider only one person at a time in the room, as we did in a previous work [2] in which we used a tracking with probabilistic and appearance model as in [8].

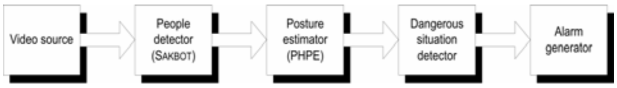

Figure 1. A simplified scheme of the overall framework

The overall framework is composed by several modules, as shown in Fig. 1. In particular three internal processing modules are in pipeline: a *people detector* [1] that produces for each frame the list of people present in the scene; a *posture estimator* used to compute the current posture for each person detected in the previous step; a *dangerous situation detector* able to identify anomalous situations by means of temporal analysis of the people posture and to generate a suitable alarm.

The posture estimator module provides posture classification frame by frame. In this work, no past information are exploited to identify the current posture. We consider only four main postures: standing (a), sitting (b), crouching (c), and lying down (d). This module will be accurately described in the next section.

The posture history is used in the *dangerous situation detector* module. In particular, in Fig. 2 the Finite State Machine used for the recognition of a fall is shown. A person is assumed to be detected in standing position at the beginning (in an indoor environment the people often enter in the rooms standing). The transition between *moving* and *static* states (see Fig. 2) depends on the average motion of the tracked blob, while the transition between *moving* states is guided by changes in the posture classification. After a too long permanence in the laying down static state an alarm is generated. Note that there is no connection from standing to laying states: this transition can produce a warning because probably the system is failing.

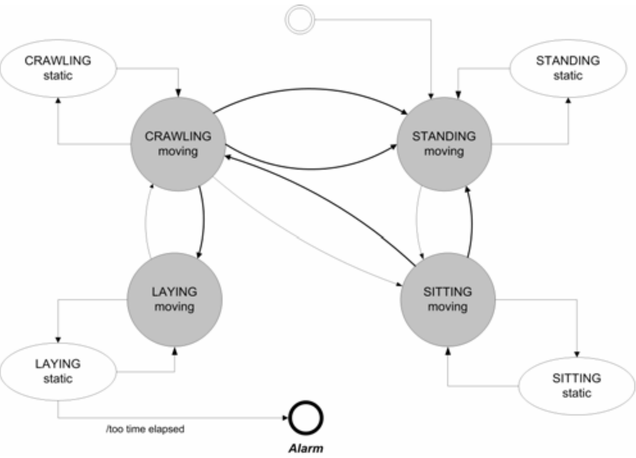

Figure 2. Finite State Machine for the identification of a fall

The alarms corresponding to dangerous situations (in our case the falls) can be managed in several ways. For example, a control center can be advised and connected through a video-audio link with the assisted person. Besides, a vocal message or a SMS can be sent to a relative or a neighbor on their cell phone or PDA, and, in this last case, a link for a low-bandwidth video connection can be provided in order to assert person's conditions. Finally, all the events can be saved on a database for further processing.

3. PROJECTION HISTOGRAMS BASED POSTURE ESTIMATOR

In order to recognize human body posture, we used a knowledge classification inspired to the one proposed by [5]. As aforementioned, we discriminate four main postures. Since the silhouette of people sitting with a frontal, left or right view are very different, internally the system splits each state in three view-based subclasses: *front-view*, *left-view* and *right-view*. The block diagram of the *Projection Histograms-based Posture Estimator* (PHPE) is reported in Fig. 3.

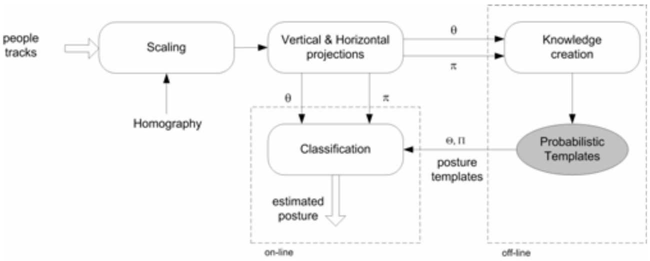

Figure 3. Block diagram of the PHPE module

Differently from [5] we use camera calibration to scale detected blobs in order to have a posture classifier invariant to people position in the room. In fact, the main lack of the direct use of the silhouette of the people to estimate the posture is the dependence on the distance between people and camera. The work [5] suggests a normalization phase after the computation of the projection histograms, but, in this way, without a relation with the 3D space, many errors can arise: for instance, the two postures *standing* frontal and *crouching* frontal can be easily confused. To this aim we exploit camera calibration to compute the distance *d* of the people from the camera and we use this value to scaling the tracks. More details can be found in [3].

3.1 Projection histograms and probabilistic templates

We start with a blob *B* as a cloud of 2D points contained into a bounding box with size *(Bx,By),* representing the points internal to the person silhouette. As in [5] we define horizontal and vertical projections (respectively π and θ) the following cardinality (indicated with #) of horizontal and vertical sub-sets:

$$\theta(x) = \#\left\{(x_p, y_p)\in B \middle| x_p = x\right\} \text{ where } x\in[0, Bx-1] \tag{1}$$

$$\pi(y) = \#\left\{(x_p, y_p)\in B \middle| y_p = y\right\} \text{ where } y\in[0, By-1] \tag{2}$$

Fig. 4(a) shows an example of histograms computed.

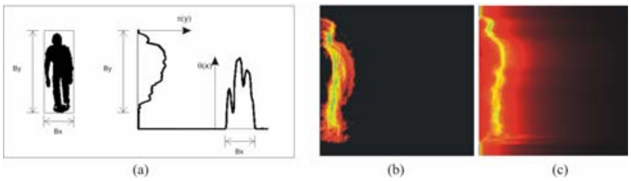

Figure 4. Example of projection histograms and probability maps for *Standing* frontal posture. Brighter colors correspond to higher probabilities.

We defined a learning phase where the system constructs two probability maps (horizontal and vertical one) for every posture (class), using the respective silhouette projections. A training set of T_i 2D blobs referred to the *i*-th class is given for each *k* posture class:

$$B_t^i = \left\{(x,y), x\in[0, Bx_t^i - 1], y\in[0, By_t^i - 1]\right\}, t = 1..T_i \tag{3}$$

For each B_i^t the couple $P_i^t = (\theta_i^t, \pi_i^t)$ of projection histograms is computed as in equations (4) and (5). We construct the couple $\Theta_i(x,y)$ and $\Pi_i(x,y)$, (with $x\in[0,Bx-1]$, $y\in[0,By-1]$) of 2D probability density maps of the i-th state as follow:

$$\Theta_i(x,y) = \frac{1}{T_i}\sum_{t=1}^{T_i} g(\theta_i^t(x), y) \quad \Pi_i(x,y) = \frac{1}{T_i}\sum_{t=1}^{T_i} g(x, \pi_i^t(y)) \tag{4}$$

where

$$g(s,t) = \frac{1}{|s-t|+1} \tag{5}$$

The density functions $\Theta_i(x,y)$ and $\Pi_i(x,y)$ describe a priori conditional probabilities: $\Theta_i(x,y)$ is the probability (conditioned to the fact of being in the status i) that the histogram θ has $\theta(x) = y$, that means, in other words, to have *y* points in the blob at the coordinate *x*. Similarly for $\Pi_i(x,y)$. Fig. 4(b) shows two probability maps computed for the posture *standing* frontal. The function in eq. (5) is used to avoid misclassifications due to sparse probability maps. Each point of a map is increased according with its distance from the histogram. The number 1 at the denominator is inserted to avoid dividing by zero.

In the classification phase the computed projection histograms are compared with the probability maps stored into the system. For each posture *i*, a measure of similarity S_i is extracted. In [5] the reference templates are histograms instead of maps, and S_i is computed with a logarithmic likelihood formula. By making that, all the points of the template histograms have the same weight. Differently, with the adoption of the probability maps, the most reliable parts of the histograms are highlighted. Let $\tau_i = (\Theta_i, \Pi_i)$ be a probability map couple for the class *i* and $P = (\theta, \pi)$ the projection silhouette of the track to be classify. We consider the two similarity values S_i^θ and S_i^π obtained as:

$$S_i^\theta = \frac{1}{B_x}\sum_{x=0}^{B_x-1}\Theta_i\left(x, \theta(x)\right) \qquad S_i^\pi = \frac{1}{B_y}\sum_{y=0}^{B_y-1}\Pi_i\left(\pi(y), y\right) \tag{6}$$

The final score S_i is computed as the correlation between the two scores and the posture estimator module selects as final posture the one with the maximum final score S_i.

$$S_i = S_i^\theta \cdot S_i^\pi \tag{7}$$

4. EXPERIMENTAL RESULTS

The system has been designed to meet real-time constraints and to process a sufficient number of frames per second to be reactive and adaptive enough for possible alarm. Using standard workstations connected with a network camera we are able to process about 10 fps. Here we report some results on videos acquired in different contexts (Table 1).

Luca1	Luca2	Roberto1	Roberto2
1209 frames	735 frames	294 frames	416 frames
320 x 240	320 x 240	360 x 288	384 x 288

Table 1. Frames from the videos adopted

In particular, we describe three tests:

- the *efficacy* in posture detection *over the same video where the training phase has been performed*: these results could be interesting for a domotic surveillance application, supposing that an initial training is done in the specific context on the specific person, as a sort of initial calibration of the system (Table 2) ;
- the *efficacy* in posture detection by using a *different training set* obtained from the *same camera system* (Table 3, first row);
- the *efficacy* and the generality of the model in posture detection *on different videos* (different camera, different scene, different actors) w.r.t. the training set (table Table 3, second row).

Video	Frames	Correct	Wrong	Efficacy %
Luca1	1209	1167	42	96,53%
Luca2	735	723	8	98,37%
Roberto1	294	289	1	98,30%
Roberto2	416	412	4	99,04%

Table 2. Efficacy rate for test 1

Video	Frame	Correct	Wrong	Efficacy %
Luca1	1209	1197	13	99,01%
Roberto2	416	387	29	93,03%

Table 3. Efficacy rate for tests 2 and 3; the template exploited in the test is obtained with Luca2

The system exhibits a quite robustness (about 90%) in every test. Principally, the two mistaken postures are *standing* and *crouching*, because the transitions between them are very difficult to classify also for human observer.

5. CONCLUSIONS

The paper discusses initial results of detecting human posture for surveillance and behavior monitoring in home environments. The discussed approach proved to be reliable and robust if the working constraints are satisfied.

In conclusion, in this paper we present an intelligent surveillance system, improved by an initial learning phase and able to detect dangerous situations for people in their home. The system will be capable to allow low-bandwidth live connection with the remote site and to communicate with audio output with the people in the home in order to have a feedback of the recognized situation. This can be the basis for future paradigms of home-human interactions that will improve autonomy of people with some disabilities, increasing their safety and at the same time allowing programs of tele-rehabilitation.

This work is supported by the project “Domotics for disability” of “Fondazione Cassa di Risparmio di Modena” (Italy). The authors are also thankful to Luca Panini for his valuable help in ground-truth tests.

6. REFERENCES

[1] Cucchiara, R., Grana, C., Piccardi, M., Prati, A.: "Detecting Moving Objects, Ghosts and Shadows in Video Streams", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, n. 10, pp. 1337-1342, 2003.

[2] Cucchiara, R., Grana, C., Prati, A., Vezzani, R.: “Computer Vision Techniques for PDA Accessibility of In-House Video Surveillance”, in press on *ACM Multimedia 2003 - First ACM International Workshop on Video Surveillance*, Berkeley (CA), USA, Nov. 2-8, 2003

[3] Cucchiara, R., Prati, A., Vezzani, R.: “Domotics for disability: smart surveillance and smart video server" in *8th Conference of the Italian Association of Artificial Intelligence - Workshop on "Ambient Intelligence"*, Pisa, Italy, pp. 46-57, Sep 23-26, 2003

[4] Fujiyoshi, H., Lipton, A.J.: “Real-Time Human Motion Analysis by Image Skeletonization”. In *Fourth IEEE Workshop on Applications of Computer Vision*, pp. 15-21, 1998.

[5] Haritaoglu, I., Harwood, D., Davis, L.S.: “Ghost: A Human Body Part Labeling System Using Silhouettes”. In *Proc. Fourteenth International Conference on Pattern Recognition*, Brisbane, pp. 77-82, vol. 1, August 1998

[6] Haritaoglu, I., Harwood, D., Davis, L.S.: “W^4: Real-Time Survelliance of People and Their Activities”. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809-830, August 2000.

[7] Moeslund, T.B., Granum, E.: “A Survey of Computer Vision-Based Human Motion Capture”. *Computer Vision and Image Understanding*, vol. 81, pp. 231-268, 2001.

[8] Senior, A., Hampapur, A., Tian, Y.-L., Brown, L., Pankanti, S., and Bolle, R.: “Appearance models for occlusion handling.” *Proc. Second International workshop on Performance Evaluation of Tracking and Surveillance systems*, 2001.

[9] Wren, C.R., Azarbayejani, A., Darrel, T., Pentland, A.P.: “Pfinder: Real-Time Tracking of the Human Body”. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780-785, July 1997.