Assessing Temporal Coherence for Posture Classification with Large Occlusions

Rita Cucchiara, Roberto Vezzani
D.I.I. - University of Modena and Reggio Emilia - Italy
{*cucchiara.rita, vezzani.roberto*}*@unimore.it*

Abstract

In this paper we present a people posture classification approach especially devoted to cope with occlusions. In particular, the approach aims at assessing temporal coherence of visual data over probabilistic models. A mixed predictive and probabilistic tracking is proposed: a probabilistic tracking maintains along time the actual appearance of detected people and evaluates the occlusion probability; an additional tracking with Kalman prediction improves the estimation of the people position inside the room. Probabilistic Projection Maps (PPMs) created with a learning phase are matched against the appearance mask of the track. Finally, an Hidden Markov Model formulation of the posture corrects the frame-by-frame classification uncertainties and makes the system reliable even in presence of occlusions. Results obtained over real indoor sequences are discussed.

1. Introduction

This work focuses on the problem of people posture classification in cluttered environment, with large and long-lasting occlusions of the people. This is typical of indoor surveillance, when the camera is placed close to the targets and the moving people are often occluded by other people or static objects (e.g., desks, chairs). Our goal is to detect the most probable posture also when the people's silhouette is not completely visible by assessing the temporal coherence of the data with probabilistic models. The temporal coherence is taken into account at three levels:

i) the motion of a person and of its body's parts is not rigid and thus is not easy predictable with basic tracking methods; nevertheless it is continuous and without abrupt changes: thus probabilistic tracking techniques are particularly suitable to integrate appearance coherence along the time [14, 7]; ii) the motion of single body's parts (e.g. the feet) could be predictable in a short period; the prediction with the temporal coherence of their position can be exploited in case of occlusions; iii) the classification of the posture cannot be provided frame by frame. Even if probabilistic models based on a learning phase are adopted, an instant classification is reliable enough only when the people visual appearance is good enough. In case of occlusions and posture transitions the classification must be reenforced with temporal reasoning.

Accordingly, in this paper we present an integrated approach mixing probabilistic and appearance based tracking with Kalman filter. Kalman is used in case of occlusions to predict the position of people's feet in order to provide an adequate shape scaling in a calibrated environment. Thus, the instant posture estimation is computed with probabilistic posture maps (PPMs) integrated in a Hidden Marcov Model (HMM) to reinforce the temporal coherence of the model. The novel approach has been tested in a working indoor surveillance system: the adoption of HMMs with mixed probabilistic and Kalman tracking improves the robustness of posture classification in most of the situations.

2. Related work

Recently, an increasing number of computer vision projects deals with detection and tracking of human posture. An exhaustive review of proposals addressing this field is the work of Moeslund and Granum in 2001[11], where about 130 papers are summarized and classified according with several taxonomies.

The posture classification systems proposed can be differentiated by the more or less extensive use of a 2D or 3D model of the human body [11]. From one side, some systems use a *direct* approach and base the analysis on a detailed human body model: effective examples are Pfinder [17], W^4 [6], and the Cardboard Model [9]. In many of these cases, an incremental predict-update method is used, retrieving information from every body part. Many systems use complex 3D models, with special equipment, such as 3D laser scanners [16], or multiple video cameras [10]. They are often too expensive for many indoor surveillance applications. A second way consists in an *indirect* approach that, whenever the monitoring of single body parts is not necessary, exploits less, but more robust, information about the body. Most of them extract a minimal set of low level features and exploit them inside suitable classifiers. One

Proceedings of the IEEE Workshop on Motion and Video Computing (WACV/MOTION'05)
0-7695-2271-8/05 $ 20.00 IEEE

frequent example is the use of neural networks, as in [4]. However, the use of NN presents several drawbacks due to scale dependency and unreliability in the case of occlusions. In [13], a Universal EigenSpace approach is proposed: this presents insensitivity to clothing, but it assumes that most of the possible postures (with most of the possible occlusions) have been learned, and this is far from being realizable.

A large class of approaches are based on human silhouette analysis. The work of Fujiyoshi et. al. [5] uses a synthetic representation (Star Skeleton) composed by outmost boundary points. A similar approach is proposed in [1] where a skeleton is extracted from the blob by means of morphological operations and then processed using a HMM framework. This approach is very promising and has the characteristic of also classifying the motion type, but it is very sensitive to segmentation errors and in particular to occlusions. Moreover, no scaling algorithm to remove perspective distortion is proposed making this approach unfeasible for our target application.

In [6], Haritaoglu et al. add to W^4 framework some techniques for human body analysis using only information about the silhouette and its boundary. They first use a hierarchical classification in main and secondary postures, processing vertical and horizontal projection histograms from the body's silhouette. This histograms are normalized to a fixed size without exploiting any 3D information. Then, they locate body parts on the silhouette boundary's corners. In [3] the posture classification with projection histograms is improved by a learning phase to generate probabilistic posture maps. Posture classification is provided frame by frame and verified with a deterministic state transition graph. The transition time is not related to the probability of the previous state and to the similarity of the current projection histograms with respect to the templates. Thus, it cannot cope with many conditions of uncertainness during posture transitions. Another approach based on silhouette analysis is reported in [8] where a 2D complex model of the human body is matched with the current silhouette by genetic algorithms. In addition to the problems of segmentation errors and occlusions, this approach also suffers from dependency of the model on the view. Instead, the scaling of the model in the 3D environment is mandatory, since the people posture recognition is strictly related with the size and the proportions of the body shape.

3. System Overview

The defined approach is composed by several operators as in Fig. 1. The people posture classification asks for the computation at each frame t and for each person in the scene of its blob, its appearance image, and a measure of occlusion probability. The blobs can be segmented with standard background suppression algorithms, as [7, 2, 15]. Since in indoor environments shadows and background changes frequently occur, we included in the segmentation process a shadow suppression module and we adopt a dynamically adaptive model of the background. The objects larger enough and satisfying some geometrical requirements are classified as potential people. In case of occlusions and frequent posture changes, people posture cannot be reliably provided with the static information extracted in the current frame only. Conversely, we aim at exploiting the temporal coherence of the posture. Thus, we suppose to have, after the segmentation and the tracking, in addition to the current blob B (Fig. 2.b), the *appearance image AI* (or *temporal template*) and the *probability mask* of the track (Fig. 2.c). AI is obtained with a temporal integration of the color images of the blobs, while the probability mask associates to each point of the map a probability value (between 0 and 1) that indicates its reliability (See Fig. 2.c). Probabilistic and appearance based tracking is frequently adopted for objects with non rigid motion and variable shape, like people [14, 7, 3]. Even if the probabilistic tracking is out of the scope of this paper, we briefly describe the algorithms we implemented. In particular, at each frame the known tracks are matched against the segmented visual objects. The best track correspondence is computed with a Gaussian function [14], that measures the similarity in the RGB space between pixels of the track and of the blob. The measure of similarity is also exploited to estimate the occlusion probability that we call *visibility*. Then the probability mask P_M and the appearance image AI are updated pixel by pixel. For not occluded pixels x of the tracks, $P_M(x)^t = \lambda \cdot P_M(x)^{t-1} + (1-\lambda)$ and $AI(x)^t = \lambda \cdot AI(x)^{t-1} + (1-\lambda) I(x)^t$. For not visible pixels $P_M(x)^t = \lambda \cdot P_M(x)^{t-1}$ and $AI(x)^t = AI(x)^{t-1}$. $I(x)^t$ is the color value of the frame in x and λ is a value lower than one (e.g. $\lambda = 0.9$). In case of occlusion the probability

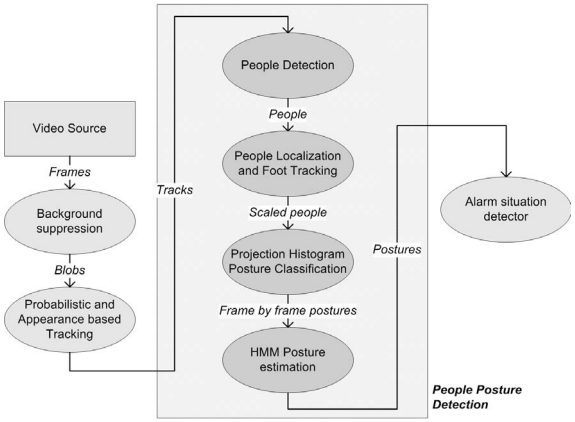

Figure 1. Block diagram of the system

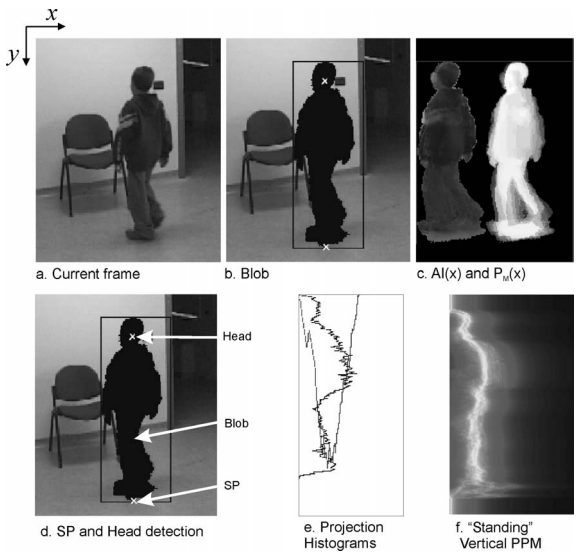

Figure 2. Posture detection phases

mask is "freezed" for each pixel x ($P_M(x)^t = P_M(x)^{t-1}$).

Posture classification is provided exploiting appearance features of the silhouettes and, in particular, the two projection histograms, that are matched against posture probabilistic models. The projection histograms $\theta(x)$ and $\pi(y)$ describe how the blob's shape is projected onto the x ad y axes respectively, as in Fig. 2.e.

Projection histograms are simple and coarse features, but discriminant enough if the entire blob is visible. In addition, their computation is very fast, suitable for real time surveillance. However, these descriptors suffer from three limitations: i) they are too sensitive to the unavoidable non-rigid movements of the human body; ii) they depend on the silhouette's size; iii) they become not reliable in presence of occlusions. The first problem is partially solved with the adoption of a learning process capable of generalizing the peculiarity of a training set of postures. The second drawback is overcame with an initial scaling correction in 3D, improved with a Kalman-based prediction of the support point (SP). The support point is the contact point between the person and the floor; it is used to obtain the 3D position of the person exploiting the homography relations of the floor with respect to the camera plane (known by means of the calibration). Although we adopt an enhanced tracking and probabilistic posture models, when occlusions occur the lower level modules could be affected by errors and noise. To make more reliable the output of the posture classification we have introduced a further step, represented by a HMM framework that models the slowness of the transitions between two postures and, moreover, measures the *reliability* of the posture detected (through evaluating the differences between the HMM state probabilities).

The computed posture, the measure of its *reliability* by means of HMM, and the position of the people and their heads in the 3D environment can be further processed to assess the indoor situation, to generate alarms or to log the people behavior.

In the next sections the details of these steps are analyzed with a special focus on the occlusion handling.

4. Support Point Tracking

The correct estimation of the *SP* position also during an occlusion is a key requirement for a correct people scaling. The *SP* is normally coincident with the foot position, but they could differ when the person is lying down on the floor. If the camera has no roll angle and the pan angle is low enough, the support point can be estimated taking the maximum y coordinate and a mean of the x coordinates of the lowest part of the blob (see Fig. 2.d). This simple algorithm fails in presence of occlusions: if the occlusion interests the feet of the person, in fact, the y coordinate of the SP becomes not valid. The SP detection cannot be provided neither on the blob nor on the probability mask: the first because is uncomplete and the second because is unprecise and unreliable especially in the body parts with frequent motion as the legs are. To this aim, we have introduced a dedicated tracking for the SP coordinates: two independent constant-velocity Kalman filters (for the two components x and y respectively) are adopted. The first is implemented in a standard manner, while the Kalman filter for the y coordinate of SP, is used in two modalities: when the person is *visible* the filter considers both the forecast and the measure (as usually), while when the person is *occluded* the measure of the y coordinate is ignored and the filter exploits only the forecast to estimate the current position.

The two filters have the same parameters and exploit the constant velocity assumption. Using the well-known notation of the discrete Kalman filter

$$\begin{aligned} x(k+1) &= \mathbf{\Phi} \cdot x(k) + v \\ z(k) &= \mathbf{H} \cdot x(k) + w \end{aligned} \tag{1}$$

the matrixes adopted are reported in Eq. 2.

$$x(k) = \begin{pmatrix} pos_k \\ vel_k \end{pmatrix} \quad H = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix} \quad Q = \begin{pmatrix} 0 & 0 \\ 0 & \gamma \end{pmatrix}$$

$$z(k) = \begin{pmatrix} pos_{k-1} \\ pos_k \end{pmatrix} \quad \Phi = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad R = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \tag{2}$$

The Measurement Noise Covariance Matrix $R(k)$ (of the gaussian variable w) is computed assuming that the two measured positions could be affected by a noise that is independent frame by frame and time constant, while the Process Noise Covariance $Q(k)$ of the variable v assumes that

Figure 3. a) *SP* estimation. b) Trajectories obtained with the blob (black) and with the Kalman Filter (white)

the process noise affects only the velocity terms. In our implementation, we set $\lambda = 200$ and $\gamma = 0.5$. The results obtained by the introduction of this two independent filters during a strong occlusion are visible in Fig. 3.

Known the SP position and the homography of the floor plane, we can compute the distance between the person and the camera. Supposing that the image of the person entirely lies on a plane parallel to the camera, we can project the blob and the tracking images into a metric space by a linear scaling procedure.

5. Posture Classifier

The set of postures recognized by our system is:

$$Post = \{S_j | j = 1..np\} = \{ST_F,\ ST_L,\ ST_R,\ CR_F,\ CR_L,\ CR_R,\ SI_F,\ SI_L,\ SI_R,\ LA_F,\ LA_L,\ LA_R\} \quad (3)$$

that is obtained by splitting four $Main\_postures = \{\boldsymbol{ST}anding, \boldsymbol{CR}ouching, \boldsymbol{SI}tting, \boldsymbol{LA}ying\}$ into three view-based subclasses (**F**rontal, **L**eft-headed, and **R**ight-headed).

The frame by frame posture estimation, at each time t compares the *projection histograms* ($PH^t = ((\vartheta(x); \pi(y))^t$) (Fig.2.e) of each tracked person with the np Projection Probabilistic Maps ($PPM^j = \left(\Theta^j(x,s), \Pi^j(t,y)\right), j = 1..np$) (an example of PPM for Θ in standing position is in Fig.2.f) created through a supervised machine-learning phase. The probabilistic approach included in the PPMs allows us to filter the useless moving parts of the body (such as the arms and the legs) that can generate misclassifications of the posture. The PPMs, created with a training set T including images of different actors in different postures, integrate the histograms as follows:

$$\Theta^j(x,s) = \frac{1}{\|T^j\|} \cdot \sum_{i \in T^j} g\left(\theta_i(x), s\right) \quad (4)$$

$$\Pi^j(t,y) = \frac{1}{\|T^j\|} \cdot \sum_{i \in T^j} g\left(t, \pi_i(y)\right) \quad (5)$$

where $\left\|T^j\right\|$ is the number of frames T^j with posture j in T and $g(i,j)$ is a generalization function such as:

$$g(i,j) = \frac{1}{|i-j-|+1}. \quad (6)$$

Projection histograms are normally computed over the current blob B [7, 3]. Like for the support point computation, in case of occlusions the blob is uncomplete and the projection histograms could be unreliable. In this case, the appearance masks are adopted instead of the blobs to compute the histograms. The appearance masks, in fact, keep a recall of the occluded parts of the person's silhouette. The disadvantage of the appearance masks is a possible lower reaction in the detection of posture changes, because in such situations the transition is perceivable only after the adaptation of the mask. Nevertheless, it is an affordable cost considering that such an approach can correctly classify the posture of an almost completely occluded person, as in the Fig. 4.

Given the Projection Histograms PH of a blob B (or, as in our case, an appearance image AI), the output of the comparisons with the np PPMs, corresponding to the S_j postures, is a vector of probabilities $\mathbf{P} = \{P_j, j = 1..np\}$:

$$P_j = P(PH|S_j) = P(\vartheta_B|S_j) \cdot P(\pi_B|S_j) = \prod_x \Theta^j_B(x, \vartheta_B(x)) \cdot \prod_y \Pi^j_B(\pi_B(y), y) \quad (7)$$

As a frame by frame posture classification we can select the posture S_i such that $i = \underset{j=1..np}{argmax}\, P_j$.

5.1. Temporal Coherence Posture Classifier

Although the improvements given by the use of appearance mask instead of blobs, a frame by frame classification is not reliable enough. A posture transition graph could modelled with deterministic transition graphs. However, the introduction of probabilistic models can guarantee the temporal coherence of posture transition.

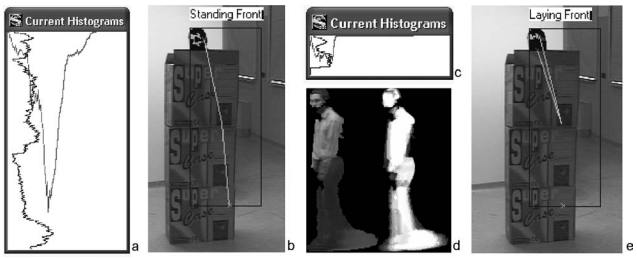

Figure 4. PH computed on the Appearance Image (a,b,d) and on the blob only (c,e).

In fact, the transition time must be related to the probability of the previous states and to the similarity of the current projection histograms with respect to the templates. For instance, if the current position is very similar to a particular posture, the switch to the corresponding state must be faster than the case in which all the posture are equally probable.

A typical solution is given by Hidden Markov Models. Using the notation adopted by Rabiner in his famous tutorial [12], we define the followings sets:

- *The set* $\mathbf{S}$*, composed by four states:*

$$\mathbf{S} = \{S_1, S_2, S_3, S_4\} = Main\_Postures \tag{8}$$

- *The initial state probabilities* $\mathbf{\Pi} = \{\pi_i\}$: the initial probabilities are set equal for each state ($\pi_i = \frac{1}{4}$). To introduce the hypothesis that a person enters a scene upright, it is possible to set the vector $\mathbf{\Pi}$ with all the elements equal to 0 except for the element corresponding to the standing state (set to the value 1). However, the choice of the values assigned to the vector $\mathbf{\Pi}$ affects the classification of the first frames only, and then it is negligible for our purpose.

- *The matrix* $\mathbf{A}$ *of the state transition probabilities:* computed as a function of a reactivity parameter α. We have considered the probabilities to remain in the same state and to pass to another state equal for each posture. Then, the matrix $\mathbf{A}$ has the following structure:

$$\mathbf{A} = \mathbf{A}(\alpha) = \{A_{ij}\}, A_{ij} = \begin{cases} \alpha & i = j \\ \frac{1-\alpha}{3} & i \neq j \end{cases} \tag{9}$$

In our system we use $\alpha = 0.9$.

To complete, the Observation Symbols and Observation Symbol Probability distribution $\mathbf{B}$ have to be defined. As observations symbols we assume the projection histograms. Even if the observation set is numerable, we do not compute explicitly the matrix $\mathbf{B}$, but we estimate frame by frame the values b_j for each state j as

$$b_j = P_j = P(PH|S_j). \tag{10}$$

Then, at each frame, we compute the probability of being in each state with the forward algorithm [12].

At last, the HMM input has been modified to keep into account the visibility status of the person. In fact, if the person is completely occluded, the appearance image is only a remember of the person and its posture reliability must decrease with the time. In such a situation, we could set $b_j = \frac{1}{N}$ as the input of the HMM. In this manner, the state probability tends to a uniform distribution (that models the increasing uncertainty) with a delay that depends on the previous probabilities: higher the probability to being in a state S_j is and higher is the time required to lose this certainty. To menage simultaneously the two situations and to cope with the intermediate cases, (i.e., partial occlusions), we have define a generalized formulation of the HMM input:

$$b_j = P(PH|S_j) \cdot \frac{1}{1 + n_{fo}} + \frac{1}{N} \cdot \frac{n_{fo}}{1 + n_{fo}}. \tag{11}$$

where n_{fo} is the number of frames for which the person is occluded. If n_{fo} is zero (i.e., the person is visible) b_j is computed as in Eq. 10, otherwise it tends to a uniform distribution with the increasing of n_{fo}.

6. Experimental results

The system has been tested in a prototype working online for indoor surveillance and over several indoor videos, in which different actors interact each others and with the furniture generating occlusions. In particular, the results of the system over two videos (named "Boxes" and "TwoActors") are reported.

The video "Boxes" shows a person walking behind a stack of boxes (Fig. 3); the corresponding results are reported in Fig. 5. During the occlusion (highlighted in Fig. 5.c) the person's blobs are reduced to the head only and the classification frame by frame obtained with the computation of the projection histograms directly over the blobs is wrong (Fig. 5.a). In Fig. 5.b, instead, are reported the results obtained computing the projection histograms over the appearance maps and adopting the HMM framework described in the previous section. In addition to a correctly classification of the postures, the HMM framework also indicates the reliability of the output: when the person is visible, the temporal coherence reinforces the probability to be standing disadvantaging the other postures, while during the occlusion the classification uncertainty increases.

In the video "TwoActors" the people interact each other (Person1 walks behind the television at frame 100 and then ahead Person2 at frame 500) and change their posture (in particular Person2 falls down at the frame 1000). The results of the system over this video are reported in Fig. 6. The

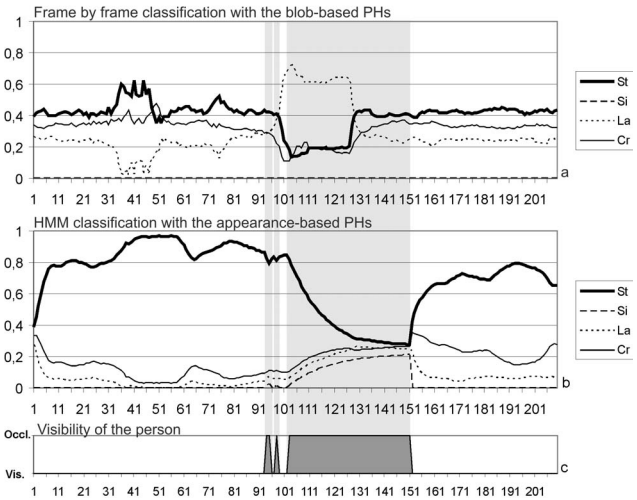

Figure 5. Results on the video "Boxes"

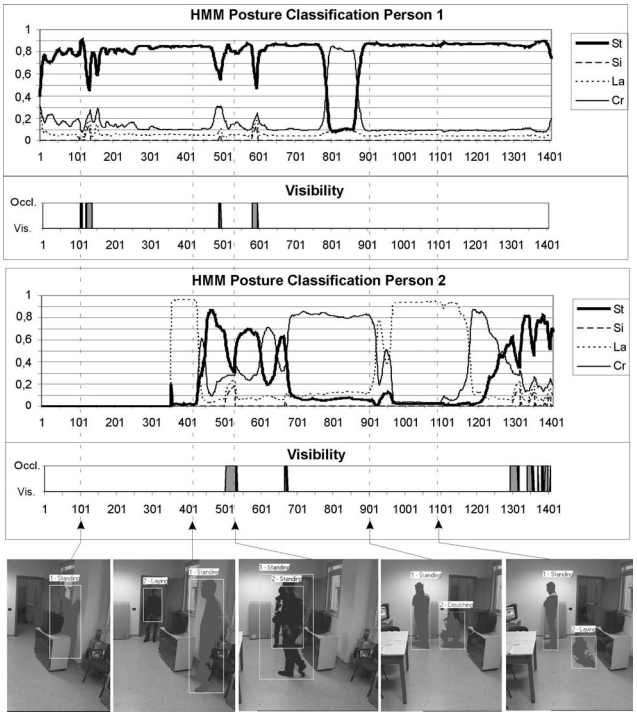

Figure 6. Results on the video "TwoActors".

posture estimated are always right, except during the entering of the Person2 (frames 350-400), when the low-level module wrongly segments the blob.

7. Conclusions

In this paper a People Posture Classification especially developed to cope with occlusions is presented. The occlusions are solved exploiting the temporal coherence of the people appearance, of their position and of their posture. A probabilistic object tracking (to cope with object splitting and merging), a Kalman filter (to estimate the correct position of the person inside the room), and an HMM formulation of the classification phase (to take into account the classification uncertainties) make the system reliable even in presence of occlusions. The HMM slows down the reactivity at frame level, but solves many uncertain cases and gives a measure of the classification reliability. The model can be extended with other postures or improved by adopting other low level features.

References

[1] I.-C. Chang and C.-L. Huang. The model-based human body motion analysis system. *Image and Vision Computing*, 18(14):1067–1083, Nov. 2000.

[2] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts and shadows in video streams. *IEEE Trans. on PAMI*, 25(10):1337–1342, Oct. 2003.

[3] R. Cucchiara, C. Grana, A. Prati, G. Tardini, and R. Vezzani. Using computer vision techniques for dangerous situation detection in domotics applications. *Proc. of IEE Intelligent Distributed Surveillance Systems (IDSS-04)*, pages 1–5, Feb. 2004.

[4] J. Freer, B. Beggs, H. Fernandez-Canque, F. Chevriet, and A. Goryashko. Automatic recognition of suspicious activity for camera based security systems. In *Proc. of European Convention on Security and Detection*, pages 54–58, 1995.

[5] H. Fujiyoshi and A. Lipton. Realtime human motion analysis by image skeletonization. In *Fourth IEEE Workshop on Applications of Computer Vision*, 1998.

[6] I. Haritaoglu, D. Harwood, and L. Davis. Ghost: a human body part labeling system using silhouettes. In *Proc. of Int'l Conf. on Pattern Recognition*, volume 1, pages 77–82, 1998.

[7] I. Haritaoglu, D. Harwood, and L. Davis. W4: real-time surveillance of people and their activities. *IEEE Trans. on PAMI*, 22(8):809–830, Aug. 2000.

[8] C. Hu, Q. Yu, Y. Li, and S. Ma. Extraction of parametric human model for posture recognition using genetic algorithm. In *Proc. of IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 518–523, 2000.

[9] S. Ju, M. Black, and Y. Yacob. Cardboard people: A parameterized model of articulated image motion. In *2nd Int'l Conf. on Automatic Face and Gesture Recognition*, 1996.

[10] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisition and tracking using voxel data. *Int'l Journal of Computer Vision*, 53(3):199–223, July 2003.

[11] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, Mar. 2001.

[12] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. of the IEEE*, volume 77, pages 257–286, Feb. 1989.

[13] M. Rahman, K. Nakamura, and S. Ishikawa. Recognizing human behavior using universal eigenspace. In *Proc. of Int'l Conf. on Pattern Recognition*, pages 295–298, 2002.

[14] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, and R. Bolle. Tracking people with probabilistic appearance models. In *Proceedings of International Workshop on Performance Evaluation of Tracking and Surveillance (PETS) systems*, 2002.

[15] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on PAMI*, 22(8):747–757, Aug. 2000.

[16] N. Werghi and Y. Xiao. Recognition of human body posture from a cloud of 3d data points using wavelet transform coefficients. In *Proc. of IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 70–75, 2002.

[17] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: real-time tracking of the human body. *IEEE Trans. on PAMI*, 19(7):780–785, July 1997.