2004 8th International Conference on
Control, Automation, Robotics and Vision
Kunming, China, 6-9th December 2004

Human posture recognition in video sequence using Pseudo 2-D Hidden Markov Models

Lim Hock Wyi Aloysius*
*National University of Singapore
3 Science Drive 2, Singapore 117543
hockwyi@yahoo.com
huangzy@comp.nus.edu.sg

†Guo Dong
†DSO National Laboratories
20 Science Park Drive
Singapore 118230
gdong@dso.org.sg

*Huang Zhiyong ‡Tele Tan
‡Curtin University of Technology
GPO BOX U1987 Perth
Western Australia, Australia 6845
teletan@cs.curtin.edu.au

Abstract— **This paper describes a video surveillance system capable of recognizing human postures from video sequences. The system comprises of two key modules: human detection and posture classification. In the module of human detection, human blobs are extracted by the technique of background subtraction. An adaptive background model is employed to characterize the dynamics and complexity of outdoor scenes based on the *Mixture of Gaussians*. In order to formulate the variations of human postures, Pseudo 2D Hidden Markov Models (P2DHMM) is employed for representing and recognizing human postures based on its '2-D elastic matching' property. It is trained to differentiate human postures and tolerate the variations of the same human posture using *Embedded Viterbi* and *Segmental K means* algorithms. In the classification of human postures, observation sequence is extracted from current image frame. The probabilities of observation sequence corresponding to each P2DHMM model are computed by *Doubly Embedded Viterbi* optimization, and human blob is classified as the human posture with the highest likelihood.**

I. INTRODUCTION

Video surveillance has been a vital instrumentation for detecting illegal or suspicious activities, where recorded images can be of critical value to the investigators if such activities occurred. Often, reliance on human discernment is accurate but is unreliable and tedious, since human's concentration is not consistent. With the rise of awareness in security, there arises a need for a reliable and 'intelligent' system harnessing the best outputs from the latest computer vision techniques. Such a system is able to provide an early warning for threat events based on the recognition of specified suspicious human activities.

A. State of the art

There have been two successful systems for detecting and tracking the moving objects, e.g. Detection of Events for Threat Evaluation (DETER) [1] and DARPA Video Surveillance and Monitoring (VSAM) [2]. In DETER, the threat is detected by observing the trajectories of moving objects. The VSAM combines multiple sensors covering a wide area of surveillance into a coherent system, where moving objects are detected and tracked.

The detection of moving objects is essential for video surveillance applications. Background subtraction is an efficient and sensitive approach for the detection of unusual motion in the fixed scene by comparing incoming frame to the model of scene background. For modeling outdoor background scenes, dynamic background subtraction are widely used as outdoors constantly undergoes fast and slow changes in the background. In [3], a real-time tracking system *Pfinder* was built for reliable human detection in relatively static indoor environment. The background scene was modeled by Gaussian-based unimodel based on YUV color components. It is updated by a simple adaptive filter to compensate the lighting changes. In [4], The on-line mixture of K Gaussians is applied to model the variation in the background. It is robust for illumination changes, repetitive motions of scene elements, and has well detection in the cluttered region. In [2], a hybrid background model was proposed based on three-frame differencing and adaptive background subtraction. The model is simple and fast in the detection of moving object, but it is inadequate for illumination changes and repetitive motions of scene elements. The non-parametric estimation was used to estimate the probability of pixel values based on the accumulated sample in [5]. The model adapted quickly for the scene changes, but a huge amount of memories were required to accumulate

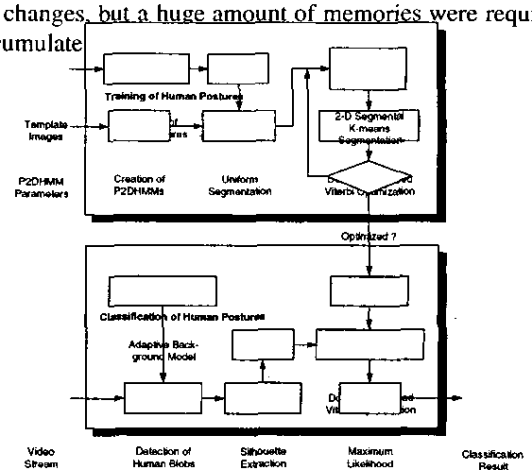

Fig. 1. Architecture of Video Surveillance System

The recognition of human activities is a classification process, whose output is merely a decision regarding the class

0-7803-8653-1/04/$20.00 © 2004 IEEE

to which each activity belongs. In [6], human is identified based on body shape and gait. The cyclic gait analysis is used to extract key frames from various sequences, and subject classification is performed by nearest neighboring match and correlation scores. *Recurring Motion Image* (RMI) method [7] attempts to model and identify activities based on the repetitive changes of blobs. Several human activities with violence are detected based on motion trajectory and limb orientation in [8]. The human head is tracked for the detection of jerk based on temporal derivative of *Acceleration Measure Vectors* . The silhouette information of human posture is used to identify the activity of armed robbery in [9]. The classic holdup position of armed robbery is recognized by skeleton segmentation and the identification of arm position. As human activities are quite diverse, simple classification, e.g. thresholding, is sufficient enough for the classification of human activities. The *Hidden Markov Model* (HMM) technique is an effective tool used for pattern classification due to its robustness and rich mathematical structure. In [10], HMM is trained by the edge information of segmented regions. It works well for occluded objects and across different scales but not for non-rigid (e.g. humans) and rotated objects. Another method uses ergodic 2D HMM [11] to model silhouettes is rotational invariant but is highly complex due to its use of fully connected 2D HMMs. Assuming dimensional distortions are independent each other, *Pseudo 2D Hidden Markov Model* (P2DHMM) was employed for face recognition with the promising performance in [12].

B. Description of Problem

In current video surveillance systems, much work has been carried out on background modeling, change detection and object tracking. The recognition of human activities, however, has not gained enough attentions. A video surveillance system is proposed in the paper, which is able to detect unusual human activities from video sequences based on the modules of human detection and activity classification. The paper is organized as follows: In section 2, we discuss the architecture of our proposed system. The modules of background subtraction and activity recognition are presented in section 3 and 4 respectively. Finally, experimental results are presented in section 5, followed by the conclusions.

II. Overview of System

The video surveillance system consists of two modules, i.e. *Training of Human Postures* and *Classification of Human Postures*. It is illustrated in Figure 1. The adaptive background model is used to detect moving objects from video stream. In our system, outdoor scene is modeled by the mixture of Gaussians, it is robust for illumination changes, repetitive motions of scene elements, and has well detection of moving objects in the cluttered region. The perimeter control is used to extract human blobs from moving objects based on the aspect of human body and position of footprint. The classification of human activities is implemented by P2DHMM model based on silhouette features of human postures. It is trained by the image sample related to different human activities.

III. Background Subtraction

We use Mixture of Gaussians method [4] for adaptive background subtraction. Each pixel is modeled by a mixture of K Gaussian distributions, where K is determined by computational power. In our case, $K = 3$. The probability of observing the current pixel in current frame is given by,

$$P(X) = \sum_{i=1}^{K} \omega_i \eta(X, \mu_i, \Sigma_i) \quad (1)$$

where η is a Gaussian density function, ω_i is the estimated fraction of Gaussian distribution, μ_i is the mean value and Σ_i is the covariance matrix of the ith mixture at current frame. If the pixel value is matched to one of Gaussian models (i.e. pixel value is within 3σ of μ of the ith Gaussian model. The parameters of this Gaussian model is updated by,

$$\mu_t = (1-\rho)\mu_{t-1} + \rho(M_{k,t}) \quad (2)$$

$$\sigma_t^2 = (1-\rho)\sigma_{t-1}^2 + \rho(X_t - \mu_t)^T(X_t - \mu_t) \quad (3)$$

$$\omega_{k,t} = (1-\alpha)\omega_{k,t-1} + \alpha(M_{k,t}) \quad (4)$$

where ρ and α are constant learning parameters, and $M_{k,t}$ is set to 1 for matched model and 0 for remaining models. If the pixel value is not matched by all Gaussian models, the model with minimum likelihood is replaced by a new Gaussian model.

In outdoor environment, various moving objects may exist in the background scene. The parameter control is used to extract human blobs from moving objects. Three parameters are used for the extraction of human blobs: blob size, blob aspect and position of footprint. Figure 2 shows that human blobs are successfully detected in different and extreme background situations of night, rain and day.

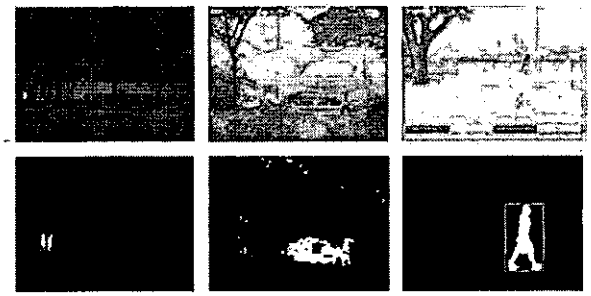

Fig. 2. Detection of human blobs in different scenes

IV. Classification of Human Activities

The human blobs are classified as usual or unusual activities in this module. In this paper, P2DHMM model is employed for the classification of human activities, whose 2-D '*elastic matching*' property is well suitable to represent the variations of human activities.

A. Feature Extraction

The size of human blobs is not constant in the image. In order to perform the proper classification, all human blobs are normalized to the same size. The sampling window is used to extract 1-D observation vectors from the silhouette of human blob. The observation vector is a column-vector that contains 2-D DCT coefficients of image block in the sampling window. Figure 3(a) shows a $W \times H$ the silhouette of human blob. The $P \times L$ sampling window is used to scan the image left to right, top to bottom. The observation vectors can be generated by moving the sampling window with the following procedure,

1) Slide the sampling window from left to right with the sampling step Q till the sampling window reach the right edge of the image.
2) Move the sampling window back to the left edge of the image, slide down it with the sampling step M, and continue the step 1).
3) Repeat the step 2) till the whole image is scanned by sampling window.

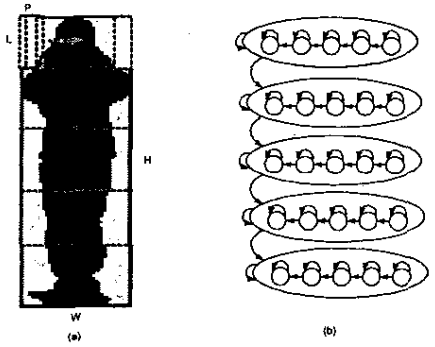

Fig. 3. The Classification of human activities by P2DHMM Model. (a) Feature Extraction. (b) Architecture of P2DHMM.

B. Structure of P2DHMM

The P2DHMM is a stochastic automata with the special two-dimensional arrangement of the states. Figure 3(b) shows the structure of P2DHMM used for the classification of human activities. There are 5 superstates along the vertical direction, each superstate consists of a one-dimensional HMM along the horizontal direction. There are 5 states in each superstate. In the horizontal direction, transitions are only allowed among the states of the superstate. In the vertical direction, transitions occur among the superstates. The P2DHMM model can be symbolized by $\eta = \{N, \mathbf{A}, \mathbf{\Pi}, \mathbf{\Lambda}\}$,

N the number of superstates in the vertical direction.
$\bar{\mathbf{S}} = \{\bar{S}_j : 1 \le j \le N\}$ the superstate set.
$\mathbf{A} = \{a_{kj} : 1 \le k, j \le N\}$ the superstate transition probability matrix, where $a_{kj} = P(q_{y+1} = \bar{S}_j | q_y = \bar{S}_k)$.
$\mathbf{\Pi} = \{\pi_j : 1 \le j \le N\}$ the initial superstate probability distribution, where $\pi_j = P(q_1 = \bar{S}_j)$
$\mathbf{\Lambda} = \{\lambda^j : 1 \le j \le N\}$ the parameter set of left-right 1D HMM within each superstate. In the jth superstate, each λ^j is specified by the standard 1D HMM parameters:
N^j the number of states in the horizontal direction with the jth superstate.
$\mathbf{A^j} = \{a^j_{ki} : 1 \le k, i \le N^j\}$ the state transition probability matrix, where $a^j_{kj} = P(q_{x+1,y} = S^j_i | q_{xy} = S^j_k)$.
$\mathbf{B^j} = \{b^j_i(o_{xy}) : 1 \le i \le N^j\}$ the observation symbol probability distribution for each state, where $b^j_i(o_{xy}) = P(o_{xy} | q_{xy} = S^j_i)$
$\mathbf{\Pi^j} = \{\pi^j_i : 1 \le i \le N^j\}$ the initial state probability distribution, where $\pi^j_i = P(q_{1y} = S^j_i)$.

C. P2DHMM Training

Given the observation sequence of hand posture, P2DHMM learning encodes this observation sequence in such a way that the model should be able to identify it if one observation sequence has many characteristics similar to it. Two P2DHMM models are used to represent usual and unusual human postures. The learning procedure of P2DHMM can be illustrated as follows:

1) Uniform Segmentation: Given the structure of P2DHMM model η, observation sequence is uniformly segmented to obtain initial estimates of model parameters. The overall observations are firstly segmented into N vertical super states, those observations corresponding to each of super states are uniformly segmented from left to right into N^j states.

2) 2-D Segmental K-means Segementation: The model parameters of P2DHMM are estimated by the extension of the segmental k-means algorithm to two dimensions as follows:

$$a^j_{ki} = \frac{\text{number of transitions from } S^j_i \text{ to } S^j_k}{\text{number of transitions from } S^j_i} \quad (5)$$

$$\mu^j_i = \text{mean of sample vectors in state } i \text{ of super state } j \quad (6)$$

$$V^j_i = \text{covariance matrix of sample vectors in state } i \text{ of super state } j \quad (7)$$

$$a_{ij} = \frac{\text{number of transitions from } \bar{S}_i \text{ to } \bar{S}_j}{\text{number of transitions from } \bar{S}_i} \quad (8)$$

3) Doubly Embedded Viterbi Optimization: The doubly embedded Viterbi algorithm is used to search the optimal state sequence in P2DHMM learning. In each row of observation vectors, Viterbi algorithm is used to compute the probabilities of states and observations, which represent the super state probability of this row. Given the super state probabilities, super state transition probabilities and initial superstate probabilities, Viterbi algorithm is used again to decode the model from the top to the bottom.

Define the states $Q = \{\mathbf{q_y}, q_{1y}, \ldots, q_{xy}, \ldots, q_{Xy} : 1 \le y \le Y\}$ corresponding to the observation vectors O, where $\mathbf{q_y}$ is the superstate for the yth row, and q_{xy} represents the state assigned for the observation vector o_{xy}. Given the P2HMM

model η, the best single state sequence is searched to maximize $P(Q|O,\eta)$ using the Viterbi algorithm, which is equivalent to maximizing $P(Q,O|\eta)$. Using logarithmic reformulation, define the quantity $D_y(j)$ as the highest logarithmic probability along the single path at the yth row, which accounts for the observation sequence of the first y rows and ends in superstate $\overline{S}_j$,

$$D_y(j) = \max_{\mathbf{q}_1,\ldots,\mathbf{q}_{y-1}} \{\ln P(\overline{S}_j, o_1, \ldots, o_y|\eta)\} \quad (9)$$

where, $\overline{S}_j = \{\mathbf{q}_1, \ldots, \mathbf{q}_y\}$.

The equation (9) involves the computation of maximum likelihood $P_j(y) = \ln P(O_y|q_y = \overline{S}_j)$, which is the logarithmic probability of row y in superstate j, it is obtained by another execution of the Viterbi algorithm. Define the array $\gamma_y(j)$ to keep track the optimum superstates that maximize the equation (9). The doubly embedded Viterbi algorithm has the following procedure:

1) Initialization

$$D_1(j) = \ln \pi_j + P_j(1) \quad (10)$$

$$\gamma_1(j) = 0 \quad (11)$$

where, $1 \le j \le$ N.

2) Recursion

$$D_y(j) = \max_{j-2\le k\le j} [D_{y-1}(k) + \ln a_{kj}] + P_j(y) \quad (12)$$

$$\gamma_y(j) = \arg \max_{j-2\le k\le j} [D_{y-1}(k) + \ln a_{kj}] \quad (13)$$

where, $2 \le y \le Y$ and $1 \le j \le N$.

3) Termination

$$P^* = \max_{1\le j\le N} [D_Y(j)] \quad (14)$$

4) Superstate and State Sequence Backtracking

$$\mathbf{q}_y = \begin{cases} \arg \max_{1\le j\le N} [D_Y(j)] & \text{if } y = Y \\ \gamma_{y+1}(\mathbf{q}_{y+1}) & \text{otherwise} \end{cases} \quad (15)$$

where, $y = Y, Y-1, \ldots, 1$.

$$q_{N_y} = \chi_{\mathbf{q}_y}(y) \quad (16)$$

$$q_{xy} = \psi^{\mathbf{q}_y}_{x+1,\,y}(q_{x+1,\,y}) \quad (17)$$

where, $x = X-1, \ldots, 1$

The $\chi_{\mathbf{q}_y}(y)$ is the last state of the optimum path of O_y in superstate $\mathbf{q}_y$, it is given by another execution of Viterbi algorithm in the corresponding 1-D model $\lambda^{\mathbf{q}_y}$.

D. Classification of Human Activities

In the classification of human activities, human blobs are firstly detected from the image. They are normalized to a standard size, depending on their original aspect ratio. The observation vectors are extracted from the silhouette of human blobs. The probabilities of observation sequence corresponding to closest P2DHMM model are computed by doubly embedded viterbi optimization, and human activities are classified usual or unusual activities based on the highest likelihood. Figure 4 shows typical postures trained for this paper, where suspicious postures are highlighted.

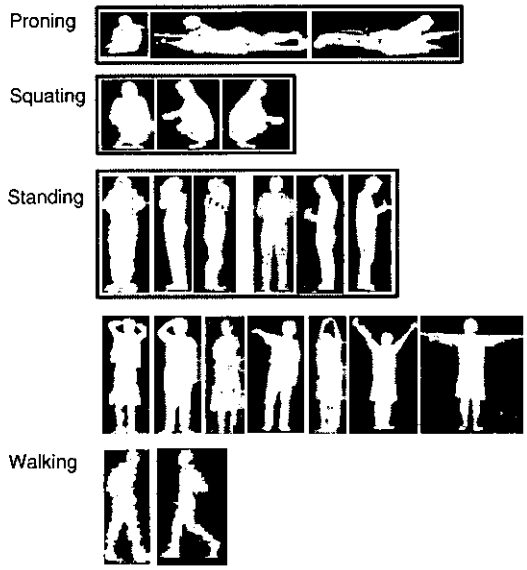

Fig. 4. Training of human activities. Typical human postures with normal and suspicious behaviors. The postures with suspicious behaviors are highlighted.

V. Experimental Results

The proposed video surveillance system has been tested in different outdoor environments. Some results are shown in Figure 5. The suspicious activities, such as postures of taking note, taking photo, are well detected.

VI. Conclusions

A video surveillance system is presented in this paper. It is able to detect unusual activities from video sequences based on four computational modules: *background subtraction*, *extraction of human blobs*, *classifier training* and *classification of human activities*. In background subtraction, *Mixture of Gaussians* are used to model outdoor scene background. This model adapts quickly for the changes of background scene, and it is able to handle the side effects of illumination and extraneous motions. The P2DHMM model is used for the classification of human activities. This probabilistic model provides a flexible approach for the training of human activities. It is able to discriminate the differences among different activity patterns while being robust to variations for the same

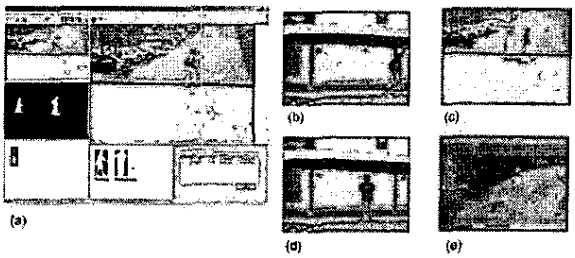

Fig. 5. Detection of unusual human activities. (a) GUI of the detection system detailing the extracted blobs, the database of trained activities and records of past suspicious activities (b) normal human walking. (c) The suspicious activity is detected from two human activities: taking photo posture and normal walking .(d) The suspicious activity is detected: taking note. (e) The suspicious activity is detected: taking note posture

activity. We have tested the system by detecting suspicious activities from multiple postures: of which we have success in recognizing taking note and taking photo. For a complete video surveillance system, it will be further improved to identify more human activities.

REFERENCES

[1] I. Pavlidis, V. Morellas, P. Tsiamyrtzi, and S. Harp, "Urban surveillance systems: From the laboratory to the commercial world," *Proceedings of the IEEE*, vol. 89, no. 10, pp. 1478–1497, Oct. 2001.

[2] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithm for cooperative multi-sensor surveillance," *Proceedings of the IEEE*, vol. 89, no. 10, pp. 1456–1477, Oct. 2001.

[3] T. D. C. R. Wren, A. Azarbayejani and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 780–785, July 1997.

[4] C. Stauffer and W. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[5] D. H. A. Elgammal, R. Duraiswami and L. S. Davis, "Background and foreground modeling using nonparameteric kernel density estimation for visual surveillance," *Proceedings of IEEE*, vol. 90, no. 7, pp. 1151–1163, July 2002.

[6] R. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," in *Proc. (IEEE) Conf. on Automatic Face and Gesture Recognition*, Washington, DC USA, May 2002, pp. 351–356.

[7] O.Javed and M. Shah, "Tracking and object classification for automated surveillance," in *Proceedings of IEEE*, vol. 90, Denmark, May 2002, pp. 343–357.

[8] A. Datta, M. Shah, and N. D. V. Lobo, "Person-on-person violence detection in video data," in *Proc. (IEEE) Conf. on Pattern Recognition*, Aug. 2002, pp. 438–438.

[9] J. Dever, N. D. V. Lobo, and M. Shah, "Automatic visual recognition of armed robbery," in *Proc. (IEEE) Conf. on Pattern Recognition*, Aug. 2002, pp. 451–455.

[10] M. Bicego and V. Murino, "2d shape recognition by hidden markov models," in *Proc. (IEEE) Conf. on Image Analysis and Processing*, Palermo, Italy, Sept. 2001, pp. 20–24.

[11] W.Abd-Almageed and C. Smith, "Hidden markov models for silhouette classification," in *Proceedings o the 5th Biannual World Automation Congress*, vol. 13, June 2002, pp. 395–402.

[12] A. V. Nefian and M. H. Hayes, "A embedded hmm-based approach for face detection and recognition," in *Proc. (IEEE) Conf. on Acoustics, Speech and Signal Processing*, Phoenix, AZ, USA, Mar. 1999, pp. 3553–3556.