Automatic surveillance analyzer using trajectory and body-based modeling

Automatic Surveillance Analyzer Using Trajectory and Body-based Modeling

Weilun Lao[1], Jungong Han[1] and Peter H.N. de With[1, 2]
[1] Eindhoven University of Technology, P.O.Box 513, 5600MB Eindhoven
[2]Cyclomedia Technology B.V., P.O.Box 68, 4180BB Waardenburg, The Netherlands

***Abstract*—With the continuous improvements in computer-vision techniques, automatic low-cost video surveillance gradually emerges for consumer applications. Successful trajectory estimation and human-body modeling facilitate the semantic analysis of human activities in video sequences. We propose a fast analyzer for surveillance video, which aims at automatic analysis of human behavior and semantic events. Our analyzer employs visual cues to track moving persons and classify human-body postures from a monocular video. It consists of three processing steps: (1) multi-person detection, (2) persons tracking with trajectory estimation, and (3) posture classification. We show attractive experimental results, highlighting the system efficiency and classification capability.**

I. INTRODUCTION

Visual surveillance based on human-behavior analysis has been investigated worldwide as an active research topic [1]. As a consumer video application, automatic surveillance system requires high accuracy and its computation complexity needs significant reduction to obtain real-time performance. More specifically, only the trajectory estimation is not sufficient. The postures of the persons can provide important information for understanding their activities. Therefore, accurate detection and efficient recognition of various human postures contribute to event-based analysis.

Significant research has been devoted to event-based surveillance in the past few years. In [2], shape and pose are utilized for pedestrian monitoring. The input frame is warped to the training view and processed using a corresponding view-based model. However, it can only detect simple motion patterns (e.g. walking) with strong prior or learnt model. In [3], a pixel-wise event-representation method is proposed to construct feature images, in which each blob corresponds to a visual event. Then blob-level features are transformed into subspaces to model probabilistic appearance manifolds for each event. However, it cannot meet the practical requirement for online processing. In this paper, we propose a fast scheme for human-motion analysis, which deals with various human activities composed of sub-events (individual posture). Besides, we propose a simple but effective shape descriptor to represent the human silhouette. Only binary-shape information is utilized while texture/color or an explicit body model is not required. The temporal consistency of the human activity is also considered.

We focus on developing a robust surveillance system from monocular video sequences. Unlike existing proposals only estimating the trajectory, we attempt to further analyze the individual posture and model the multi-person interaction. The objective is to achieve event-based semantic analysis, exploiting interaction modeling. The semantic event (i.e. the bank robbery) is inferred and the alarm is triggered. Our detector achieves near real-time speed with promising results.

This research is part of the European ITEA Cantata on content analysis.

II. SYSTEM DESCRIPTION

When performing a trajectory-based estimation and body-based detection, we intend to capture the human motion and analyze the silhouette using temporal filtering. The block diagram of our proposed scheme is shown in Figure 1.

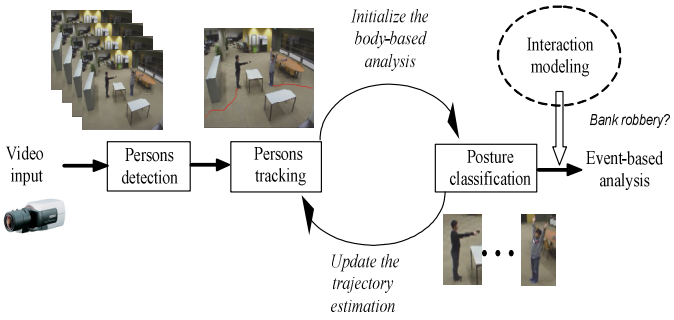

Figure 1: Framework of the proposed system

A. *Multi-person detection*

The first module of our system shown in Figure 1 is multi-person detection, whose purpose is to extract potential persons appearing in the scene. Firstly, we perform a pixel-based background subtraction, where the scene model has a probability density function for each pixel separately. A pixel from a new image is considered to be a background pixel if its new value is well described by its density function. After the background modeling, the next step is to e.g. estimate appropriate values for the variances of the pixel intensity levels from the image, since the variances can vary from pixel to pixel. Pixel values often have complex distributions and thus more elaborate models are needed. The Gaussian Mixture Model (GMM) is employed for the background subtraction. We apply the algorithm from reference [4] to produce the foreground objects using a Gaussian-mixture probability density with shadows removal. The parameters for each Gaussian distribution are updated in a recursive way. Furthermore, the method can efficiently select the appropriate number of Gaussian distributions during pixel processing, so as to fully adapt to the observed scene. Afterwards, we use a k-Nearest Neighbor (k-NN) classifier to recognize persons. The classifier utilizes two features commonly used for object classification: area, and the ratio of the bounding box attached to each detected object. This approach is simple but efficient, and it contributes significantly to the tracking and avoids a complex procedure for training data. The non-person objects and image noise can be effectively removed.

B. *Multi-person tracking*

The second step is to track every detected person in the frames. In this trajectory-estimation step, we utilize the broadly accepted mean-shift algorithm for tracking persons, based on their individual appearance model represented as a color histogram. When the mean-shift tracker is applied, we extract every new person entering the scene and calculate the corresponding histogram model in the image domain. In subsequent frames for tracking that person, we shift the

978-1-4244-2559-4/09/$25.00 ©2009 IEEE

person object to the location whose histogram is the closest to the previous frame. After the trajectory is located, we can conduct the body-based analysis at the location of the person in every frame. When the trajectory is obtained, we can also estimate the position and speed of the persons involved in the video scene. The label of walking or standing can be therefore assigned to each individual person. From our previous work in [5], we can also adopt the Double Exponential Smoothing (DES) operator to track moving persons, which runs approximately 135 times faster than the popular Kalman filter-based predictive tracking algorithm with equivalent prediction performance.

C. *Posture classification*

Unlike conventional algorithms [6][7][8] that calculate the *Hu* moments or detect the body parts along the contour, we propose a simple but effective approach to analyze the silhouette.

First, every detected person is adapted to a *w*×*h* pixels template in a normalization phase (we choose *w=80* and *h=180*). Then within the template, we apply the horizontal and vertical projections. Because either the horizontal or vertical projection itself is non-orthogonal, the projections along the vertical and horizontal axis with 180 and 80 dimensionalities are redundant. Therefore, a *Principal Component Analysis* (PCA) is used to obtain a more compact, and accurate information in each frame. In the vertical projection, the shape vector of 180 elements is divided into 3 parts and thus a feature matrix of 60×3 is obtained for each frame. Then the dimensionality of each feature matrix is reduced to 2×3 after performing PCA. Afterwards, a vector of 6×1 is concatenated from the above matrix. Although there are different options for the matrix form, we adopt the division scheme of 60×3 to achieve the balance between computation cost and recognition rate. Similarly, a vector of 8×1 is concatenated from the horizontal projection. Hence, in total we obtain a 14-D vector to represent the human silhouette. We name the above feature *HV-PCA*. It is defined by

$$obv_i = (P_w(\underset{(x,y)\in S}{H_i}(x,y)), P_h(\underset{(x,y)\in S}{V_i}(x,y)))^T \quad (1)$$

where obv_i is the observed feature vector of the silhouette in the frame *i*, and H_i and V_i refer to the horizontal and vertical projection, respectively, positions *(x,y)* represent the pixels of the silhouette within the template *S,* and *P(.)* indicates the PCA implementation. Then every obv_i is set as an observation input to the CHMM (Continuous Hidden Markov Models) classifier for parameters learning or testing. After training the classifier for each posture, we can estimate the posture type in the test sequences. Finally every given posture is classified into one of the following types: *left-pointing*, *right-pointing*, *squatting*, *raising hands overhead* and *lying*. Furthermore, we can apply expert knowledge for semantic event analysis. For example, in the study case of bank-robbery detection, we can infer that bank robbery occurs when person A is labeled as "raising hands overhead" after person B is detected "pointing".

Table 1: Posture classification results

Posture classification	4-Hu [6]	Skel. [7]	Skel2.[8]	HV-PCA
Left pointing	84%	82%	88%	92%
Right pointing	88%	78%	84%	90%
Squatting	78%	56%	60%	84%
Raising hands overhead	84%	66%	72%	86%
Lying	80%	56%	64%	76%

Figure 2: Example of our detector result. The trajectory of every person is visualized after persons tracking. The postures are estimated and the semantic event is highlighted after interaction modeling

III. EXPERIMENTS AND CONCLUSIONS

We have performed our analyzer using more than 30 and 50 video sequences of various single/multi-person motion (15 frames/s) for training and testing, respectively. It achieves a 98% accuracy rate on multi-person detection, 96% detection rate on multi-person tracking, where the criterion is that at least 80% of the human body is included in the detection window. The posture-classification results using different methods are summarized in Table 1. The proposed HV-PCA feature achieves the accuracy rate of around 86% and generally outperforms other proposals. Our system is efficient, achieving a near real-time performance (6~8 frames/s for 320×240 resolution, with a P4 3-GHz PC). The postures classification is important for event-based analysis when the heuristic rules are defined. The Figure 2 shows a detection example of a simulated bank-robbery event. These results can be readily translated into burglary detection for home-use.

As our human-motion analyzer is efficient and achieves fast performance, we can facilitate its application in a surveillance system or further analyze the human behavior based on interaction modeling.

REFERENCES

[1] T. B. Moeslund, A. Hilton and V. Kruger, "A survey of advances in vision-based human motion capture and analysis", *Computer Vision and Image Understanding*, vol. 104, pp. 90-126, 2006.

[2] G. Rogez, J.J. Guerrero and C. Orrite, "View-invariant human feature extraction for video-surveillance applications", *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp.1-8, 2007.

[3] P. Cui, L. Sun, Z. Liu and S. Yang, "A sequential Monte Carlo approach to anomaly detection in tracking visual events", *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp.1-8, 2007.

[4] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction", *Pattern Recognition Letters*, vol. 27, pp.773-780, 2006.

[5] J. Han, D. Farin, P. H. N. de with and W. Lao, "An automatic analyzer for sports video databases using visual cues and real-world modeling", *Proc. IEEE International Conference on Consumer Electronics,* pp. 477-478, 2006.

[6] H. Li, S. Wu, S. Ba, S. Lin and Y. Zhang, "Automatic detection and recognition of athlete actions in diving video", *Proc. Int. Conf. Multimedia Modeling*, LNCS 4352, pp. 73-82, 2007.

[7] H. Fujiyoshi, A. Lipton and T. Kanade, "Real-time human motion analysis by image skeletonization", *IEICE Trans. Information and System*, vol. 87, pp. 113-120, 2004.

[8] P. Peursum, H. Bui, S. Venkatesh and G. West, "Robust recognition and segmentation of human actions using HMMs with missing observations", *EURASIP Journal on Applied Signal Processing*, vol. 13, pp. 2110-2126, 2005.