Real-Time Multi-person Tracking in Video Surveillance

Wei Niu, Long Jiao, Dan Han, and Yuan-Fang Wang

Department of Computer Science
University of California
Santa Barbara, CA 93106

Abstract

In this paper, we briefly summarize our video surveillance research framework. We then survey current research on human activity recognition, and present our current work on real-time multi-person tracking. By applying adaptive background subtraction, foreground regions are first identified and segmented. A clustering algorithm is then used to group the foreground pixels in an unsupervised manner to estimate the image location of individual persons. A Kalman filter is used to keep track of each person and a unique label is assigned to each tracked individual. Based on this approach, people can enter and leave the scene at random. Abnormity, such as silhouette merging, is handled gracefully and individual persons can be tracked correctly after a group of people split. Experiments demonstrate the real-time performance and robustness of our system working in complex scenes.

1. Introduction

There has been a surge in the number of surveillance cameras put in service in the last two years since the September 11th attacks. Closed Circuit Television (CCTV) has grown significantly from being used by companies to protect personal property to becoming a tool used by law enforcement authorities for surveillance of public places. However, several important research questions must be addressed before we can rely upon video surveillance as an effective tool for crime prevention. In this paper, we briefly summarize our multi-camera video surveillance system framework that comprises schemes for spatio-temporal data fusion, event representation, and recognition. Our system can track multiple cars and detect suspicious driving behaviors such as circling and zigzag patterns [21]. We will then report our results on multi-person tracking.

The rest of the paper is organized as follows. Section 2 presents a short survey of people identification and activity recognition techniques in video surveillance. Section 3 presents our video surveillance framework. Current work on multi-person tracking is described in section 4. Section 5 presents the experimental results. Finally, section 6 offers the concluding remarks and directions of future work.

2. Background

Computer analysis of human actions is gaining increasing interests, especially in video surveillance arenas where people identification and activity recognition are important. Using two important metrics: preciseness of the analysis outcome and the required video resolution to achieve the desired outcome, human identification and activity recognition can be classified into three categories. At one extreme, which is often characterized by high video resolution and a small amount of scene clutter, high fidelity outcome is achievable [2, 6, 7, 10, 11, 12]. Many gait recognition techniques fall in this category, which aim to identify individuals against a pre-established gait database. At the other extreme, which is characterized by low video resolution and potentially significant scene clutter, it is often not possible to achieve highly discriminative outcome. Instead, the goal is often to detect the presence, and identify the movement and interaction of people through "blob" tracking [5, 9, 14]. In the middle of the spectrum, it is possible to refine the "blob" representation of a person through hierarchical, articulated models [1, 3, 4, 8, 13, 15]. This allows main body parts, such as head, arms, torso, and legs, to be individually identified to specify the activities more precisely. Some representative research works in these three scenarios are briefly surveyed below.

2.1 High Resolution

For near field application, in which high fidelity video is available, analysis results can be revealing enough to ID individuals. Many techniques in face, gesture, and gait recognitions fall in this category. For video surveillance, gait recognition may be one of the most interesting problems. Interested readers can refer to [20, 1, 7, 4, 12] for a discussion of face and gesture recognition.

Approaches to gait recognition can be classified roughly as either model-based or model-free. One example of the model-based method is [9], which fits the movement of the thighs to an articulated pendulum-like template motion pattern. Assuming the motion of thighs as a periodic signal, coefficients of the Fourier series of the signal are used as the feature for recognition, and a Genetic Algorithm is used to perform a heuristic search of the parameter space. Because a traditional CANNY edge detector is employed for edge detection with little domain-specific information used for discriminating leg from other body parts, non-leg edges can also be included in the feature description. This implies that the approach may be sensitive to noise in video processing. Another example is [16], in which seven ellipses are used to represent different parts of the silhouette of a person. For each ellipse, the centroid, aspect ratio of the major and minor axes, and the orientation of the major axis are extracted. For each image frame, a "region feature" is formed by combining these parameters of the seven ellipses. Given a gait sequence, two kinds of features are computed over time. One feature is the mean and standard deviation of the region features, combined with one additional parameter: the height (relative to the body length) of the centroid of the whole silhouette. Together, they provide a "gait average appearance feature." The other feature is computed based on the magnitude and phase of

the Fourier transform of the region features in the sequence, which gives a "gait spectral component feature." In recognition, the authors reported that the average appearance feature shows better results if clothing information is available in the gait library, otherwise the spectral feature is more reliable. However, both features could be affected significantly in the presence of noise in the silhouette. And to combine these two features will be a possible improvement.

One example of the model-free methods is [3], which first scales the sizes of blobs in a gait sequence to a standard one, then maps the sequence to a 2D feature consisting of the N*N matrix of the difference (which is calculated by accumulating the difference of pixel intensity values in the blobs) between each pair of images in the sequence (where N is the size of the sequence). The principal components analysis (PCA) is then used to reduce the dimensionality of the feature space. Finally, classification is performed based on the k-nearest-neighbor rule in the reduced space. [14] uses as features the smoothed and down-sampled versions of the width vector and differenced width vector that represent the projection profile of the width of a body silhouette along the vertical (height) direction. Dynamic time warping (DTW) is used in matching gait sequences to adapt to the changes of walking speed. In [10], various features are extracted from the gait sequence, such as the swing of the hands/legs, the sway of the upper body and static features like height. For different kinds of features, different methods such as DTW and HMM are used for classification. For example, the features for the swing of the hands/legs are projection vectors, and are matched by DTW. And the HMM is used to represent the leg dynamics. The results of these classifications are combined by Sum, Product and MIN rules to achieve a decision fusion. This approach improves the overall recognition performance. [3], [14] and [10] are not view invariant and need camera calibration information. [15] proposes a view invariant method to synthesize the side view from any other arbitrary view using a single camera if the person is far enough from the camera. By using the perspective projection model and the optical flow based structure from motion equations, the azimuth angle of the original view is estimated, and a video sequence at the new side view is synthesized. This approach can be combined with other gait analysis techniques for efficient and invariant recognition.

2.2 Medium Resolution

Here, the goal is to recognize generic activities such as the movement of arms and legs, instead of trying to tie the action to a particular actor/actress like in gate recognition.

For example, [5] used the MHI (motion-history images) to record both the segmentation result and the temporal motion information. The MHI is a *single* image composed of superimposing a sequence of segmented moving objects weighed by time. The most recent foreground pixels are assigned the brightest color while past foreground pixels are progressively dimmed. This allows the summarization of information on both the spatial coverage and the temporal ordering of the coverage of an activity. The MHI does not use any structure to model human. A vector of seven moment values is computed for each MHI. Activities are recognized by finding the best match of the moment vectors between the query MHI and the training patterns.

[6] does not segment all human body parts. Instead, it just tracks the head and two hands. It uses CHMMs (coupled HMMs) to analyze a sequence of position information and to recognize complex actions (T'ai Chi gestures). CHMM is composed by two or more traditional HMMs. In [6], two HMMs are used, and they represent the movements of the left and right hands, respectively. Two HMMs are coupled together to represent the interactions between them. For example, consider the action of clapping hands. Suppose that we have two HMMs, each representing the motion of a single hand If the state of the left hand HMM is to move toward the center of the body, it then implies the next state of the right hand HMM should be to either move toward or just come to the center of the body. CHMM thus provides a method to analyze the movements of multi-body parts together, and is more powerful than a single HMM since it can describe the relationships between several HMMs. But when the number of coupled HMMs increases, the complexity increases exponentially.

[2] uses a 2-D stick model to represent torso, arms, and legs. The recognition algorithm consists of two phases. The first phase computes the angles between connected body parts such as the upper arms and the lower arms. The algorithm matches the angular trajectories with trained data and accumulates the best matches into a hash table. The second phase then computes the vote for a whole motion sequence and identifies the activity as the one in the database receiving the largest number of votes.

The above three algorithms all use 2D information. It is also possible to employ explicit 3D models. Some current research is focused on tracking human body parts using generic 3-D models, e.g., head and hands. There are three difficulties in building a 3-D model [11]: 1.) It may be difficult to recover the depth information from 2-D images, 2.) feature extraction in three dimensional space can be challenging, and 3.) the degree of freedom of 3D motion can be huge. [11] builds a 3D model from multiple views. They initialize the 3D model at the first frame, then estimate the model state by the Kalman filter and physical forces which pull the template model to confirm to the real object pose observed in image. The information extracted from image includes depth information from stereo-correlation and the silhouettes. The distance between extracted features and those predicted by 3D model is computed. And the strength of physical force is determined by the distance of deviation between image features and the 3D model. The above process is computed iteratively until it converges.

Finally, [17], [19], etc. make a compromise between 2D and 3D analysis. The 2D information in [17] includes a set of feature points, e.g., head, right and left hands, and the contour of the human body. The 3D model used is a skeleton model. It describes the bone structure of a person represented by joints and vertex points, such as hands and head. Given 2D information, 3D structure can be inferred through inverse kinematics. But it is often ambiguous to infer 3D structure just from a single view point. To resolve

the ambiguity, the method fuses the model from multiple views. To build and track a 3D model is difficult. A hybrid 2D-3D model may be a reasonable compromise.

2.3 Low Resolution

For far field application, the goal of people tracking is often to detect the presence, and identify the movement and interaction of multiple persons through "blob" tracking. The VSAM system [8] tracked the human body as a whole blob. They use a hybrid algorithm by combining adaptive background subtraction with a three-frame differencing technique to detect moving objects, and use the Kalman filter to track the moving objects over time. A neural network classifier is trained to recognize four classes: single person, group of persons, vehicles, and clutter. They also use linear discriminant analysis to further provide a finer distinction between vehicle types and colors. The VSAM system is very successful at tracking humans and cars, and at discriminating between vehicle types. But it did not put much emphasis on activity recognition; only gait analysis and simple human-vehicle activity recognition are handled.

The W4 system [13] is designed for outdoor surveillance tasks and particularly for night-time or other low light level situations. It operates on monocular grayscale video imagery. Foreground regions are detected by a combination of background analysis and simple low level processing of the resulting binary image. A second-order motion model including the velocity and acceleration terms is used to track the overall body motion and the motions of body parts. Region splitting and merging are handled and individual body parts such as head, hands, torso and legs are tracked in order to understand actions. The W4 system represents a good first step to the problem of recognizing and analyzing human activities, but the cardboard model used in W4 to predict body pose and position is restricted to upright people. In order to recognize and track people in other generic poses like crawling, the convex hull-like representations of the silhouettes of people should be incorporated

Paragios and Deriche incorporated geodesic active contour and level sets method for the detection and tracking of moving objects [18]. The geodesic active contour objective function is minimized using a gradient descent method: the curve is propagated towards the object boundaries under the influence of boundary, intensity, and motion-based forces using a PDE, which is implemented using a level set approach where topological changes are naturally handled. But this approach is limited to boundary-based information, region-based tracking modules could be incorporated to increase robustness. Moreover, a direct implementation of level set method is computationally very expensive. The necessity of real time detection should lead to a multi-scale approach.

3. System Framework

We have built a system for video surveillance of vehicular motion. This system recognizes driving patterns by analyzing vehicular trajectories. Multiple cameras are deployed in a parking lot to increase the spatial coverage. Each camera is calibrated and registered to a common frame of reference. The trajectory of a vehicle is integrated from the tracking results of several cameras. With the help of camera registration information, the trajectory can be fused in the world coordinate system. The recognition algorithm takes trajectories as input and consists of two steps. The first step is to transform the numeric trajectory data into semantic descriptions such as turns and stop. The second step uses SVM and HMM to recognize the motion patterns. We have achieved reasonable tracking and classification performance. For more details, please see http://www.cs.ucsb.edu/~longjiao/demo.mpg for a demo.

Our current work is focused on generalizing this framework for human action recognition. While the processing flow of detection-representation-recognition might be valid for recognizing vehicular and human activities, the detailed tracking, representation and recognition schemes must be adapted for deformable silhouettes and human figures. We will present our preliminary results on people tracking.

4. Multi-person Tracking

Segmenting and tracking multiple people in real-time is a challenging but important problem in video surveillance. Our video surveillance system can automatically detect moving objects, and classify the moving objects into semantic categories such as car and human. If the tracked object is a person, a unique label will be assigned to this person, information such as height, centroid and the average intensity value of the region occupied by the tracked person will be recorded as the signature. The Kalman filter is used to track each person. As a result, multiple people can enter and leave the scene at random. Abnormity, such as silhouette merging, is handled and individuals can be tracked correctly after a group of people split. If the same person re-enters the view in a reasonably short period of time, our system recognizes the reentry and assign the same label to the person.

4.1 Object Detection

Background Subtraction Adaptive Background subtraction techniques have been used extensively to detect the foreground regions [8]. With the assumption of a static background, the adaptive background subtraction technique can adapt to slow changes such as illumination changes by recursively updating the background model. Let $B_n(x)$ represents the current background intensity value at pixel x and $I_n(x)$ represents the current intensity value at pixel x, then x is considered as a foreground pixel if:

$$\left|I_n(x) - B_n(x)\right| > T_n(x) \qquad (1)$$

While $B_0(x)$ is initially set to be the first frame and $T_0(x)$ is initially set to some empirical non-zero value, both $B_0(x)$ and $T_0(x)$ are updated over time.

Foreground Object Detection After binarizing the image derived from background subtraction, a filter is applied to the image to get rid of the noise. Then a fast connected-component operator is applied to the image to locate the foreground regions. The operator assigns a clustering label to each pixel according to the labels of its

upper-left and left neighbors. After this self-clustering procedure, small regions are eliminated, and big regions are considered as interesting objects.

4.2 Object Tracking

The second order Kalman filter (state including position, velocity, and acceleration) is used to model the motion of each person in the scene. The Kalman filter works in two stages: prediction and correction.

Suppose that there is an object moving in the scene, whose tracked image trajectory in local camera reference frame is described as $\mathbf{p}(t) = [x(t), y(t)]^T$. The system state vector for this observed object is $\mathbf{x}(t) = [\mathbf{p}(t), \dot{\mathbf{p}}(t), \ddot{\mathbf{p}}(t)]^T$ where $\mathbf{p}(t)$ represents the position, $\dot{\mathbf{p}}(t)$ the velocity and $\ddot{\mathbf{p}}(t)$ the acceleration of the tracked object. By using Taylor series expansion, the state prediction equation is

$$\mathbf{x}(t^-) = \mathbf{A}\mathbf{x}(t-\Delta t) + \mathbf{w}(t)$$

$$\begin{bmatrix} \mathbf{p}(t^-) \\ \dot{\mathbf{p}}(t^-) \\ \ddot{\mathbf{p}}(t^-) \end{bmatrix} = \begin{bmatrix} \mathbf{I}_2 & \Delta t\mathbf{I}_2 & \frac{\Delta t^2}{2}\mathbf{I}_2 \\ \mathbf{O}_2 & \mathbf{I}_2 & \Delta t\mathbf{I}_2 \\ \mathbf{O}_2 & \mathbf{O}_2 & \mathbf{I}_2 \end{bmatrix} \begin{bmatrix} \mathbf{p}(t-\Delta t) \\ \dot{\mathbf{p}}(t-\Delta t) \\ \ddot{\mathbf{p}}(t-\Delta t) \end{bmatrix} + \mathbf{w}(t) \quad (2)$$

where $\mathbf{I}_2$ and $\mathbf{O}_2$ represents 2×2 identity and null matrices, respectively. $\mathbf{w}(t)$ represents the model prediction uncertainty, and is assumed to be a random variable with a zero mean and a covariance matrix $\mathbf{Q}(t) = E[\mathbf{w}(t)\mathbf{w}^T(t)]$.

In the prediction stage, the foreground object position is computed by extrapolating the state of the Kalman filter from the previous frame to the current frame. Then the correspondence between the predicted foreground object position and the foreground object position detected by background subtraction is computed. The measurement equation is simple: We estimate the person's position, velocity, and acceleration by taking equivalent finite difference from the corresponding trajectory:

$$\mathbf{z}(t) = \mathbf{H}\mathbf{x}(t) + \mathbf{v}(t) \quad (3)$$

$$\begin{bmatrix} \mathbf{p}(t) \\ \frac{\mathbf{p}(t) - \mathbf{p}(t-\Delta t)}{\Delta t} \\ \frac{\mathbf{p}(t) - 2\mathbf{p}(t-\Delta t) + \mathbf{p}(t-2\Delta t)}{\Delta t^2} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_2 & \mathbf{O}_2 & \mathbf{O}_2 \\ \mathbf{O}_2 & \mathbf{I}_2 & \mathbf{O}_2 \\ \mathbf{O}_2 & \mathbf{O}_2 & \mathbf{I}_2 \end{bmatrix} \begin{bmatrix} \mathbf{p}(t) \\ \dot{\mathbf{p}}(t) \\ \ddot{\mathbf{p}}(t) \end{bmatrix} + \mathbf{v}(t)$$

$\mathbf{z}(t)$ represents the external observation (the image trajectory), and $\mathbf{v}(t)$ represents the uncertainty in such an observation. Here we assume $\mathbf{v}(t)$ is zero mean with a covariance matrix $\mathbf{R}(t) = E[\mathbf{v}(t)\mathbf{v}^T(t)]$.

In correction stage, the new measurement (foreground object's position computed by background subtraction) is incorporated to update the state of Kalman filter model by the following process:

$$\mathbf{x}(t^+) = \mathbf{x}(t^-) + \mathbf{K}(t)(\mathbf{z}(t) - \mathbf{H}\mathbf{x}(t^-)) \quad (4)$$

The weighting factor $\mathbf{K}(t)$ comes from summarizing the following three equations:

$$\mathbf{K}(t) = \mathbf{E}(t^-)\mathbf{H}^T(t)[\mathbf{H}(t)\mathbf{E}(t^-)\mathbf{H}^T(t) + \mathbf{R}(t)]^{-1} \quad (5)$$

$$\mathbf{E}(t^-) = \mathbf{A}\mathbf{E}(t-\Delta t)\mathbf{A}^T + \mathbf{Q}(t) \quad (6)$$

$$\mathbf{E}(t^+) = [\mathbf{I} - \mathbf{K}(t)\mathbf{H}(t)]\mathbf{E}(t^-) \quad (7)$$

where $\mathbf{E}(t) = E[(\mathbf{x}(t) - \hat{\mathbf{x}}(t))(\mathbf{x}(t) - \hat{\mathbf{x}}(t))^T]$ is the error covariance matrix in the state estimation process.

Handling Emerging Objects Several special cases may happen while matching the predicted foreground object position with the detected foreground object position. When a foreground region is detected while there is no corresponding object in the past, it may signal a new person just entering the field of view. If this new foreground region can be tracked successfully for several frames, it will then be considered as a new person, with a unique label assigned. Information such as height, centroid, and average intensity value of the person is recorded as the signature. A Kalman filter model is initialized to track this new person in the ensuing frames.

Handling Merging and Splitting of Objects If two or more tracked objects merge into one foreground region, temporary occlusion may have happened. In order to track each person individually under temporary occlusion, the prediction of Kalman filter model and the information of matched foreground region are incorporated to update each of the occluded person's Kalman model.

$$\begin{bmatrix} \mathbf{p}(t^-) \\ \dot{\mathbf{p}}(t^-) \\ \ddot{\mathbf{p}}(t^-) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_2 & \Delta t\mathbf{I}_2 & \frac{\Delta t^2}{2}\mathbf{I}_2 \\ \mathbf{O}_2 & \mathbf{I}_2 & \Delta t\mathbf{I}_2 \\ \mathbf{O}_2 & \mathbf{O}_2 & \mathbf{I}_2 \end{bmatrix} \begin{bmatrix} \mathbf{p}(t-\Delta t) \\ \dot{\mathbf{p}}(t-\Delta t) \\ \ddot{\mathbf{p}}(t-\Delta t) \end{bmatrix} + \begin{bmatrix} (\mathbf{1}-\boldsymbol{\alpha})_2\,\mathbf{z}(t) \\ 0 \\ 0 \end{bmatrix} \quad (8)$$

Both $\boldsymbol{\alpha}_2$ and $(\mathbf{1}-\boldsymbol{\alpha})_2$ are 2×2 diagonal matrices with the diagonal elements being α, which is an empirical number between 0 and 1.

After a group of people split, each person can still be tracked correctly. In next session, we will show the experimental results that handled the temporary occlusion.

5. Experimental Results

We implemented the multi-person tracking system in C++ using a Windows XP PC as the implementation platform. For 354x240 resolution color images, our system runs at 25Hz using a 1.8GHz Celeron processor. We recorded 3 minutes video using a Samsung SCD MinDV camcorder. Our system has the capability of tracking multiple persons in complex scenes and it never lost tracking. Each occluded person can be tracked accurately before, during, and after occlusion. Figure 1 shows the tracked persons before, during, and after occlusion. Figure 2 shows the two persons leave and re-enter the scene.

6. Conclusion and Future Work

In this paper, we describe the framework of our video surveillance system and provide the algorithms and implementation results of our current work on multi-person tracking. Our system works in real-time. It can track multiple persons in the camera's field of view accurately.

Temporary occlusion can be handled successfully and if the same person re-enter the field of view after a reasonably short period of time, our system can recognize the re-entry and assigns the same label to the person.

Our future work will focus on estimating the 3D trajectory of each moving object using multi-camera data fusion, analyzing the multiple-person interaction, and detecting suspicious behaviors.

7. Acknowledgement

This work was supported in part by an NSF grant, IRI-9908441. Part of the work was jointly done with Edward Chang, Guan Wu, and Yi Wu.

References

[1] S. Basu, I.A. Essa, A.P. Pentland, "Motion Regularization for Model-based Head Tracking", International Conference on Pattern Recognition, 1996.

[2] J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram, "Human Activity Recognition Using Multidimensional Indexing", PAMI, Vol. 24, No. 8, August 2002.

[3] C. BenAbdelkader, "Gait as a Biometric for Person Identification in Video Sequences", Dissertation, University of Maryland, 2001.

[4] M. J. Black, A. D. Jepson, "EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation", ECCV, 1996

[5] A. F. Bobick, and J. W. Davis, "The Recognition of Human Movement Using Temporal Templates", PAMI, Vol. 23, No. 3, 2001.

[6] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition", proceedings of CVPR 97.

[7] S. Chandrasekaran, B. S. Manjunath, Y. F. Wang, J. Winkeler, and H. Zhang, "An Eigenspace Update Algorithm for Image Analysis", Graphical Models and Image Processing, Vol.59, pp.321–332, 1997.

[8] R. Collins, A. Lipton, *et al*. A System for Video Surveillance and Monitoring. CMU-RI-TR-00-12, Robotics Institute, CMU, May, 2000.

[9] D. Cunado, J.M. Nash, M.S. Nixon, and J. N. Carter, "Gait extraction and description by evidence- gathering," Proc. of the International Conference on Audio and Video Based Biometric Person Authentication, pp. 43–48, 1995.

[10] N. Cuntoor, A. Kale and R. Chellappa, "Combining Multiple Evidences for Gait Recognition", ICASSP 2003.

[11] Q. Delamarre, O. Faugeras, "3D Articulated Models and Multi-View Tracking with Physical Forces", CVIU Vol. 81, pp328-357, 2001.

[12] D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey", Computer Vision and Image Understanding, Vol.73, No.1, pp.82-98, 1999.

[13] I. Haritaoglu, D. Harwood, and L. Davis. W4: Who, when, where, what: A real time system for detecting and tracking people. In Third Face and Gesture Recognition Conference, pages:222-227, 1998.

[14] A. Kale, N. Cuntoor, B Yegnanarayana, A.N Rajagopalan, R. Chellappa, "Gait analysis for human identification", Proceedings of the 3rd International conference on Audio and Video Based Person Authentication, 2003.

[15] A. Kale, A. K. Roy Chowdhury and R. Chellappa, "Towards a View Invariant Gait Recognition Algorithm", AVSS, 2003.

[16] L. Lee and W.E.L. Grimson, "Gait analysis for recognition and classification," Proceedings of the IEEE Conference on Face and Gesture Recognition, pp. 155–161, 2002.

[17] E. Ong, and S. Gong, "A Dynamic Human Model Using Hybrid 2D-3D Representations in Hierarchical PCA Space", BMVC99.

[18] N. Paragios, R. Deriche. "Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects", PAMI, Vol. 22, No. 3, March 2000.

[19] Sidenbladh, H., Black, M. J., and Fleet, D.J. "Stochastic tracking of 3D human figures using 2D image motion". ECCV, pp. 702-718 2000.

[20] L. Wiskott, J.M. Fellous, N. Krüger, C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching", PAMI, pp. 775-779, 1997.

[21] G. Wu, Y. Wu, L. Jiao, Y.-F. Wang and E. Chang. "Multi-camera spatio-temporal fusion and biased sequence-data learning for security surveillance", ACM International Conference on Multimedia, Berkeley, November 2003.

(a):before occlusion (b): in occlusion

(c): in occlusion (d): after occlusion

Figure 1 Before, in, and after temporary occlusion

(a): 1 leaving (b): 2 leaving

(c): 1 re-enter (d): 2 re-enter

Figure 2 Two persons leave and re-enter the scene