2010 Fourth Pacific-Rim Symposium on Image and Video Technology

Tracking Groups of People in Presence of Occlusion

Hamed Kiani Galoogahi
School of Computing
National University of Singapore
Singapore, Singapore
hkiani@comp.nus.edu.sg

***Abstract*—This paper addresses the problem of people group tracking in presence of occlusion as people form groups, interact within groups or leave groups. Foreground objects (a person or a group of people) from two consecutive frames are matched based on appearance (RGB histogram) and object location (2D region) similarity. While tracking, this method determines and handles some events such as objects merging and splitting using forward and backward matching matrices. The experimental results show that the proposed algorithm is efficient to track group of people in cluttered and complex environments even when total or partial occlusion occurs.**

Keywords- tracking, occlusion, RGB histogram, group splitting and merging, background subtraction

I. INTRODUCTION

People tracking in surveillance applications plays an important role to analyze people actions and behaviors such as standing in a forbidden area, running, jumping, and hiding. In addition, combining such information from two or more people may provide useful information about the interaction and behavior of the group. However, solving people tracking problem might be very difficult especially when moving people vary dynamically in a cluttered scene. Furthermore, the occlusion that occurs when people move in groups or interact with other people is another cause that makes the people tracking problem more difficult. Due to the importance of people tracking for visual surveillance, many algorithms in this area have been proposed [1-3]. Nevertheless, most of these methods just focus on tracking an individual person without considering that while tracking people might form groups, interact with each other, merge to larger groups or separate from groups. In addition, tracking people in unconstrained and cluttered environments requires robust methods that address the tracking problems caused by the partially or complete occlusions that usually causes considerable difficulty to many tracking schemes.

This paper focuses on group tracking when people merge to a larger group or leave groups by (1) updating the tracking model when an object (an object can be a person or a group of people) merges/splits into/to a larger/small objects and (2) utilizing the information of a group to address the occlusion problem while tracking a person in groups. Generally, the proposed method performs group tracking by detecting and following five possible events - *entering*, *merging*, *splitting*, *leaving* and *corresponding* - between two sets of objects between two consequent frames. The events detection is done according to appearance (RGB color histogram) and 2D region similarity. The reminder of this paper is organized as follows. Section 2 reviews some related works. Section 3 and 4 describe the group tracking problem and how the proposed algorithm solves it. Presenting the result of experiments and analyzing the performance of the proposed algorithm are done in section 5. Some final comments and conclusions are made in section 6.

II. RELATED WORK

Some systems for video surveillance perform people tracking based on the fact that each extracted connected component that belongs to foreground corresponds to a moving object [4,5]. Even though many efficient people tracking method have been proposed, tracking multiple moving people as a group in cluttered and dynamically varying scene is a challenging problem in the area of automated video surveillance. The problem of people tracking either in a group or individually can be defined as determining the appearance and location of a foreground objects in the sequence of frames. The challenges associated with people tracking can be caused by the people similarity in shape, color, size or occlusion by other people or background component. Generally, object tracking can be divided into four categories [6] as (1) Region based tracking, (2) Active contour based tracking, (3) Feature based tracking and (4) Modal based tracking.

In region based approaches [7,8], tracking is performed based on the variation of the image regions in motion. This approach does not require the whole object segmentation and feature extraction. Instead, it models a person by one or more small regions such as head, torso and the four limbs. Therefore, by tracking each small object, the moving human is tracked. However, this approach suffers from computational complexity, as it matches a window with all candidate windows in the next frame. Moreover, it cannot reliably handle occlusion between objects [6]. Using monochromatic image data, Hydra [9] attempts to detect the heads of people in groups and track them through occlusions. It uses silhouette-based shape models and temporal texture appearance models. Although effective in many situations, these 2D appearance models will not cope well with large

978-0-7695-4285-0/10 $26.00 © 2010 IEEE
DOI 10.1109/PSIVT.2010.80

rotations in depth during occlusions. Hydra is therefore quite effective at tracking people through occlusions as they walk past one another but it does not cope well when people leave a group in a different direction from that in which they entered it.

In contrast to region based tracking, active contour based tracking methods [10, 11] use very simple bounding contours to represent object's outline, which are updated dynamically in successive frames [6]. Moreover, using active contours as object descriptor increases the efficiency and reduces computational complexity. In the case of occlusion, even under disturbance or partial occlusion, these algorithms may track objects continuously. For example, the trackers based on 2D active shape models which have been used in [10,12] can only cope with moderate levels of occlusion. However, it is highly sensitive to the initialization of tracking that makes it inefficient start tracking automatically.

Model-based approach [13] requires developing a 2D or 3D model of human and tracking components of model. This is a robust approach for tracking and performs well under occlusion, but requires high computational cost. The Kalman filter has been used for object tracking frequently. It presumes that the behavior of a moving object could be characterized by a predefined model [14], and the model is usually represented in terms of its state vector. However, in a real world environment, we often face a situation where the predefined behavior model falls apart. It is possible that a target may disappear totally or partially due to occlusion by other objects. In addition, generally in object tracking, the sudden deformation of a target itself can cause the failure of the predefined behavior model of the Kalman filter.

In feature-based tracking [12], features of image objects are extracted for matching in sequence of frames. In this method, several features of objects are used in feature-vector for matching, such as size, position, velocity, ratio of major axis of best-fit ellipse [14], orientation, coordinates of bounding box etc. The feature-vectors can be compared by several techniques such as Euclidean distance [14] and correlation-based approach [6]. The histogram of RGB color components of image objects can also be used as feature and those histograms are compared for matching [15].

III. Problem Definition

The main goal of people group tracking is tracking people in unconstrained and cluttered environments as they form groups, interact and part from the group in presence of occlusion. In this section, after presenting some basic notations and definitions, we will formulate the people group tracking as follow.

A. Notations and Definitions

- f_t: f_t refers to the frame of an input video at time *t*. A frame is a 2D mapping $f_t : X \to v$ from $X = [x, y]$ to values $v = (r, g, b)$. (x, y) is the coordinate of a pixel in 2D space and (r, g, b) is the pixel value in RGB color space.
- Object: an object is a connected component belongs to foreground that can be a person or a group of people.
- h_i^t: $h_i^t =< n_b^{t,i} >,\ b = 1,...,B$ is the histogram of object O_i^t where $n_b^{t,i}$ is the number of pixels belong to object O_i^t whose value is that of the b^{th} bin. *B* is the number of bins.
- r_i^t: For each extracted object O_i^t of frame f_t, the 2D region r_i^t of this object on 2D Euclidean space is the (x, y) coordinates of pixels that belong to object O_i^t.
- $|f_t|$: is the number of objects that belong to the foreground of frame f_t.

Definition1. Object signature: given an object O_i^t, $\varphi_i^t =< r_i^t, h_i^t >$ is defined as the object signature of O_i^t where h_i^t is the object's histogram and $r_i^t = \{(x_n, y_n) \in O_i^t \mid 1 \le n \le N, 1 \le m \le M\}$ is the 2D region O_i^t in frame f_t. *N* and *M* are the size of f_t.

Definition2. Frame Model: given the frame f_t with *i* foreground objects $O_i^t, 1 \le i \le |f_t|$ and their corresponding object signatures φ_i^t, Frame Model $FM(f_t)$ is defined as $FM(f_t) = \{\varphi_i^t, 1 \le i \le |f_t|\}$.

B. Problem Formulation

Given two frames f_{t-1} and f_t with their corresponding frame models $FM(f_t) = \{\varphi_i^t, 1 \le i \le |f_t|\}$ and $FM(f_{t-1}) = \{\varphi_j^{t-1}, 1 \le j \le |f_{t-1}|\}$, respectively, for each objects O_j^{t-1} in frame f_{t-1} find the corresponding object O_i^t in frame f_t that maximizes the weighted similarity *S*:

$$S = \sum_{\substack{\forall \varphi_j^{t-1} \in FM(f_{t-1}) \\ \forall \varphi_i^t \in FM(f_t)}} w_r S_r^{i,j} + w_h S_h^{i,j} \quad (1)$$

where

$$S_r^{i,j} = \begin{cases} Overlap(r_i^t, r_j^{t-1}) & if \quad r_i^t, r_j^{t-1} \quad overlaps \\ -\infty & otherwise \end{cases} \quad (2)$$

and

$$S_h^{i,j} = \sum_{b=1}^{B} \frac{\sqrt{n_b^{t,i} \times n_b^{t-1,j}}}{(\sum_{k=1}^{B} n_k^{t,i}) \times (\sum_{k=1}^{B} n_k^{t-1,j})} \quad (3)$$

$S_h^{i,j}$ is Bhattacharyya coefficient that computes the histogram similarity between two objects and $Overlap(r_i^t, r_j^{t-1})$ is a function that returns the area of overlapped region between r_i^t and r_j^{t-1}. w_r and w_h are weights, $0 \le w_r, w_h \le 1, \quad w_r + w_h = 1$.

According to (1), the group tracking problem is performed by matching objects from the current frame with those of the previous one based on a two-criterion similarity measure. The first criterion is based on the 2D regions similarity of objects, S_r in (1). Basically, using the 2D region of objects as a matching criterion in two consecutive frames is inspired of this fact that the visual motions of objects are normally small relative to their spatial extents.

The second criterion which is used as matching criterion is RGB color histogram similarity. In fact, the aim of using appearance similarity based on color histogram is to enhance the group tracking process while matching based on 2D region is not efficient, especially when a group splits to smaller objects or a person leaves a group.

For example, in Fig. 1(a) two persons 1 and 2 can be tracked efficiently using 2D region criterion until they merge to new group, Fig. 1(b). After merging, persons 1 and 2 can be tracked as the new group 1+2 until they leave the group. However, after splitting and leaving the group, tracking persons 1 and 2 using only 2D region criterion is not feasible, because the new 2D regions at Fig. 1(c) don't overlap the last 2D regions at Fig. 1(a). Therefore, the algorithm performs tracking process using RGB histogram criterion (appearance similarity) to find the corresponding objects when using 2D region is not efficient.

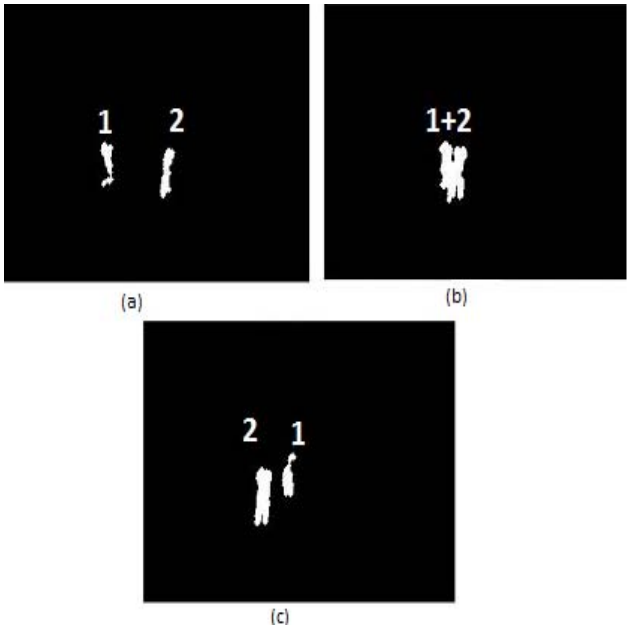

Figure 1. Detected objects (a) before forming a group, frame #128, (b) after forming a group, #213, (c) after splitting, frame #307.

IV. POPOSED METHOD

The proposed algorithm described here contains two steps: foreground segmentation and tracking. At foreground segmentation, to simplify the problem, we assume that the camera is stationary and that the background model is static. Generally, in this step given current frame f_t and a static background image reference I_b in RGB color space, as inputs, the aim is to extract moving objects that belong to foreground using a simple background subtraction operator.

The output of the subtraction operator on current frame f_t and the background image reference I_b is a binary foreground mask I_f which is calculated as below:

$$I_f(x,y) = \begin{cases} 1 & Dis(f_t, I_b)_{(x,y)} \succ \tau_{DisSim} \\ 0 & Otherwise \end{cases} \quad (4)$$

where, τ_{DisSim} is a predefined dissimilarity threshold and $Dis(f_t, I_b)_{(x,y)}$ is the Euclidean distance between *r*, *g* and *b* values of f_t and I_b on each corresponding pixel as follow:

$$Dis(f_t, I_b)_{(x,y)} = [(f_t^r(x,y) - I_b^r(x,y))^2 + (f_t^g(x,y) - I_b^g(x,y))^2 + (f_t^b(x,y) - I_b^b(x,y))^2]^{0.5} \quad (5)$$

Using the binary foreground mask, the coherent pixels (with value 1 at I_f) are grouped together as extracted objects by seeded region growing approach inspired by [16]. The idea used in this approach is similar to seeded region growing, but different in terms of number of regions and choosing seeds. We try to grow one region at a time until all connected neighboring pixels are considered and then start growing another region. After finding all image objects, smaller ones are discarded [17]. The minimum size of objects is determined by some heuristics and zoom of the camera. In our experiments, a minimum object size of 200 to 300 pixels worked well.

After segmentation, for each extracted object O_i^t, the object signatures $\varphi_i^t = < r_i^t, h_i^t >$ are calculated to construct the frame model $FM(f_t)$ which is required for tracking phase.

Let f_{t-1} and f_t be two consecutive frames. Suppose foreground segmentation step results in *U* objects in the first frame and *V* in the second one. Generally, the group tracking process for these two sets of objects is performed according to five different events: (1) *entering*, (2) *leaving*, (3) *merging*, (4) *splitting* and (5) *corresponding*. To find and control the events between both sets of objects, one 2D region matching matrix and two event detection vectors are defined as below:

Definition3. Region Match Matrix (RMM): Given two frame models $FM(f_t)=\{\varphi_i^t, 1\le i\le |f_t|\}$ and $FM(f_{t-1})=\{\varphi_j^{t-1}, 1\le j\le |f_{t-1}|\}$, a region match matrix $(RMM_{t-1}^t)_{U\times V}$ is defined as:

$$RMM_{t-1}^t[j,i]=\begin{cases}1 & if \quad r_i^t, r_j^{t-1} \quad overlap\\ 0 & otherwise\end{cases} \tag{6}$$

where $U=|f_{t-1}|$ and $V=|f_t|$.

Each entry $RMM_{t-1}^t[j,i]$ represents that the 2D region of object O_j^{t-1} overlaps object O_i^t or does not.

Definition4. Backward Matching Vector (BMV): given a region match matrix $(RMM_{t-1}^t)_{U\times V}$ over two frame models $FM(f_t)=\{\varphi_i^t, 1\le i\le |f_t|\}$ and $FM(f_{t-1})=\{\varphi_j^{t-1}, 1\le j\le |f_{t-1}|\}$, the backward matching vector BMV_t^{t-1} is a *U*1* column vector which defined as:

$$BMV_t^{t-1}[j]=\sum_{i=1} RMM_{t-1}^t[j,i], 1\le j\le U \tag{7}$$

Each entry $BMV_t^{t-1}[j]$ represents the number of objects at frame model $FM(f_t)$ that overlap the object O_j^{t-1} in frame model $FM(f_{t-1})$.

Definition5. Forward Matching Vector (FMV): given a RMM $(RMM_{t-1}^t)_{U\times V}$ over two frame models $FM(f_t)=\{\varphi_i^t, 1\le i\le |f_t|\}$ and $FM(f_{t-1})=\{\varphi_j^{t-1}, 1\le j\le |f_{t-1}|\}$, the forward matching vector FMV_{t-1}^t is a *1*V* row vector which defined as:

$$FMV_{t-1}^t[i]=\sum_{j=1} RMM_{t-1}^t[j,i], 1\le i\le V \tag{8}$$

Each entry $FMV_{t-1}^t[i]$ represents the number of objects at frame model $FM(f_{t-1})$ that overlap the object O_i^t in frame model $FM(f_t)$.

The group tracking algorithm performs tracking by finding the events among objects sets that belong to frame f_{t-1} and frame f_t according to the values of the backward and forward matching vectors. In addition to *BMV* and *FMV*, to preserve the history of the tracked objects from previous frames, the algorithm uses a temporary backup list that contains the information of new objects, object signatures and object labels.

Given *BMV*, *FMV* and backup list, tracking and event detection are performed as follow:

Entering: an object O_i^t enters to scene at current frame f_t if $FMV_{t-1}^t[i]=0$. In this case, for each object O_{BLst}^k in backup list, the histogram similarity between O_{BLst}^k and the object O_i^t is calculated using Bhattacharyya coefficient, $S_h^{i,k}$. If there is an object O_{BLst}^l with $S_h^{i,l}\succ T$ then $Label(O_t^i)=Label(O_{BLst}^l)$, the *T* value is initialized to 0.75 in our experiments. That means the entering detected object O_i^t had been detected in previous frames and leaved the scene for a while. The algorithm updates the histogram and 2D region of object O_{BLst}^l in backup list by information of object O_i^t.

Otherwise, the object O_i^t is detected as a new object. Thus, the algorithms generates and assigns a new label to the object O_i^t and updates the backup list using the histogram, 2D region and label of object O_i^t.

Corresponding: two objects O_i^t and O_j^{t-1} are corresponding if $FMV_{t-1}^t[i]=1$ and $BMV_t^{t-1}[j]=1$. In this case, $Label(O_t^i)=Label(O_{t-1}^j)$, and the histogram and 2D region of object O_{t-1}^j in backup list is updated by histogram and 2D region of object O_i^t.

Splitting: an object O_{t-1}^j splits to two or more objects O_i^t if $BMV_t^{t-1}[j]\succ 1$. When a splitting event is detected, each object O_i^t resulted by splitting event is tracked as an *entering* object.

Merging: two or more objects O_{t-1}^j merge to one object O_i^t if $FMV_{t-1}^t[i]\succ 1$. In this case, the new group O_i^t is tracked as a new group that contains merged objects O_{t-1}^j and $Label(O_t^i)=\bigcup_{\forall O_{t-1}^j} Label(O_{t-1}^j)$. In fact, when two or more objects merge and form a group, their information is kept but it is the new group that is tracked through the following frames.

Leaving: an object O_{t-1}^j leaves the previous frame f_{t-1} if $FMV_{t-1}^t[i]=0$. In this case, the information of this object is kept to be used for the next frames.

Generally, after finding the events, the tracking algorithm updates each new object using the information stored in the old ones, RGB color histogram, 2D region and the label of object which is used for tracking through frames. In addition, if a group splits or an object appears again after several frames, the algorithm uses object's appearance

characteristics – RGB color histogram- to identify correctly the two splitting or the new objects.

V. EXPERIMENT RESULTS

This method is implemented in MATLAB 7 running on a Pentium IV 2.8 GHz PC with 2 GB memory. The image frames extracted from video had a size of 240x320. We performed several experiments on different scenarios such as tracking (1) a single person walking in the video, (2) multiple people as individual walking in the video and (3) a group of people walking as a group, where the people may be occluded partially or totally, and they may merge to a new group or leave the group. A subset of results is shown here. As the first experiment, we tested a video with two people who walk individually, form a two persons group and then split. The result is shown in Fig. 2 with a bounding box shown for each tracked individuals and groups.

As can be seen in Fig. 2(a), this algorithm tracks multiple people efficiently when they walk individually based on 2D region matching, in this case two events *corresponding* and *entering* have been detected. Moreover, by detecting different events such as *merging* and *splitting* using *FMV* and *BMV* vectors, it can perform tracking the new group in presence of partial, Fig. 2(b), and total, Fig. 2(c), occlusions. Furthermore, Fig. 2(d) represents that RGB histogram matching is appropriate when the 2D region matching is not efficient for individual tracking.

For second experiment, we test the proposed algorithm on a more complicated experiment on a video with three persons who walk individually, Fig. 3(a1), form two and three persons groups with complete and partial occlusions, Fig. 3(a2-a5), and then split up again to three individuals, Fig. 3(a6).

Fig. 3(a3) shows that regardless an object is an individual or a multi people group, the proposed algorithm performs tacking without any preprocessing to find whether a tracked object is an individual or a group. That means this algorithm is efficient for both individual and group tracking. In addition, this experiment states that while tracking a person in a group, the amount of occlusion does not influence the performance of tracking, tracking person 3 in Fig. 3(a4,a5).

VI. CONCLUSION

In this paper, we present a new algorithm for people group tracking in a cluttered and unconstrained environment when partial or total occlusions occur. For object segmentation and moving object extraction, we use a simple background subtraction operator on RGB color space, assuming that the background model is static. While tracking, objects from two consecutive frames are matched based on finding five possible events of objects-*entering*, *merging*, *splitting*, *corresponding* and *leaving*- that might happen while people tracking. For two consecutive frames with two sets of foreground objects, finding events is performed using forward and backward matching vectors based on 2D region criterion. In order to increase the performance of tracking when 2D region criterion is not efficient while *entering* or *splitting*, this algorithm uses RGB histogram similarity to continue tracking process. The results shows that this algorithm is robust in presence of occlusion and can perform people group tracking when group changes due to merging and splitting events.

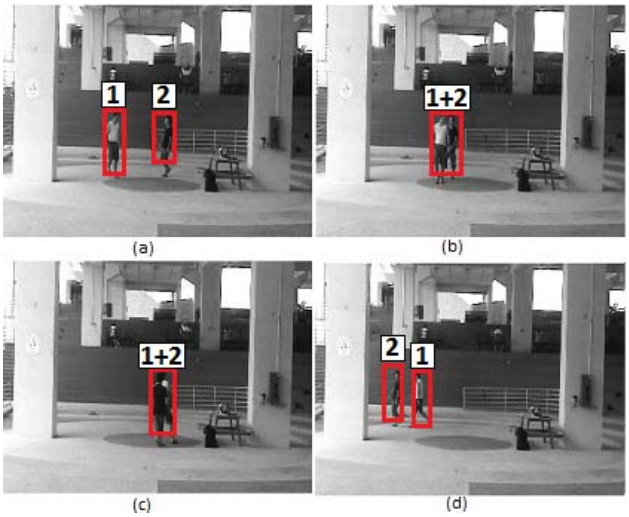

Figure 2. Tracking persons (a) before forming group using 2D region criterion, (b) within a group with partial occlusion, (c) within a group with total occlusion, (d) after splitting using RGB histogram criterion.

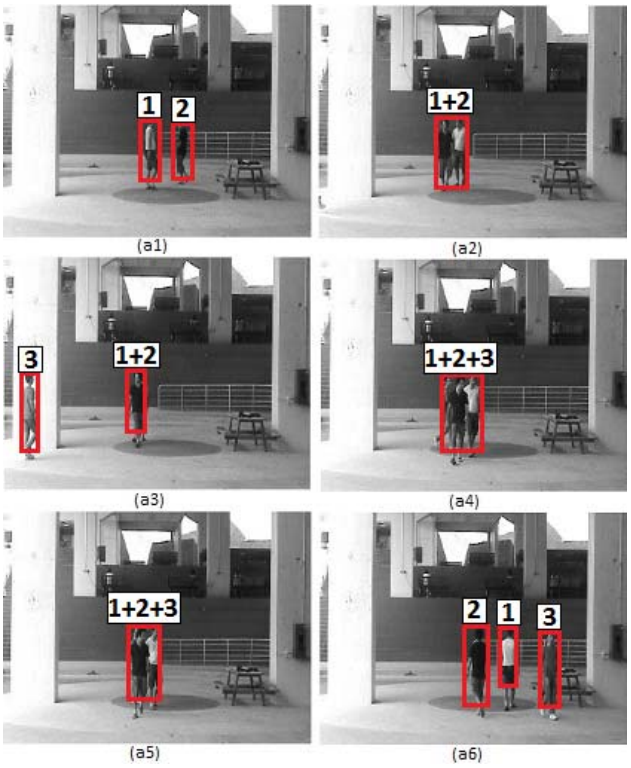

Figure 3. (a1,a2) tracking two persons, individually and as a group, (a3) tracking a group and an individual, (a4,a5) tracking a three person group with total and partial occlusion, (a6) tracking three person individually after splitting.

REFERENCES

[1] Bremond and M. Thonnat, "Tracking multiple non-rigid objects in a cluttered scene. In Proceedings of the 10th Scandinavian Conference on Image Analysis (SCIA '97), vol. 2, pp. 643-650, Finland, 1997.

[2] D. M. Gavrila and L. S. Davis, "Tracking of humans in action: A 3-D model-based approach", In ARPA Image Understanding Workshop, (Palm Springs). Pp. 737-746, 1996.

[3] Q. Cai, A. Mitiche, and J. K. Aggarwal, "Tracking human motion in an indoor environment", In Proceedings of the 2nd International Conference on Image Processing (ICIP'95), pp. 215-218, 1995.

[4] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions", In International Conference on Vision Systems, 1999.

[5] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking", In IEEE Conference on Computer Vision and Pattern Recognition, pp. 246–252, 1999.

[6] W. Hu, T. Tan, L. Wang, S. Maybank, "A survey on visual surveillance of object motion and behaviors", Systems, Man and Cybernetics, Part C, vol. 34, 3rd ed., 2004.

[7] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y Tsin, D. Tolliver, N. Enomoto, "A system for video surveillance and monitoring", Technical Report CMU-RI-TR-00-12, CMU, 2000.

[8] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, H. Wechsler, "Tracking groups of people. Computer Vision and Image Understanding", vol. 80, pp. 42-56, 2000.

[9] I. Haritaoglu, D. Harwood, L. S. Davis, "Hydra: multiple people detection and tracking using silhouettes", International Conference on Image Analysis and Processing, pp. 280- 285, 1999.

[10] M. Isard, A. Blake, "CONDENSATION—Conditional Density Propagation for Visual Tracking", International Journal of Computer Vision, vol. 29, 1th ed., pp. 5-28, 1998.

[11] N. Paragios, R. Deriche. Geodesic active regions for motion estimation and tracking, ICCV 99.

[12] R. Polana, R. Nelson, "Low level recognition of human motion (or how to get your man without finding his body parts)", IEEE Workshop on Motion of Non-Rigid and Articulated Objects, 1994.

[13] I. A. Karaulova, P. M. Hall, A. D. Marshall, "A hierarchical model of dynamics for tracking people with a single video camera", In Proc. British Machine Vision Conf, pp. 262-352, 2000.

[14] L. Xu, J. L. Landabaso, B. Lei, "Segmentation and tracking of multiple moving objects for intelligent video analysis", BT Technology Journal, vol. 22, 3rd ed., 2004.

[15] D. Comaniciu, V. Ramesh, P. Meer, "Real-time tracking of non-rigid objects using mean shift", Computer Vision and Pattern Recognition, 2000.

[16] R. Adams, L. Bischof, "Seeded region growing", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.16, 6th ed., pp. 641-647, 1994.

[17] M. Xu, T. J. Ellis, "Partial observation vs. blind tracking through occlusion", In Proc of BMVC'2002, Cardiff, pp. 777-786 , 2002.