

Tracking Multiple People for Video Surveillance

M. A. Ali, S. Indupalli and B. Boufama

School of Computer Science, University of Windsor

Windsor, ON N9B 3P4, Canada

Email: ali12w, indupal, boufama@uwindsor.ca

Abstract

This paper addresses the problem of detecting and tracking multiple moving people in a complex environment with unknown background. In this paper, we propose a new correlation-based matching technique for feature-based tracking. Our method was compared with two existing matching techniques, namely the normalized Euclidean distance and histogram-based matching. Experimental results on real-images suggest that our correlation-based approach is more accurate and efficient than the other two approaches.

1. Introduction

Automatic visual surveillance in dynamic scenes (both in indoor and outdoor environment) has recently got a considerable interest to researchers [9]. Technology has reached a stage where mounting video camera is cheap causing a widespread deployment of cameras in public and private areas [22]. Finding available human resources to sit and watch the imagery is too expensive for most organizations to afford the cost of human operators [3]. Moreover, surveillance by operators is error prone due to fatigue, negligence and lack of ubiquitous surveillance. Therefore, it is important to develop an accurate and efficient automatic video analysis system for monitoring human activity that will create enormous business opportunities. It will allow us to detect unusual events in the scene and warrant the attention of security officers to take preventive actions [22]. The purpose of visual surveillance is not to replace human eyes with camera, but to accomplish the entire surveillance task as automatic as possible [9]. Other applications of automatic video surveillance include preventing theft at parking and shopping areas [22], detecting robbery in bank and secured places [3], detecting camouflage [19] etc.

The automatic video surveillance system has two major components, they are detecting moving objects and tracking them in sequence of video images. The accuracy of these components largely affects the accuracy of overall surveillance system. Detecting moving regions in the scene and separating them from background image is a challenging problem. In the real world, some of the challenges associated with foreground object segmentation are illumination

changes, shadows, camouflage in color, dynamic background and foreground aperture [20]. Foreground object segmentation can be done by three basic approaches: frame differencing, background subtraction and optical flow. Frame differencing technique does not require any knowledge about background and is very adaptive to dynamic environments [3], but suffers from the problem of foreground aperture due to homogeneous color of moving object. Background subtraction can extract all moving pixels, but it requires perfect background modeling. It is extremely sensitive to scene changes due to lighting and movement of background object. Optical flow is the most robust technique to detect all moving objects, even in the presence of camera motion, but it is computationally expensive and cannot be used for real-time systems.

Tracking multiple moving people in cluttered video sequences is another challenging problem in the area of automated video surveillance. It is the building block of understanding high-level events and complex actions such as detection of walking, running, dancing, stalking etc. The problem of tracking can be stated as determining the appearance and location of a particular object in the sequence of frames. The challenges associated with tracking are similarity of people in shape, color and size, proximity of other people and occlusion by other people or background component. Tracking also requires proper management of appearance or disappearance of objects (which changes total number of objects being tracked).

Object tracking methods can be divided into 4 groups [9], they are:

1. Region-based tracking
2. Active-contour-based tracking
3. Feature-based tracking
4. Model-based tracking

In region-based approach [13, 3], tracking is performed based on the variation of the image regions in motion. This approach does not require computation of image blobs and feature extraction, but it 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 [9]. In addition, it fails to a match an object when it moves beyond a region. In contrast to region-based tracking, objects are more simply described in active contour-based tracking [14, 11]. Here, bounding contours are used to represent object's outline, which are updated dynamically in successive frames [9]. This approach is

sensitive to initialization and limited to tracking precision. Model-based approach [12] 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. In feature-based tracking [15], features of image blobs are extracted for matching in sequence of frames. In this method, several features of blobs are used in feature-vector for matching, such as size, position, velocity, ratio of major axis of best-fit ellipse [22], orientation, coordinates of bounding box etc. The feature-vectors can be compared by several techniques such as Euclidean distance [22] and correlation-based approach [7]. The histogram of RGB color components of image blobs can also be used as feature and those histograms are compared for matching [4].

In this paper, we propose a new method for matching features of blobs in conjunction with a tracking system. Our system is briefly as follows: the background is modeled by statistical method and updated continuously. Foreground object segmentation is performed by background subtraction and K-means clustering. We used HSV color space to minimize cast shadows. After finding legitimate blobs, features are extracted and compared with features of blobs in the previous frame using Pearson correlation-based approach. Best matched blob is identified by considering maximum correlation coefficient.

This paper is organized as follows: section 2 describes some previous work in this area. Section 3 describes our proposed tracking system in details. Section 4 and 5 show some experimental results and compare the results. The paper concludes in section 6 with some future research direction.

2. Related Work

2.1 Background Modeling and Foreground Object Segmentation

Most of the work on foreground object segmentation is based on three basic methods, namely frame differencing, background subtraction and optical flow. Only background subtraction requires modeling of background. It is faster than other methods and can extract maximum features pixels. In [3], Collins *et al.* used a hybrid of frame differencing and background subtraction for effective foreground segmentation. In the literature, a lot of work has been done on modeling dynamic background. Researchers usually use Gaussian [13], a mixture of Gaussian [17], kernel density function [6] or temporal median filtering techniques for modeling background [23]. Background can be modeled in different color spaces for resolving some of the challenges associated with background issues. In [2], Chen *et al.* modeled the background in HSV color space to eliminate shadows. Horprasert *et al.* [8] used a color constancy model for shadow detection, assuming that

the chromaticity remains same while only intensity differs between shadow and background.

2.2 Tracking

A feature-based object tracking algorithm requires useful feature selection, feature extraction, feature matching and proper handling of object's appearance and disappearance. In [3], Collins *et al.* described a basic object tracking algorithm and tracking hypothesis. An effective management of object entry and exit was proposed by Stauffer [18]. Most of the works on tracking use a prediction on features in the next frame and compare the predicted value with estimated value to update the model. Usually a model like Kalman filter [22] is used for prediction.

Regarding matching of features, Xu *et al.* used a scaled Euclidean distance function for matching [22]. In [3], Collins *et al.* used a correlation function for matching regions in motion. Haritaoglu *et al.* used sum of absolute difference (SAD) as correlation score during matching [7]. In [4], the authors proposed a mean-shift technique to calculate most probable target position. They calculated similarity of objects by constructing histograms of target model and target candidates. Similarity is expressed by a metric derived from the Bhattacharyya coefficient. Some of the other techniques used in tracking are geodesic method [14], condensation method [11] and dynamic Bayesian network [9].

3. Our Tracking System

3.1 Background Modeling and Foreground Object Segmentation

To model the background, we used a statistical method that was proposed in our earlier work in [10]. The background image is constructed based on the statistical observation of pixel intensities of both the foreground and the background simultaneously. For every pixel, we developed a histogram of RGB color and considered the color with highest frequency. We used background subtraction for identifying regions where the objects are moving. We performed background subtraction in HSV color space, as HSV color space works well against shadow [16, 2]. We utilized the advantages of all the components of HSV color space to get more accurate result. We considered HSV color space for moving region segmentation; later we used RGB color space for feature calculation and histogram analysis. During background subtraction, finding a good threshold value is a major problem. If we take a smaller value for T to consider all the pixels in a moving region, then we introduce noise and shadow in the resultant image. If we increase T to remove shadow and noise, then we remove the self shadow region of the moving people and the image blob gets distorted. As suggested in [3], we can use different threshold value for different pixels and update them dynamically, instead of taking one global threshold. This approach is

computationally expensive. To avoid these problems, we adopted K-means clustering technique for segmenting foreground pixels from background. In this approach, first we calculate a difference matrix by subtracting background image from a frame. K-means clustering is then applied to the difference matrix to separate all the pixels into 2 clusters, a background cluster and a foreground cluster. This approach is highly efficient and eliminated the requirement of threshold. After finding the moving regions, the noise is removed by morphological operations (erosion and dilation).

3.2 Feature Extraction

3.2.1 Finding image blobs

The coherent pixels are grouped together as image blob by seeded region growing approach inspired by [1]. 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 neighbouring pixels are considered and then start growing another region. After finding all image blobs, smaller ones are discarded [21]. The minimum size of blobs is determined by some heuristics and zoom of the camera. In our experiments, a minimum blob size of 200 to 300 pixels worked well.

3.2.2 Finding features of blobs

In our method, we considered following significant features of blobs for matching during Euclidean distance-based approach and correlation-based approach:

- Size of blob
- Average of individual RGB components
- Coordinate of center of blob
- Motion vector

Size of the blob is represented as total number of pixels in the blob. The motion vector is calculated by taking the difference between coordinates of centers of blobs in two consecutive frames. Histogram of RGB color components was used during histogram-based matching. In the histogram, we considered a bin size of 10 and hence, there were a total of 26 bins for each color component. The size of the bin was taken based on heuristics. A bin size of 1 is not computationally feasible. Moreover, they are very sensitive to slight variations of color. Taking a large bin size will work poorly during matching. As large sized blob have larger frequency count in histogram and vice versa, we normalized the values of the histogram within 1 by dividing the value by size of blob. All other features are also normalized to 1 before matching. For example, the size of the blob is divided by total size of images (240x320 in our case). Similarly the average color components are divided by 256 to normalize within 1. The coordinates of center of blob are normalized by dividing each dimension of image.

3.3 Tracking People

We developed our tracking system based on the basic tracking algorithm proposed by Collins *et al.* [3], which is as follows:

1. Predict positions of known objects
2. Associate predicted objects with current objects
3. If tracks split, create new tracking hypothesis
4. If tracks merge, merge tracking hypotheses
5. Update object tracking models
6. Reject false alarms

Object type classification is not discussed in this paper. We assumed that the objects are all human. The classification of objects can be done before or after object tracking. Most of the tracking system is built on the basis of this algorithm, and therefore use prediction of features in the next frame. It reduces the search space, but predicting features requires use of a predictor like Kalman filter. It requires significant computation time to build and update the model. In our system, we skipped the prediction of features to save computation time; rather we compared features obtained in the previous frame with features obtained in the current frame.

3.3.1 Matching blobs

Tracking is performed by matching features of blobs in the current frame with the features of the blobs in the previous frame. Suppose we have N number of blobs in the current frame and M number of blobs in the previous frame. We do an exhaustive matching among N blobs in the current frame with M blobs in the previous frames, so a total of $N \times M$ matching is required. As we do not have a lot of objects in the scene, this exhaustive matching is not time consuming. In our experiments, we applied 3 kinds of matching techniques and compared their results. They are normalized Euclidean Distance, Pearson correlation coefficient and sum of absolute histogram difference. In the case of normalized Euclidean distance and Pearson correlation coefficient, the features are represented as feature vector and matching is performed on those feature vectors. Suppose we have two feature vectors E_i and E_j , $i=1$ to N , $j=1$ to M , the normalized Euclidean distance is calculated by equation 1.

$$Dist(E_i, E_j) = \sqrt{\frac{1}{d} \sum_{k=1}^d (E_{ik} - E_{jk})^2} \quad \dots (1)$$

Here d is the dimension of the vector (which is 8 in our case; 1 for size, 3 for 3 color components, 2 for coordinates of center of blobs and 2 for motion vector). The problem of Euclidean distance is that, the feature which has a higher value dominates others. To solve this problem, Xu *et al.* [22] suggested to use Mahalanobis distance, but it is computationally expensive. To avoid this problem and to give importance to significant features, we used different weight factors for different

features. For example, during tracking people, size of the blobs will be close to each other. So, this feature should be given less weight (0.05 in our case) compared to other strong cues like color and center of blobs (0.15 in our case). The motion vectors are given weight 0.15 for each. The best match is found by considering smallest Euclidean distance.

To calculate the correlation between features, we used Pearson correlation coefficient, which has been used successfully in Bioinformatics for finding similarity of gene expression [5]. The formula for Pearson correlation coefficient takes many forms; one of them is showed in equation 2.

$$Corr(E_i, E_j) = \frac{1}{d} \sum_{k=1}^d \frac{(E_{ik} - \bar{E}_i)(E_{jk} - \bar{E}_j)}{\sigma_{E_i} \sigma_{E_j}} \quad \dots (2)$$

Here \bar{E}_i and \bar{E}_j indicates mean and σ_{E_i} and σ_{E_j} indicates standard deviation, calculated by

$$\sigma = \frac{\sum_{k=1}^d (E_k - \bar{E})^2}{d}$$

Here d indicates dimension of the vector. Higher value of the coefficient indicates higher chances to be best matched candidate.

We implemented the histogram-based matching technique using following equation of SAD [7].

$$Hist(H_i, H_j) = \sum_{r=1}^n \sum_{s=1}^n \sum_{t=1}^n |H_{irst} - H_{jrst}| \quad \dots (3)$$

Where n =number of bins and H_i and H_j are histogram of Blobs i and j . The lower the value, the higher is the similarity of histograms between blobs. After finding the summation, the result is normalized to 1 by dividing by 26^3 .

In all cases, although the minimum or maximum value indicates best match, the best matching value has to be less than or greater than a certain threshold. Otherwise, a new object hypothesis is created. The threshold values 0.08 for histogram-based matching, 0.9 for correlation-based matching and 0.08 for Euclidean-distance based matching worked well in our experiments.

4. Experimental Result

We have implemented our method in Matlab 7 running on a Pentium IV 2.79 GHz workstation and having 256 MB memory. The image frames extracted from video had a size of 240x320. We used Matlab's image processing toolbox for image I/O and morphological operations. We performed several experiments and a subset of results is shown here. At first we experimented with two people moving towards

each other. The tracking results using three different matching techniques are shown in figure 1, 2 and 3.

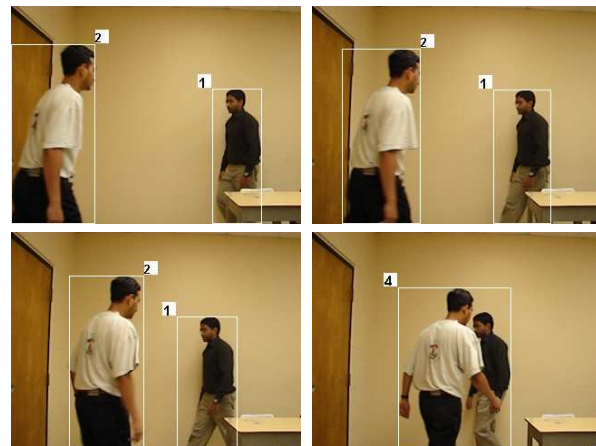


Figure 1: Tracking 2 people using histogram-based approach

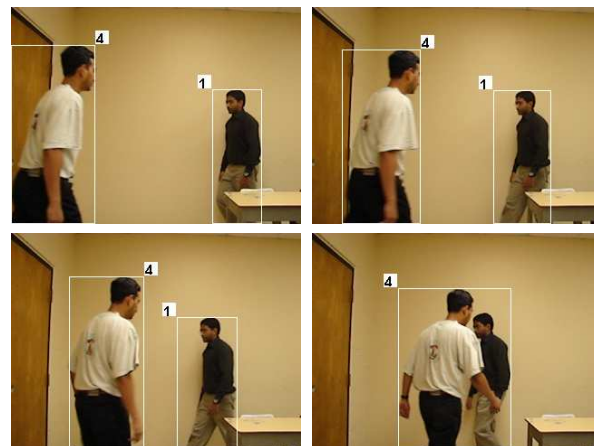


Figure 2: Tracking 2 people using correlation-based approach



Figure 3: Tracking 2 people using Euclidean distance-based approach

We also experimented with 3 people moving independently in a classroom. The results are shown in figure 4, 5 and 6.

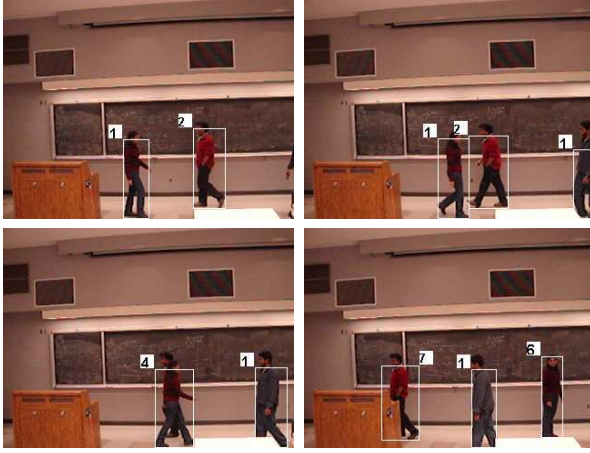


Figure 4: Tracking 3 people using histogram-based approach (frames 29, 32, 35, 37)

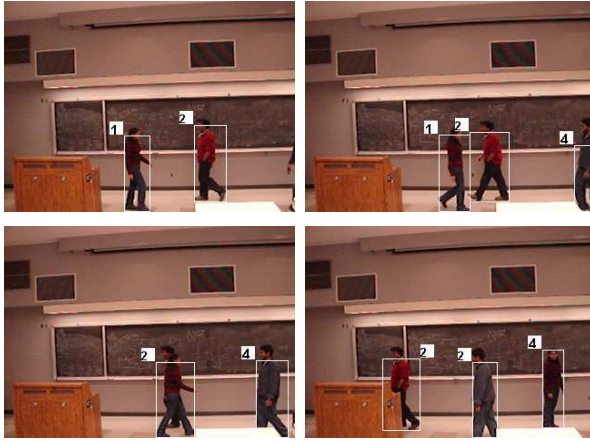


Figure 5: Tracking 3 people using correlation-based approach (frames 29, 32, 35, 37)

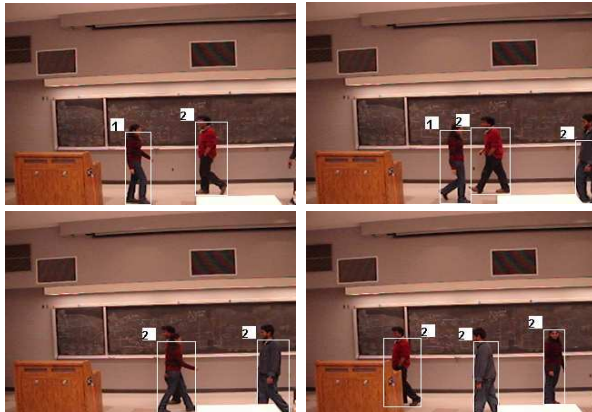


Figure 6: Tracking 3 people using Euclidean distance-based approach (frames 29, 32, 35, 37)

A subset of matching values obtained by histogram-based matching, correlation-based matching and

Euclidean distance based matching is shown in table 1, 2 and 3. The best matched values are highlighted.

Table 1: Value of Histogram Comparison (T=0.08)

Previous blobs \ Current blobs	1	2	3
Frame 32			
1	0.11350	0.0487	0.1096
2	0.0504	0.10855	0.0685
3	0.08881	0.12158	0.0361
Frame35			
1	0.14519	0.04221	
2	0.07851	0.10537	
Frame37			
1	0.13000	0.11202	
2	0.13427	0.12869	
3	0.19640	0.07942	

Table 2: Value of Correlation coefficient (T=0.9)

Previous blobs \ Current blobs	1	2	3
Frame 32			
1	0.946725	0.997585	0.863
2	0.999578	0.948117	0.815
3	0.606269	0.717308	0.999
Frame35			
1	0.720732	0.999498	
2	0.995875	0.882008	
Frame37			
1	0.790499	0.985667	
2	0.905196	0.788276	
3	0.983655	0.90053	

Table 3: Value of Euclidean Distance (T=0.08)

Previous blobs \ Current blobs	1	2	3
Frame 32			
1	0.037772	0.013068	0.078
2	0.006671	0.052386	0.077
3	0.075114	0.091296	0.007
Frame35			
1	0.07013	0.00583	
2	0.017304	0.059558	
Frame37			
1	0.049365	0.036416	
2	0.053866	0.110335	
3	0.052481	0.064524	

To verify the accuracy of our correlation-based tracking system, we performed a different kind of experiments. After tracking a person successfully, we went back to the first frame and removed one person

from the sequence of images. Figure 7 shows the result of this experiment after removing the left most people.

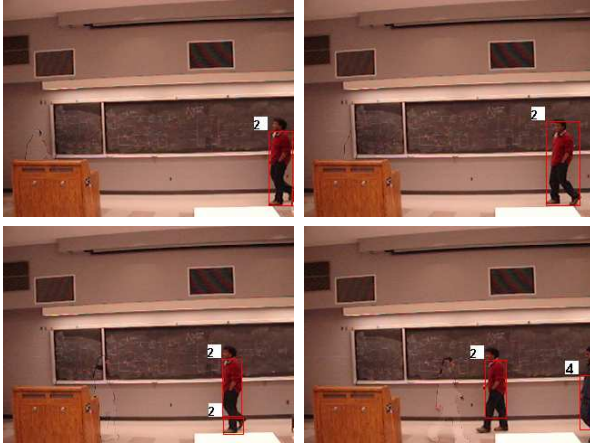


Figure 7: Sequence of frames after removing left-most people

5. Comparison of Results

The tracking method is very sensitive to error in the middle of a sequence. If any object is incorrectly tracked in the middle of sequence, this error propagates in the next frames and other objects get tracked improperly. From our experiments, we have found that correlation-based approach and histogram-based approach does not give similar results in some cases. For example, see figure 8(a) and 8(b). In these figures, only foreground images are shown. In figure 8, we can see that the image blob for right-most person is split into two parts due to segmentation error. As these blobs are close to each other and have a similar motion vector, they are considered similar image blobs when compared to previous frame. But for the case of histogram-based method, the split objects lower part best match with left-most person, so it shows that it is part of that object, which is an error. This is one of the disadvantages of histogram-based method, as it does not consider the location and motion of image blobs during matching.

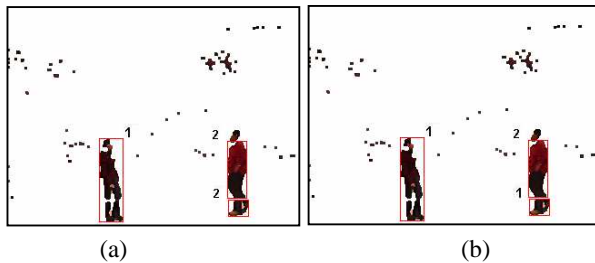


Figure 8: Foreground image obtained from (a) correlation based method (b) histogram based method

From our experiments, we also found a problem with correlation-based method. In figure 9, we can see that the person identified with blob 4 is numbered as blob 2 in the next frame. Due to lower frame rate, the object has moved a lot and reached near blob 2. So, it found

close match with blob 2 in terms of location and motion. In case of histogram (figure 10), that object was identified correctly in the next frame. These results can be verified from the matching data available in table 1 and 2. From experiments, we found that the normalized Euclidean distance-based approach performs poorly compared to other two methods in all video sequences.

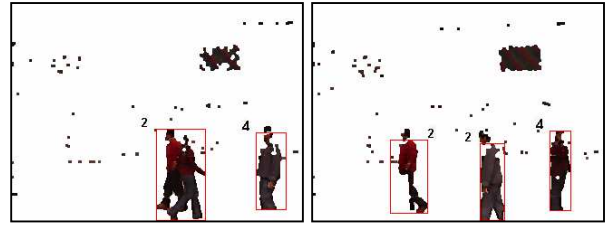


Figure 9: Foreground images in 2 consecutive frames obtained from correlation-based method

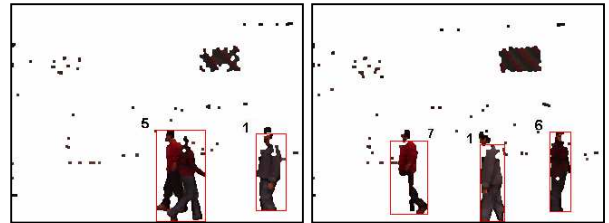


Figure 10: Foreground images in 2 consecutive frames obtained from histogram-based method

6. Conclusion and Future work

In this paper, we have presented methods of background modeling, segmentation of foreground object by background subtraction and tracking of multiple people in indoor environment. We selected background subtraction method, because it gives maximum number of moving pixels. We used feature-based tracking, as it is faster than other methods. We implemented three matching techniques and propose to use Pearson correlation coefficient for matching features, as it gives better results than histogram-based approach and Euclidean distance-based approach. The comparison of experimental results suggests that we have to combine the correlation-based approach and histogram-based approach to get more accurate result. We also have to give more emphasis on color features during correlation based matching. Future work can also be done on finding good threshold value during matching for creating new object hypothesis and minimum size of blobs. From the comparison of experimental result, we found that our correlation-based method did not work well in some cases. It might be the case that the selection of features was not accurate or the normalization technique was not correct. In future, we will investigate more on these issues.

In this paper we focused on tracking multiple people, but our correlation-based approach can be used for tracking any moving objects. Tracking can be done on

individual body parts like head, hands, legs etc for higher-level analysis of human activity.

In the future, we will address occlusions in the tracking process. As suggested by Xu *et al.* [22], we might be using blob-bounding box and motion information for detecting future occlusion and keeping track of blobs. Histogram-based approach should further be investigated for detecting partial occlusions by considering sub-blob matching.

References

- [1] Adams, R.; Bischof, L.; "Seeded region growing", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 16, Issue 6, June 1994 Page(s):641 - 647
- [2] Chen, B.; Lei, Y.; "Indoor and outdoor people detection and shadow suppression by exploiting HSV color information", 4th International Conference on Computer and Information Technology, 14-16 Sept 2004, Page(s):137 - 142
- [3] Collins, R. T.; Lipton, A. J.; Kanade, T.; Fujiyoshi, H.; Duggins, D.; Tsin, Y.; Tolliver, D.; Enomoto, N. and Hasegawa, O.; "A system for video surveillance and monitoring", *Technical Report CMU-RI-TR-00-12, CMU*, 2000
- [4] Comaniciu, D.; Ramesh, V.; Meer, P.; "Real-time tracking of non-rigid objects using mean shift", *Computer Vision and Pattern Recognition*, 2000
- [5] Dirks, W.; Yona, G.; "A comprehensive study of the notion of functional link between genes based on microarray data, promoter signals, protein-protein interactions and pathway analysis", *Technical Report*, 2003
- [6] Elgamal A.; Duraiswami R.; Harwood D. and Davis L.; "Background and foreground modelling using nonparametric kernel density estimation for visual surveillance", *Proc of the IEEE*, 90, No 7 (July 2002).
- [7] Haritaoglu, I.; Harwood, D.; Davis, L. S.; "Hydra: multiple people detection and tracking using silhouettes", *International Conference on Image Analysis and Processing*, 27-29 Sept. Page(s):280 – 285
- [8] Horprasert T.; Harwood D. and Davis L.; "A statistical approach for real-time robust background subtraction and shadow detection", *Proc of ICCV'99 FRAME-RATE Workshop (1999)*
- [9] Hu, W.; Tan, T.; Wang, L.; Maybank, S.; "A survey on visual surveillance of object motion and behaviors", *Systems, Man and Cybernetics, Part C*, Volume 34, Issue 3, Aug. 2004
- [10] Indupalli, S.; Ali, M. A.; Boufama, B.; "A Novel Clustering-Based Method for Adaptive Background Segmentation", *Technical Report TR05-027, University of Windsor*, 2005
- [11] Isard, M.; Blake, A.; "CONDENSATION—Conditional Density Propagation for Visual Tracking", *International Journal of Computer Vision* 29(1), 5–28 (1998)
- [12] Karaulova, I. A.; Hall, P. M.; Marshall, A. D.; "A hierarchical model of dynamics for tracking people with a single video camera," *In Proc. British Machine Vision Conf.*, 2000, pp. 262–352.
- [13] McKenna, S. J.; Jabri, S.; Duric, Z.; Rosenfeld, A.; Wechsler, H.; "Tracking groups of people", *Computer Vision and Image Understanding*, 80, pp 42—56 (2000).
- [14] Paragios, N.; Deriche, R.; "Geodesic Active Regions for Motion Estimation and Tracking", *ICCV 99*
- [15] Polana, R.; Nelson, R.; "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
- [16] Renno, J. P. R.; Orwell, J.; Jones, G. A.; "Evaluation of shadow classification techniques for object detection and tracking", *ICIP 2004*
- [17] Stauffer, C.; Grimson, W. E. L.; "Adaptive background mixture models for real-time tracking", *Proceedings of CVPR*, Jun 1999, pp. 246-252.
- [18] Stauffer C.; "Estimating tracking sources and sinks", *Proc of 2nd IEEE Workshop on Event Mining (in conjunction with CVPR'2003)*, 4, Madison, Wisconsin (June 2003).
- [19] Tankus, A.; Yeshurun, Y.; "Convexity-based visual camouflage breaking", *Computer Vision and Image Understanding*, 82(3):208-237, June 2001
- [20] Toyama, K.; Krumm, J.; Brumitt, B.; Meyers, B.; "Wallflower: principles and practice of background maintenance", *7th IEEE International Conference on Computer Vision*, Volume 1, 20-27 Sept. 1999 Page(s):255 - 261
- [21] Xu, M.; Ellis, T. J.; "Partial observation vs. blind tracking through occlusion", *In Proc of BMVC'2002, Cardiff*, pp 777—786 (September 2002).
- [22] Xu, L.; Landabaso, J. L.; Lei, B.; "Segmentation and tracking of multiple moving objects for intelligent video analysis", *BT Technology Journal*, Vol 22, No 3, July 2004
- [23] Zhou, Q.; Aggarwal, J. K.; "Tracking and classifying moving objects from video", *Proc of 2nd IEEE Intl Workshop on Performance Evaluation of Tracking and Surveillance (PETS'2001)*, Kauai, Hawaii, USA (December 2001).