Real-time Multiple Head Shape Detection and Tracking System with Decentralized Trackers

Jacky S. C. Yuk   Kwan-Yee K. Wong   Ronald H. Y. Chung   F. Y. L. Chin   K. P. Chow
Department of Computer Science
The University of Hong Kong
Pokfulam, Hong Kong
scyuk@cs.hku.hk , kykwong@cs.hku.hk , hychung@cs.hku.hk , chin@cs.hku.hk , chow@cs.hku.hk

Abstract

This paper presents a robust human tracking system which incorporates automatic detection of head shape objects with decentralized tracking approach. A fast and robust probabilistic shape contour matching algorithm is applied to the input image frame to detect and locate head shape objects. The detected objects are then tracked by decentralized trackers. Here, a decentralized tracker refers to the tracker that tracks exactly one object. Essentially, each newly detected object will instantiate an individual tracker, which tracks the object and destroys itself when the object disappears. Two trackers communicate with each other only when they are getting close enough. This approach simplifies the competition of targets between trackers, and is more efficient than the centralized approach whose time complexity is greatly depends on the number of tracked objects. The system has been tested with several challenging digital surveillance video sequences, and the results show the robustness and the efficiency of the system under crowded and clutter environment.

1 Introduction

In recent years, multiple targets tracking systems have been widely applied in video surveillance applications. These intelligent applications help monitoring public area, counting interested objects passing through, reporting any suspicious behavior, etc. As more and more intelligence is sought by video surveillance applications, there is an increasing demand for more robust tracking systems which can automatically detect and track multiple interested targets in real-time. For instance, real-time human detection and tracking is one of the key issues [3, 5, 9, 13, 14, 16, 17].

In general, human detection and tracking is an extremely difficult problem due to the articulation of human body. Some researchers used a fully descriptive 3D human body model to track human motions [10]. However, this approach is highly computational intensive, and therefore, is not applicable to real-time applications. A faster approach is to employ motion model to analyze and track the object motion. e.g., Haritaoglu et al. [3] proposed to identify human by analysis of periodic motion. However, it is a difficult task to capture accurate object motion in a crowded and clutter environment.

Another approach is to utilize pre-defined shape or template model in detection and tracking algorithm. [2, 6, 7, 15, 17]. For instance, Yiu et al. [15] proposed a 2.5D contour approach for vehicle tracking. Zhao and Nevatia [17], on the other hand, used multiple hypotheses, including vertical projection and omega (Ω) shape matching, to build a simplified human shape model. Their approach can successfully track multiple human objects in a crowded situation, however, not in real-time. Blake and Isard [2] demonstrated a fruitful approach of active contour tracking with a pre-defined shape space. MacCormick [6, 7] introduced a probabilistic model to further extended this approach with condensation (particle filter) [4]. The system was shown to be able to track the active contour of a specified object in a clutter environment and with a certain degree of tolerance to occlusions. However, the time complexity of this approach grows exponentially when the number of tracked objects increase, making it impractical for real-time applications.

Some researchers suggested to use training-based classifiers for the detection of desired objects, and then incorporate the detection results in tracking [8, 5, 9, 12, 13, 16]. Viola and Jones[12] proposed to perform the detection by using Adaboost classifier which is based on integral images. Okuma et al. [9] incorporated this approach into tracking with mixture particle filter [11], and demonstrated successful results in tracking inter-crossing ice hockey players. Yu and Wu [16] further extended Okuma et al.'s approach with the decentralized collaborative trackers. Their approach

Proceedings of the Sixth International Conference on Intelligent
Systems Design and Applications (ISDA'06)
0-7695-2528-8/06 $20.00 © 2006 IEEE

was shown to be able to track multiple soccer/hockey players in real-time (15 fps). However, this approach limits the number of trackers, and hence the number of tracked objects. Moreover, the performance of these classifiers greatly depends on the training sets, which are sometimes too specific and not general enough.

In this paper, we propose to combine the active contour tracking [2, 6] with decentralized trackers [16]. Unlike [16], we allow the number of trackers to be variable such that the number of tracked objects is not limited. The decentralized approach also allows the system to be more efficient to track multiple heads since the time complexity is much less dependent to the number of tracked objects. To achieve this, we first introduce a fast and robust head shape detection based on a probabilistic framework. In this framework, potential head candidates are first located by fast hough transform. The candidates are then evaluated by matching their shapes with those in the pre-defined shape space. Only those candidates with a high probability of occurrence are then tracked by decentralized trackers. Each decentralized tracker will automatically create and initialize itself, and will be destroyed when the tracked target no longer exists. We use two different shape spaces, denoted by $\boldsymbol{S_M}$ and $\boldsymbol{S_T}$, for the detection of head shape objects and active contour tracking, respectively. In general, if we do not assume any prior knowledge about the head shape when performing the detection, and allow a large $\boldsymbol{S_M}$, the detection time can be very long. On the other hand, if $\boldsymbol{S_T}$ is too small, the tracking accuracy can be affected. Therefore, a good strategy is to keep two shape spaces with $\boldsymbol{S_M} \subset \boldsymbol{S_T}$.

The details of the head shape detection and tracking are described in Section 2 and 3, respectively. We also discuss the experimental results of the proposed method in Section 4. Finally, we draw up the conclusions and discuss future works in Section 5.

2 Head Shape Detection

We use a Bayesian probabilistic approach to model the existence of a head shape. Specifically, the probability of existing a set of heads $\boldsymbol{H}$ with the corresponding location set $\boldsymbol{L}$ and a pre-defined head matching shape space $\boldsymbol{S_M}$ is formulated as:

$$P(\boldsymbol{H}|\boldsymbol{S_M}, \boldsymbol{L}) = \prod_{h_i \in \boldsymbol{H}} p(h_i|\boldsymbol{S_M}, \boldsymbol{l_i}) \prod_{h_j \in \boldsymbol{N_i}} \psi(\boldsymbol{l_i}, \boldsymbol{l_j}) \quad (1)$$

by assuming the detected $h_i \in \boldsymbol{H}$ is independent to each other. The $\boldsymbol{l_i}$ is the location of h_i. The term $\psi(\boldsymbol{l_i}, \boldsymbol{l_j})$ is a repelling function. It is defined as $1 - e^{-d(\boldsymbol{l_i}, \boldsymbol{l_j})^2/2\sigma}$ where $d(\boldsymbol{l_i}, \boldsymbol{l_j})$ is the Euclidean distance between $\boldsymbol{l_i}$ and $\boldsymbol{l_j}$, and σ is set to be half of the average head shape dimension in $\boldsymbol{S_M}$. The repelling function decreases when some detected $h_j \in \boldsymbol{N_i}$ is too close to h_i, where $\boldsymbol{N_i}$ is the set of neighborhood of h_i. With the repelling function, the system discourages the case in which multiple heads (usually heavily overlapped) are detected for a single actual head. In our actual implementation, we solve this by approximating a single head from multiple detected heads which are heavily overlapped. We group heavily overlapped h_i together, and consider the group as one head candidate with probability $\frac{1}{n} \sum_{h_i \in \boldsymbol{N_i}} p(h_i|\boldsymbol{S_M}, \overline{\boldsymbol{l_i}})$, where n is the normalizing factor and $\overline{\boldsymbol{l_i}}$ is the mean location of $h_i \in \boldsymbol{N_i}$.

By Bayes rule, the posterior probability $p(h_i|\boldsymbol{S_M}, \boldsymbol{l_i})$ is given by

$$p(h_i|\boldsymbol{S_M}, \boldsymbol{l_i}) = \frac{p(\boldsymbol{S_M}|h_i, \boldsymbol{l_i})p(h_i|\boldsymbol{l_i})p(\boldsymbol{l_i})}{p(\boldsymbol{S_M}, \boldsymbol{l_i})} \quad (2)$$

With the assumption of independence of $\boldsymbol{S_M}$ and $\boldsymbol{l_i}$, and the prior of the location $\boldsymbol{l_i}$ being uniformly distributed over the whole image, the equation can be reduced to:

$$p(h_i|\boldsymbol{S_M}, \boldsymbol{l_i}) \propto p(\boldsymbol{S_M}|h_i)p(h_i|\boldsymbol{l_i}) \quad (3)$$

A fast hough transform approach is described in Section 2.1, to locate the potential head candidate and this gives $p(h_i|\boldsymbol{l_i})$. We also introduce a shape matching algorithm in Section 2.2) to determine the likelihood, $p(\boldsymbol{S_M}|h_i)$, of the predefined shape space matching the given head.

2.1 Locating the Potential Head Candidates

Based on the observation of general head shapes in typical surveillance videos, we use an upper half circle as a shape template for locating the potential head candidates in image. Potential head shapes can then be located by applying fast hough transform [1] to the oriented sobel edges of the image. Note that the time complexity of this hough transform approach is linear. The probability $p(h|\boldsymbol{l})$ of the existence of a head given its location is then defined as:

$$p(h|\boldsymbol{l}) = \beta \frac{v_{\boldsymbol{l}}}{v_0} \quad (4)$$

where β is a scaling factor. $v_{\boldsymbol{l}}$ is the votes of hough transform at location $\boldsymbol{l}$, and v_0 is the maximum possible hough transform votes. We further apply a pre-defined threshold, τ_1, such that if $p(h|\boldsymbol{l}) < \tau_1$, no further head shape matching process described in Section 2.2 will be carried on. The value of τ_1 has been chosen from experiment, such that most of the irrelevant noises can be pruned, while the actual head candidates are retained.

2.2 Head Shape Matching

Shape matching algorithm is used to determine how likely $\boldsymbol{S_M}$ matches the potential head candidates, i.e.,

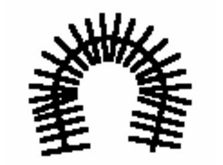


Figure 1. An example of head shape in S_M with its normals as the measurement lines

$p(\boldsymbol{S_M}|h)$. For each shape $\boldsymbol{s_j} \in \boldsymbol{S_M}$, we suggest to use measurement lines [6, 17], which are in fact the normals of the shape contour $\boldsymbol{s_j}$, to measure the likelihood $p(\boldsymbol{s_j}|\boldsymbol{h})$. $p(\boldsymbol{S_M}|h)$ is then defined as the maximum of $p(\boldsymbol{s_j}|\boldsymbol{h})$. Fig. 1 shows a particular example of a head shape in $\boldsymbol{S_M}$ with the corresponding measurement lines.

For each measurement line, it is considered to be matched if its likelihood ratio $\frac{L_{pos}}{L_{neg}}$ is larger than a threshold τ_2, in which τ_2 is calculated adaptively according to the neighboring gradient values. L_{pos} is the positive likelihood:

$$L_{pos} = \sum_{v} G(v, v_0, \sigma) g(v) \cdot n(v) \tag{5}$$

and L_{neg} is the negative likelihood:

$$L_{neg} = \sum_{v} \overline{G}(v, v_0, \sigma) g(v) \cdot n(v) \tag{6}$$

where v is a point along the measurement line. v_0 is the center of the measurement line, i.e. the intersection point of the measurement line and $\boldsymbol{s_j}$. $g(v)$ is the oriented sobel gradient at point v, and $n(v)$ is the corresponding normal vector, i.e. $g(v) \cdot n(v)$ gives the magnitude of the oriented gradient at v. $G(v, v_0, \sigma)$ is a Gaussian function centered at v_0 with variance σ. and $\overline{G}(v, v_0, \sigma) = G(v_0, v_0, \sigma) - G(v, v_0, \sigma)$ which is the inverse of the Gaussian function. We define the positive and negative likelihood in this way due to the observation that it is more likely to have a high gradient value near the center of the measurement line if the shape is matched. In contrast, when there is high gradient values farther from the center of measurement lines, it is more likely to be caused by noises due to clutter background.

Suppose s_j consists of totally M_j measurement lines, and m_j of them are matched, $p(s_j|h)$ is then defined as $\frac{m_j}{M_j}$ which is the portion of the matched measurement lines.

3 Decentralized Tracking

We use a similar decentralized tracking approach as described in [16] to perform active contour tracking [6] of each detected head. Another shape space $\boldsymbol{S_T}$, which is the superset of $\boldsymbol{S_M}$ is used for the active contour tracking. In particular, $\boldsymbol{S_M}$ consists of the most representative and general head shapes in $\boldsymbol{S_T}$, while $\boldsymbol{S_T}$ also contains deformed head shapes other than those general head shapes. This not only improves the accuracy of tracking when a tracked object undergoes shape deformation, but also retains the efficiency of the head shape detection.

As stated in previous sections, the number of trackers is not fixed. A new tracker is created and initialized for each newly detected head, and the tracker will automaticlly destroy itself when the head disappears. During tracking, a tracker evaluates itself for tracking the particular object, and only communicate with other trackers when they are getting too close. Each tracker has its own set of particles, $\{c_i, \pi_i\}_{i=1}^{n}$, where c_i is the i-th particle and π_i is its weighting. The particle set is used to estimate the probability density of current state $p(\boldsymbol{x_t}|\boldsymbol{x_{t-1}})$ from previous state. When calculating the weighting of each particle, we also incorporate the correlation rather than only use shape to improve the tracker's robustness under low contrast environment:

$$\pi_i = \lambda_1 \omega_s(\boldsymbol{x_i}) + \lambda_2 \omega_{cr}(\boldsymbol{x_i}) \tag{7}$$

where $\boldsymbol{x_i} = \{\boldsymbol{s_i}, \boldsymbol{l_i}\}$ is the state vector of c_i, with $\boldsymbol{s_i} \in \boldsymbol{S_T}$ and $\boldsymbol{l_i}$ is the particle location. $\omega_s(\boldsymbol{x_i})$ corresponds to $p(\boldsymbol{s_i}|h) = L_{pos}/L_{neg}$ which is defined in the same way as described in Section 2.2. $\omega_{cr}(\boldsymbol{x_i})$ is the typical correlation function of the head image patch. We apply λ_1 and λ_2 to weight the importance of $\omega_s(\boldsymbol{x_i})$ and $\omega_{cr}(\boldsymbol{x_i})$ respectively, such that $\lambda_1 + \lambda_2 = 1$.

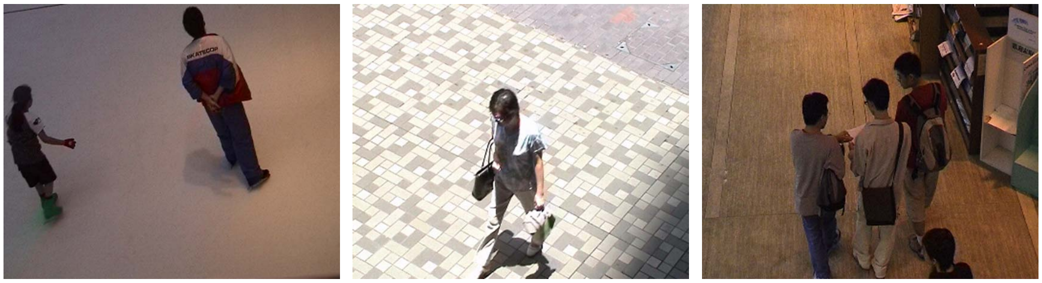
Similar to [9, 16], we also incorporate the detection results for determining the distribution of next state of the tracked head $q^*(\boldsymbol{x_t}|\boldsymbol{x_{0:t-1}})$, i.e.

$$q^*(\boldsymbol{x_t}|\boldsymbol{x_{0:t-1}}, \boldsymbol{y_t}) = \alpha q_d(\boldsymbol{x_t}|\boldsymbol{x_{t-1}}, \boldsymbol{y_t}) + (1-\alpha) p(\boldsymbol{x_t}|\boldsymbol{x_{t-1}}) \tag{8}$$

where $q_d(\boldsymbol{x_t}|\boldsymbol{x_{t-1}}, \boldsymbol{y_t})$ is the probability density function of the detection results, which is defined as $G(\boldsymbol{x_t}, \boldsymbol{y_t}, \boldsymbol{\Sigma}) p(h|\boldsymbol{S_M}, \boldsymbol{l})$. $G(\boldsymbol{x_t}, \boldsymbol{y_t}, \boldsymbol{\Sigma})$ is the Gaussian function with the covariance $\boldsymbol{\Sigma}$ and the detected state $\boldsymbol{y_t}$ as the mean. $p(h|\boldsymbol{S_M}, \boldsymbol{l})$ is the posterior probability of the head shape detection as described in Section 2. Here, α can be variated within [0,1]. When α is zero, the tracking is reduced to mixture particle filter. When α increases, the importance of the detection results also increases.

4 Experiments and Results

The proposed approach was implemented and evaluated with three challenging digital video sequences (see fig. 4), namely skating rink (3862 frames), outdoor pavement (3786 frames), and indoor passage (2299 frames). We used almost the same parameter settings for these three testing sequences except the scaling of the shapes. In particular,

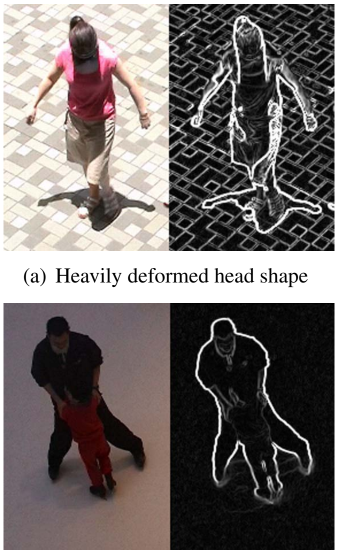


(a) Skating rink (b) Outdoor pavement with chessboard-like background and heavy sun-shine (c) Indoor passage with clutter background and crowded situations

Figure 2. Scenes of three testing sequences.

Sequences	skating rink	outdoor pavement	indoor passage
total # of people	48	26	56
# of tracked people	44	23	51
# of missed people	4	3	5
# of false positive	0	1	2

Table 1. Summary of the detection and tracking results of the proposed approach.



(a) Heavily deformed head shape

(b) Poorly contrasted image

Figure 3. Example of fail cases. The first column shows the original image, the next shows the sobel gradient magnitude map.

we allowed a newly detected head object h can initiate the creation of a new tracker only if $p(\boldsymbol{S_M}|h) > 80\%$. Another 20% was allowed to be unmatched due to the occlusions or noises. For the tracking, each tracker used at most 100 particles in order to estimate $p(\boldsymbol{x_t}|\boldsymbol{x_{t-1}})$ of each head shape object. We also set λ_1 and λ_2 in (7) to be 0.65 and 0.35 respectively. Under these settings, the system can process 15 fps on a P4 2.6GHz PC with 352x288 frame image size and around 4 to 8 objects were being tracked concurrently.

Fig. 4, 5, 6, and 7 show some screen-shots of the testing results. Each tracked target is shown by a rectangular bounding box, with its trajectory of last 50 frames in a particular color. Fig. 4 demonstrates a simple case of the tracking in the skating rink sequence. A tracker was created and initialized for a newly detected head shape object, and it successfully tracked the target which underwent sudden stop and changing of direction. The tracker was finally destroyed when the tracked target went outside the image frame. Fig. 5 shows the case of tracking inter-crossing heads. Fig. 6 shows some typical results of multiple heads tracking in the outdoor pavement sequence. The system was able to track the pedestrians which were heavily overlapped (see fig. 6(d)(e)). Fig. 7 demonstrates the system performed well in an extremely crowed and clutter environment. However, in this case, some of the head shape objects were detected and tracked with some delay. The delay is due to the fact that the system was required to gather enough confidence for the confirmation of those head objects in such a clutter background before a tracker is created. This is necessary in order to balance the detection and false alarm rates.

Table 1 summarizes the detection and tracking results. The system achieved 91.67% detection rate with 0% false alarm rate for the ice rink sequence, 88.46% detection rate with 3.85% false alarm rate for the outdoor pavement sequence, and 91.07% detection rate with 3.57% false alarm rate for the indoor passage sequence. The results show that the proposed approach performed well in the extremely crowded situations in indoor passage sequence. For the outdoor pavement sequence which contains a complex chessboard-like background and with heavy sun-shine, the system still performed fairly good with high detection rate and only one false alarm.

There are two major factors causing the fail cases. First, the appearance of head shape in the scene may not be covered in S_M. Fig. 3(a) shows an example in which the head shape was heavily deformed. This problem can be reduced by enriching the current shape space with a large head shape training sequence. Second, the contrast of head is not high enough and this results in extremely weak edge shown in fig. 3(b), the system was unable to detect the head of the child in red clothes. This case is much more challenging since it is unable to recognize the existence of the head even by human eyes without considering other body parts. In this case, a model of body part geometry [8] may help. However, this kind of complicated model can seldom perform real-time detection and tracking.

5 Conclusion and Discussion

A novel and fast head shape detection and decentralized active contour tracking approach has been proposed. We demonstrate the system based on this approach is able to detect and track multiple head shape objects in real-time. It works well even in challenging video sequences with crowded and clutter environments. Although only results of video sequence from static camera has been shown, the system itself does not depend on any static background modeling, and therefore, can be further applied to motion camera surveillance system as well (e.g. Pan-Tilt-Zoom camera).

Sometimes, some head shape like objects other than actual heads (e.g. shoulder) have been detected which causes false alarm in the current system. These cases are expected since we are focusing on auto-detection and tracking of head shape in the current work. Further verification of head objects, e.g. Adaboost detection approach [9, 16], can be incorporated in our future works to enhance the performance of the system. It should also be noted that our proposed approach effectively reduces the search space of such computationally intensive detection, and hence improve the system speed and performance.

References

[1] D. Ballard and C. Brown. *Computer Vision*. Prentice Hall, 1982.

[2] A. Blake and M. Isard. *Active Contours*. Springer, 1998.

[3] I. Haritaoglu, D. Harwood, and L. Davis. W4 : Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–819, August 2000.

[4] M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[5] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proc. 2005 IEEE Conference on Computer Vision and Pattern Recognition*, volume I, pages 878–885, San Diego, CA, USA, June 2005.

[6] J. MacCormick. *Stochastic Algorithms for Visual Tracking : probabilistic modelling and stochastic algorithms for visual localisation and tracking*. London ; New York : Springer, 2002.

[7] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. *International Journal of Computer Vision*, 39(1):57–71, 2000.

[8] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proc. 8th European Conference on Computer Vision*, volume I, pages 69–82, Prague, Czech Republic, May 2004. Springer–Verlag.

[9] K. Okuma, A. Taleghani, N. d. Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *Proc. of European Conf. on Computer Vision*, volume I, pages 28–39, Prague, May 2004.

[10] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative density propagation for 3d human motion estimation. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, volume I, pages 390–397, San Diego, CA, USA, June 2005.

[11] J. Vermaak, A. Doucet, and P. Pérez. Maintaining multimodality through mixture tracking. In *Proc. 9th IEEE International Conference on Computer Vision*, volume 2, pages 1110–1116, Nice, France, October 2003.

[12] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. 2001 IEEE Conference on Computer Vision and Pattern Recognition*, volume I, pages 511–518, December 2001.

[13] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, June 2005.

[14] S.-F. Wong and K.-Y. K. Wong. Real time human body tracking using wavenet. In *Proc. Asian Conference on Computer Vision*, pages 91–96, Jeju Island, Korea, January 2004.

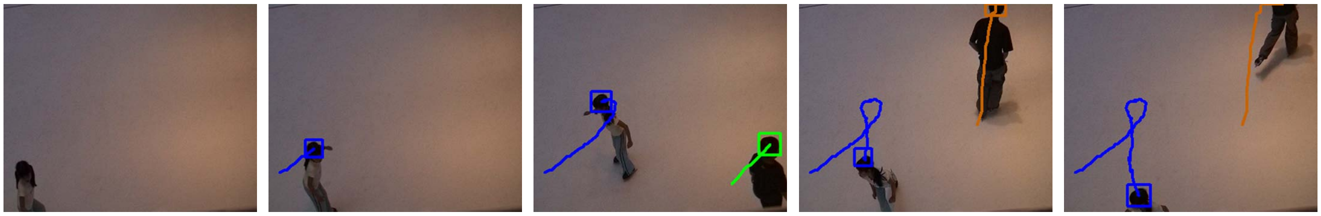
[15] B.-S. Yiu, K.-Y. K. Wong, F. Chin, and R. Chung. Explicit contour model for vehicle tracking with automatic hypothesis validation. In *Proc. International Conference on Image Processing*, volume II, pages 582–585, Genova, September 2005.

[16] T. Yu and Y. Wu. Decentralized multiple target tracking using netted collaborative autonomous trackers. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, volume I, pages 939–946, San Diego, CA, USA, June 2005.

[17] T. Zhao and R. Nevatia. Tracking multiple human in crowed environment. In *Proc. 2004 IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 406–413, Washington, D.C., USA, June 2004.

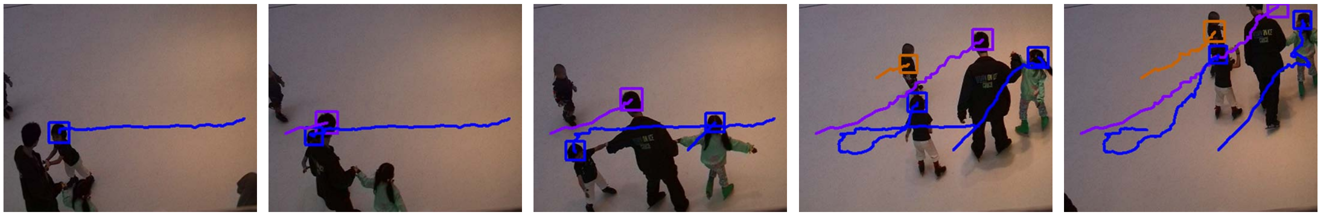
Acknowledgment

This research was jointly sonsored by Multivision Intelligence Surveillance Limited and the Innovation and Technology Commission of the Government of the Hong Kong Special Administrative Region, under the Grant UIM/167.



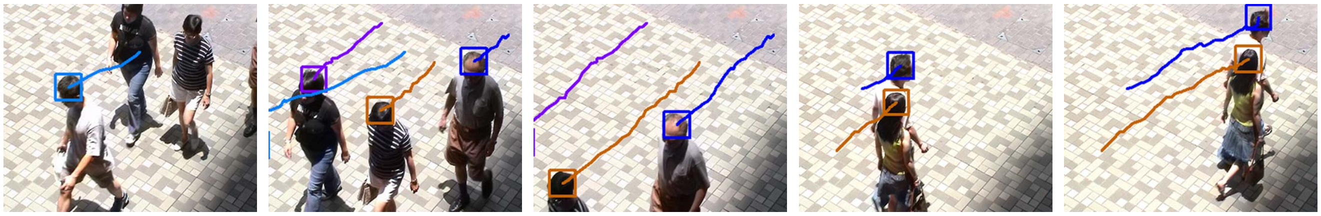
(a) frame 151 (b) frame 159 (c) frame 192 (d) frame 222 (e) frame 232

Figure 4. Typical tracking results of the ice rink sequence which shows (a)(b) creation and initiation of a tracker, (c)(d) continue tracking with sudden stop and changing of direction, (e) and destruction of the tracker.



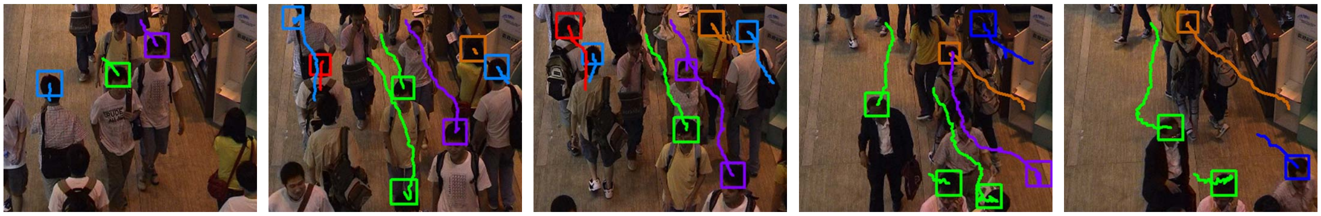
(a) frame 1663 (b) frame 1669 (c) frame 1682 (d) frame 1720 (e) frame 1750

Figure 5. Successful tracking of inter-crossing heads in ice rink sequence.



(a) frame 473 (b) frame 499 (c) frame 518 (d) frame 3309 (e) frame 3332

Figure 6. Typical tracking results of outdoor pavement sequence. (a)(b)(c)show that the system is able to track multiple head shape objects concurrently with a chessboard-like background. (d)(e) show the tracking result of two heavily overlapped pedestrian.



(a) frame 488 (b) frame 529 (c) frame 549 (d) frame 628 (e) frame 647

Figure 7. Typical tracking results of indoor passage sequence. The system performs well even in an extremely crowded and clutter environment, but with some cases being detected and tracked with small delay.