Multiple Objects Tracking across Multiple Non-overlapped Views

Ke-Yin Chen[1], Chung-Lin Huang[1], Shih-Chung Hsu[1], and I-Cheng Chang[2]

[1] Department of Electrical Engineering, National Tsing Hua University, Hsin-Chu, Taiwan
g9761587@oz.nthu.edu.tw, clhuang@ee.nthu.edu.tw, chvjohnff@gmail.com
[2] Department of Information Science and Engineering, National Don-Hwa University, Ha-Lien, Taiwan
ICChang@mail.ndhu.edu.tw

Abstract. This paper introduces a tracking algorithm to track the multiple objects across multiple non-overlapped views. First, we track every single object in each single view and record its activity as the object-based video fragments (OVFs). By linking the related OVFs across different cameras, we may connect two OVFs across two non-overlapped views. Because of scene illumination change, blind region lingering, and objects similar appearance, we may have the problem of path misconnection and fragmentation. This paper develops the Error Path Detection Function (EPDF) and uses the augmented feature (AF) to solve those two problems.

Keywords: Object tracking, Object-based Video Fragment (OVF), Augmented feature (AF), Error Path Detection Function (EPDF).

1 Introduction

Video surveillance system is constructed by a network of cameras with multiple non-overlapped views. In each camera, a period of video of each object's activity is recorded in a so-called object-based video fragment (*OVF*). This paper introduces a method to connect two *OVF*s of the same object moving across two non-overlapped views. Because of scene illumination change, blind region lingering, and objects similar appearance, the system faces the problems of *OVF* misconnection and fragmentation. Our method can detect and correct the miss-connected *OVF*s, and then reconnect the *OVF*s of the same object moving across cameras.

Lee *et al.* [2] proposed an approach for tracking objects in the cameras with overlapped field of views (FOVs) without calibration. Khan *et al.* [3] used *FOV* line constraints for tracking objects in overlapped cameras. Multi-camera tracking approaches with overlapped *FOV*s have been proposed [4, 5]. In non-overlapped views, Kettnaker *et al.* [6] presented a Bayesian solution to track objects across multiple cameras with non-overlapped views. Porikli *et al.* [7] combined spatiotemporal and appearance cues to track objects and solve the inter-camera color calibration problem.

Y.-S. Ho (Ed.): PSIVT 2011, Part II, LNCS 7088, pp. 128–140, 2011.

Black *et al* [1] used the HSI color space to improve illumination invariance. Javed *et al*. [8] present a camera network topology learning method using the path probabilities of objects. Individual tracks are found by searching the maximal posterior probability of the spatiotemporal and color appearance. Javed *et al*. [9, 10] developed the subspace of inter-camera brightness transfer functions to solve the problem of appearance change across the scenes. D'Orazio *et al*. [11] compared different methods to evaluate the color Brightness Transfer Function (*BTF*) between non overlapped cameras.

Chen *et al*. [12] proposed an unsupervised method to learn both spatiotemporal and appearance relationships for long-term monitoring. They consider the environment changes, such as sudden lighting change. Dick *et al*. [13] used a stochastic transition matrix to describe the observed pattern of people motion within and between FOVs. Ellis *et al.* [14] developed an automatic labeling method to construct the network topology. Stauffer *et al*. [15] built a correspondence model for cameras with both overlapped and non-overlapped *FOV*s.

Mehmood *et al*. [17] combined the optical flow, feature matching and shape descriptors to detect and track objects efficiently. Their method can be applied to multiple non-overlapped cameras to attain correct inter-camera correspondences. Piccardi *et al*. [18] used the Major Color Spectrum Histogram representation (MCSHR) to represent a moving object. Based on k-means clustering, the reduced color space is used to tolerate the minor changes in color between different cameras and lighting. Song *et al*. [19] combined short term feature correspondence with long-term feature dependency models to derive a path smoothness function for error correspondence correction.

This paper presents a multiple objects tracking across multiple non-overlapped views by using spatio-temporal cues and appearance cues in different views. Our system consists of (1) applying the foreground extraction method to segment the foreground object, (2) using the spatiotemporal cues and appearance connect the related OVFs across different views, (3) using Augmented Feature (*AF*) propagation method to solve the fragmentation and miss-connection problems. Different from [19], our major contributions are proposing the *Error Path Detection Function* (*EPDF*) to find the miss-connection, and using the *AF* to re-connect the *OVFs*.

2 Problem Formulation

Our problem is formulated as multiple-object tracking in non-overlapped multiple views. The camera network can be described by a graph of which each node represents the scene of a certain view. As shown in Figure 1, there are six non-overlapped views. Each scene (node) may have more than one *zone*, and each zone can be either an entrance or an exit of the scene. In Figure 1, we find four zones in node 2, and only one zone in nodes 1 and 3. Every two zones are either direct or non-direct related. Two zones in the same node or two neighboring nodes are direct-related, otherwise they are non-direct related. If two direct-related zones are in the same node, they have *intra-zone* relationship, otherwise, they have *inter-zone* relationship.

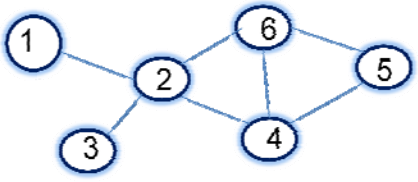

Fig. 1. The camera network topology

Any object moving between two intra-zones in the same view can be tracked and recorded as an object-based video fragment (*OVF*). The object movements between two inter-zones in different views are unknown but predictable. Our goal is to link the related *OVF*s across *inter-zones* by finding the objects in two neighboring views with the similar spatiotemporal cue and appearance cue. The linkage between two *OVF*s is called a *joint*. Here, we assume that (1) the system can track any object within one single view, (2) the cameras are synchronous, (3) each *OVF* is marked by time stamp, object appearance and location information, and (4) the zones in each scene are known priori.

The region (or linkage) between two inter-zones is a *closed blind region*. Figure 1 shows a closed blind region between nodes 1 and 2. Object leaving the exit zone of node 1 will enter the close blind region, and re-appear in the enter zone of node 2 sooner or later. We also define another blind region, called "*open blind region*" in which the object may not necessarily re-appear in any other node. For certain node adjacent to the open blind region, it has no inter-zone relationship with any other node. Object may enter or leave the scenes through the open blind region. Figure 1 also shows an open blind region in node1 or 3.

2.1 Object Tracking in Single View

First, we apply background subtracting and shadow removal to extract the foreground object when it enters the enter zone. Based on the extracted foreground object, the object model can be obtained which can be used for object tracking. In the non-overlapped scenes, each moving object appears in only one single view at any time instance. Here, we apply HS (Hue-Saturation) color histogram to model the object, and then use Mean-Shift algorithm [20] to track the moving object which is enclosed by a rectangle as shown in Figure 2. The rectangle is represented as $s = \{x, y, h, r\}$, where (x, y) represents the center of the rectangle, and (h, r) represents the height and the aspect ratio.

Fig. 2. The two video fragments of the same object

2.2 Spatiotemporal and Appearance Cues

The appearance cue of each object is modeled by the HS (Hue-Saturation) color histogram of the rectangle enclosing the moving object. The similarity measure between the observations of two objects is described by computing Bhattacharyya coefficient ρ based on the color histograms, $p(u)=\{p^{(u)}\}_{u=1\ldots m}$ and $q(u)=\{q^{(u)}\}_{u=1\ldots m}$ of two objects. Larger ρ indicates more similar between these two color histograms. The similarity distance between two objects is measured by $d = \sqrt{1-\rho[p(u),q(u)]}$. The color distribution of each object is temporally updated.

By exploiting the camera network topology, we can describe the spatiotemporal relationship between the cameras in terms of the *transition time* and the *transition probability*. The former indicates the time duration for an object moving from one exit zone to the other entry zone, and latter is the probability distribution of the transition time between two observations in two inter-zones. The spatiotemporal relationship between two inter-zones is based on the camera network topology. After the training phase, we have the transition probability for each possible linkage between two inter-zones. For inter-zones a and b of two different views, we use $P_{ab}(T)$ to describe the transition probability that people move from zone a to zone b after time T. The same object exits from zone a at time T_i and enters zone b at time T_j, then $T=T_j-T_i$.

3 Inter-zone Video Fragments Linkage

Here, we use the spatiotemporal and appearance cues of the observations to generate a preliminary linkage of *OVF*s across inter-zone. Based on the observations of *OVF*s across inter-zones, we may create the linkage of the two *OVF*s. For each zone, there is a handover list. The handover list of zone a (*i.e.*, H_a) is defined as the collection of the observations of the objects appearing in the adjacent zones of zone a.

Object A enters the zone Z_A with the observation denoted as $\boldsymbol{O}_A$ which consists of the spatiotemporal cue $O_A(st)$ and the appearance cue $O_A(app)$. The $O_A(st)$ includes the camera *id*, the zone *id*, the position, and time of appearance at the zone A as $T(O_{ZA})$. Then we find the best corresponding object with the observation $\boldsymbol{O}_h$ in the handover list H_A. Based on $O_A(st)$ and $O_A(app)$, we find the best matched one in H_A. If the highest probability exceeds a threshold, we label the new observation $\boldsymbol{O}_A$ and the observation $\boldsymbol{O}_h$ as the same object. Otherwise, object A is treated as a new object entering in the scene. The similarity between the observation $\boldsymbol{O}_A$ and the related one $\boldsymbol{O}_h$ in the handover list H_A (*i.e.*, $\boldsymbol{O}_h \in H_A$) is described as $p(O_A, O_h)$. The most likely one in H_a can be obtained as

$$\varphi = Arg\max_h \, p(O_A, O_h) \tag{1}$$

Assuming $O_A(st)$ and $O_A(app)$ are independent so that we can compute likelihood of similarity based on $O_A(st)$ and $O_A(app)$ with different weights. Equation (1) can be rewritten as

$$\varphi = Arg\max_h p(O_A, O_h)$$
$$= Arg\max_h \left[w \cdot p(O_A(st), O_h(st)) + (1-w)\, p(O_A(app), O_h(app)) \right]$$

where $p(O_A(app), O_h(app))$ is the probability of appearance similarity, and $p(O_A(st), O_h(st))$ is the probability of spatiotemporal similarity defined as

$$p(O_A(st), O_h(st)) = \sum_{\forall Z_A} \sum_{\forall Z_h} P_{Z_A Z_h}(T) \left[p(O_A(st) | Z_A)\, p(O_h(st) | Z_h) \right] \quad (2)$$

where $P_{ZAZh}(T)$ is the transition probability of travel time between two inter-zones Z_A and Z_h as $T = T[\boldsymbol{O}_{ZA}] - T[\boldsymbol{O}_{Zh}]$, and $P(\boldsymbol{O}^* | Z^*)$ is the probability of the observation $\boldsymbol{O}^*$ entering or exiting from zone Z^*. To connect two *OVF*s, we dynamically adjust the weighting for spatiotemporal features and appearance features. If the illumination changes make the appearance features unreliable for similarity measure, we will increase the weight of spatiotemporal features.

Connecting *OVF*s can be viewed as a labeling problem. Two *OVF*s assigned with the same label will be linked together indicating the activity of the same individual object. The initially cascaded OVFs are called the *OVF link*. In Figure 3, we show the ground truth of two *OVF* links for two moving objects. However, we may have six initial *OVF links*. The path fragmentation problem occurs when the connection between OVFs fails because of (1) two or more people with the same appearance, and (2) the lingering time of the designated object is longer than the others.

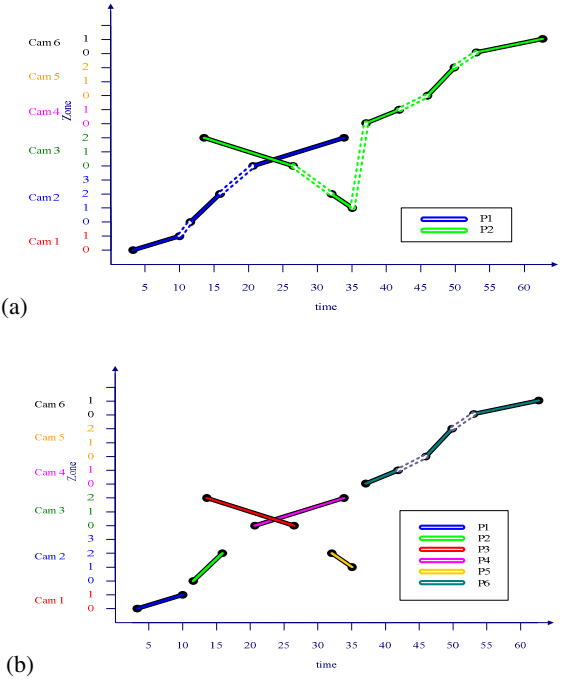

Fig. 3. (a) Ground truth, (b) six linked paths

The initially *OVF links* may have two problems: *path fragmentation* and *path misconnection.* The ***path fragmentation*** indicates that the *OVF*s of the same object are not connected across the inter-zones. It occurs due to *variant lighting* and *uncertain lingering time* in blind region. The ***path misconnection*** indicates that the *OVF*s of different objects may be miss-connected together. It occurs when people dressed in similar clothes may appear from the blind region at the same time, and appearance cue becomes not reliable. Therefore, we need to find and correct the miss-connection and solve the fragmentation.

4 Error Linkage Correction

Because of the different viewing angles and positions of the cameras, the observations from different cameras are not the same. In addition to the spatiotemporal and appearance cues, we may have another feature, the human face. The human face feature can be detected and treated as the augmented feature (*AF*) for correcting the path miss-connection. The correction process consists of four steps: (1) Calculate the error path detection function (*EPDF*) at the joints to check the validity of the linked *OVF*s, (2) Divide the *OVF* link is divided into two *OVF* sub-links at the joint if there is an error, (3) Propagate the *AF* in the same *OVF* sub-link, and (4) Re-calculate the similarity between the *OVF* sub-links for path correction.

4.1 Misconnection Detection

Path misconnection usually occurs when several objects with similar appearance pass through closed blind region at the same time. Here we propose the Error Path Detection Function (*EPDF*) to represent the possible misconnection. Two OVFs has been connected at joint and assigned to the same link L_z as the $i\text{-}1^{th}$ and i^{th} fragment as $\boldsymbol{O}_{z,i-1}$ and $\boldsymbol{O}_{z,i}$. We compare the possible connection of $\boldsymbol{O}_{z,i-1}$ and $\boldsymbol{O}_b$ with the proposed one, in which $\boldsymbol{Q}_b$ in the handover list of $\boldsymbol{Q}_{z,i-1}$ as $Q_b \in H(Q_{z,i-1})$. If the difference is not large enough, then the connection may not be correct. We use *EPDF* to identify the reliability of the connection (or *joint*) as

$$EPDF(L_{z,i}) = \begin{matrix} 0 & if \left| P(\boldsymbol{O}_{z,i-1},\boldsymbol{O}_{z,i}) - P(\boldsymbol{O}_{z,i-1}, \boldsymbol{O}_b) \right| > Thres \\ 1 & Otherwise \end{matrix}$$

where $L_{z,i}$ is defined as the ith joint of link z, i=1,…. N_z−1, and N_z represents the number of fragments in the link.

Figure 4 shows an example of three objects of which Object 1 and Object 3 have similar color appearance. Object 3 leaves zone 0 of camera 3 first and then Object 1 leaves later. We compute *EPDF* at the joint of every *OVF* link to find the error connection. Figure 5 shows the *EPDF* of three OVF links respectively, of which two *OVF* links have misconnection problem. The 1^{st} *OVF* link indicates that there are another similar appearance objects in the blind region simultaneously, so that the difference value is less than threshold. We calculate *EPDF* of every *OVF* link to find the misconnected joint.

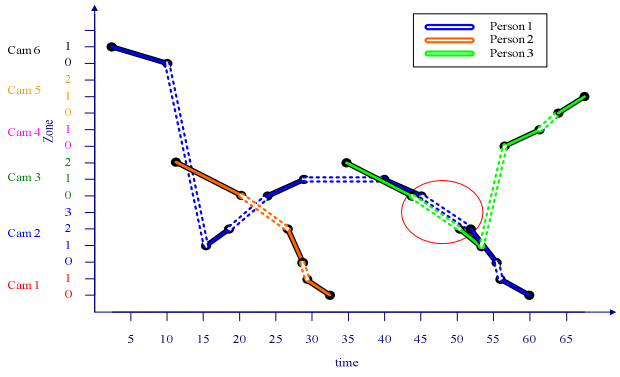

Fig. 4. The initial *OVF* links with misconnection

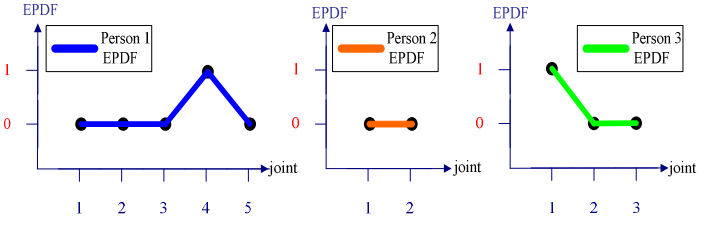

Fig. 5. The EPDF of three links

4.2 Augmented Feature Propagation

The captured observation consists of spatiotemporal and appearance cues. For some observations which cannot be obtained in each view are called the Augmented Features (*AF*). Since the camera viewing directions are different, the *AF* may not be found in every connected *OVF*. Since the connected *OVF*s are supposed to have the same *AF*, we may propagate *AF* across the connected *OVF*s.

We propagate the *AF*s to all *OVF*s in the same *OVF* link. The *AF*s of *OVF* links are used to calculate their similarity. We compare every two *OVF* sub-links, and then connect the two OVF sub-links with the highest similarity. The path misconnection and fragmentation problem can be solved by the following steps:

(1) For each *OVF* link L_z calculate *EPDF* for each joint *i*.
(2) Segment link L_z into a *OVF* sub-links S_x and S_y.
(3) Propagate the additional *AF*s to the other fragments of the same *OVF* sub-link.
(4) Establish the correspondence between the observations of every two *OVF* sub-links.
(5) If there is only one *OVF* sub-link in handover list, they can be connected directly.
(6) For each cascaded *OVF* link, re-compute the *EPDF* at every joint.
(7) Repeat the above steps until *EPDF* of this path is zero, or else it fails.

The correspondence between two *OVF* sub-links S_{ax} and S_{by} can be obtained based on the observations Q_{ax} and Q_{by}. The likelihood of the two observations is described as $p(Q_{ax}, Q_{by})$ with S_{by} in the handover list of S_{ax} as $S_{by} \in H(S_{ax})$. Assume that the spatiotemporal cue, the appearance cue and the augmented cue are independent. The most likely corresponding *OVF* sub-links can be described as

$$\varphi = Arg\max_{by}\, p(Q_{a,x}(aug), Q_{b,y}(aug))$$

where $p(Q_{a,x}(aug), Q_{b,y}(aug))$ is likelihood of the two observations based on the *AF* similarity.

Figure 6 shows the results of *AF* Propagation. The *OVF* link L_1 is divided into two *OVF* sub-links S_{11} and S_{12}, and *OVF* link L_2 is also divided into two *OVF* sub-links S_{21} and S_{22}. The *AF* is propagated in every *OVF* sub-link. In *OVF* sub-link S_{22}, the *AF* of *OVF* #2 is propagated from *OVF* #4. Therefore, S_{22}, S_{11}, and S_{21} have the same similar *AF*s. Based on the propagated *AF*s, we may compute the similarity between two *OVF* sub-links. Each *OVF* sub-link will be connected to another *OVF* sub-link with the largest similarity.

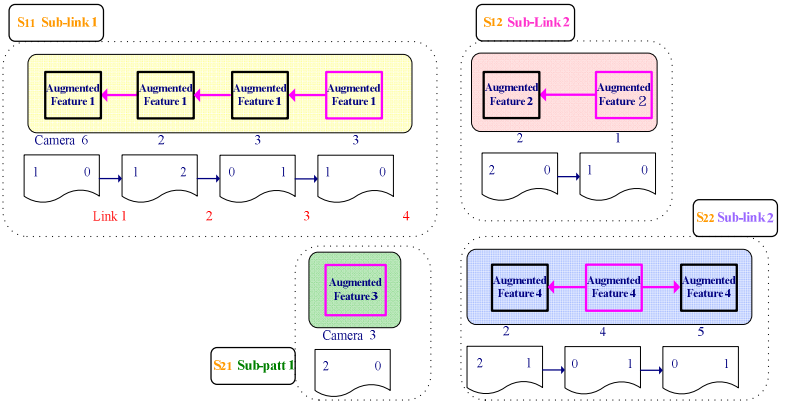

Fig. 6. The Propagation of *AF*s in *OVF* sub-links

Since the similarity between S_{22} and S_{11} is much larger, S_{22} is cascaded with S_{11}, as a new link which will retain *AF* #1 and *AF* #4 simultaneously. Once the connection is determined, the *EPDF* of 4th joint of the new link of will become zero.

There is only one *OVF* sub-link S_{21} in handover list, so that it is connected with S_{12} and become a new *OVF* link. The *EPDF* of 1st joint will be zero. Due to similar color appearance of different objects, the wrong linkages and cascaded *OVF* links are generated. As shown in Figure 7, *OVF* link 2 has two incorrect linkages, which are indicated by a pink circle.

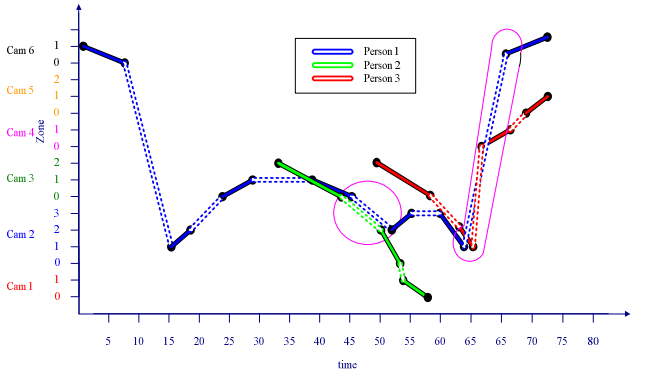

Fig. 7. Initial miss-connected linkages

Figure 8 shows that an *AF* (e.g., human faces features) is propagated to other *OVF*s in the same sub-link. Solid lines represent the initial links, and the dashed lines indicate the ground truth. Sub-link S_{22} is miss-connected with S_{21}. Sub-links S_{32} and S_{12} are not linked because of no *AF* propagation. S_{22} is more coherent with S_{11} than with S_{21}, $p(Q_{22}, Q_{11})>p(Q_{22}, Q_{21})$. Therefore, S_{22} is connected with S_{11}, and the *EPDF* is set as zero. There is only one sub-link S_{21} in S_{12} handover list, so they are connected.

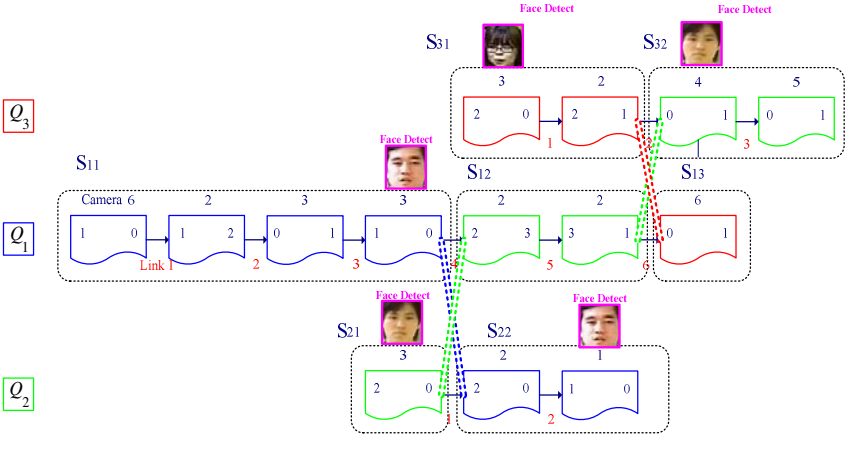

Fig. 8. The reconnection between two *OVF* sub-links

5 Experimental Results

In the experiments, we have the synchronized videos from six non-overlapped cameras. The format of image frame is 320×240×24 bits and the frame rate is 25 frames/sec. Figure 9 shows six indoor non-overlapped views in the experiment. The color histogram of the object is used as the basic feature for object tracking. Each tracked object in each view is enclosed by a rectangle block. The parameters of each block are the position, the height, and the aspect ratio of the block.

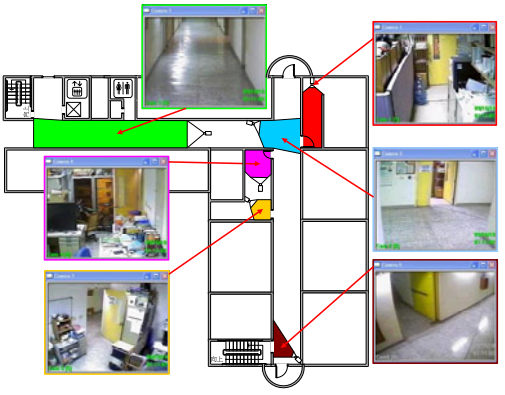

Fig. 9. The non-overlapped multi-cameras system

To illustrate the effectiveness of our system, we demonstrate three experiments.

a) Experiment 1. Three people enter in the viewing of camera 6 individually. They walk together in the blind region at the same time, and then leave the blind region individually. Each object can be tracked independently after the walking through the blind region.

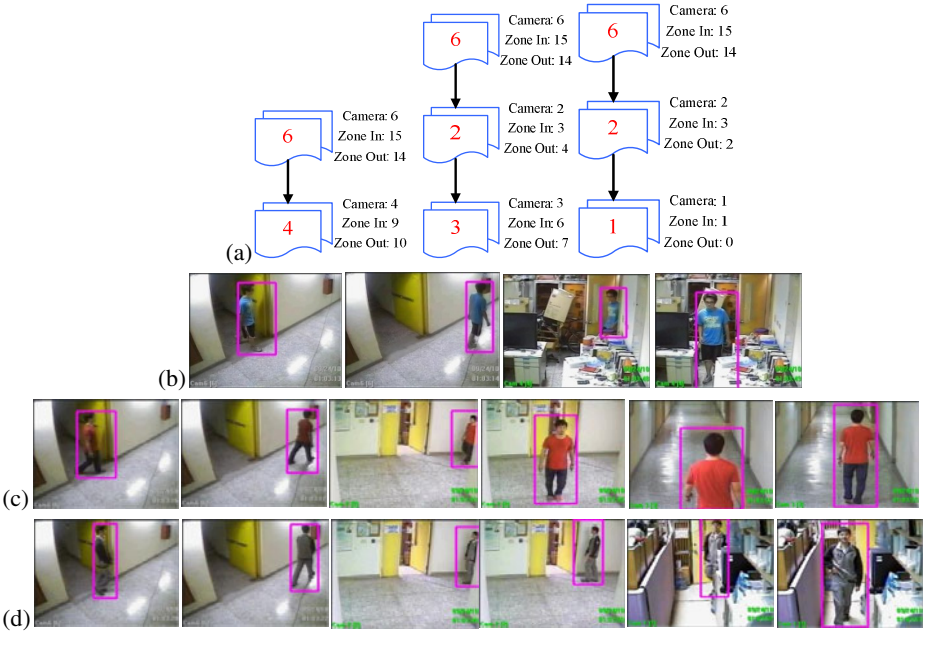

Fig. 10. The experimental results of experiment #1. (a) three different paths, (b)~(d) the frames of each OVF links.

b) Experiment 2. Two people enter the scene of camera3 individually. Object 2 leaves camera 3 first, and object 1 leaves later. But object 1 enters the scene of camera 2 first, object 3 enters later. Their spatiotemporal similarity is close, and they have similar appearances. In the initial *OVF* linkage, path miss-connection occurs when two objects leave the blind region. The *OVF* link of object 2 in camera 2 and 1 will be connected to the *OVF* link of object 1 in camera 3. The *OVF* links of object 1 in cameras 2, 4 and 5 will be connected to the *OVF* link of object 2 in camera 3.

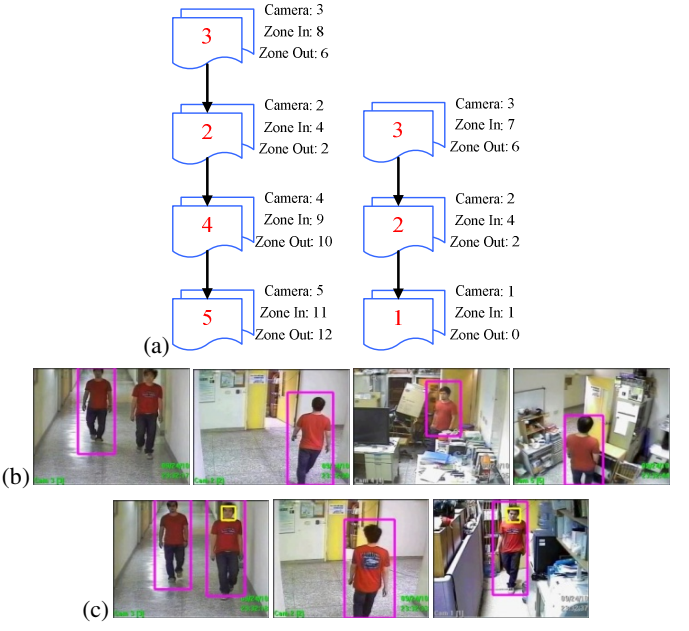

Fig. 11. The Experimental results of Experiment #2. (a) two different paths, (b)~(c) the frames of each OVF links.

c) Experiment 3. Three people with similar appearances (dressed in the same colors) appear in the scene. When they enter the same closed blind region, the path miss-connection problem occurs. *OVF* link #2 is a miss-connection. We employ the *EPDF* to find the miss-connected joints and divide the miss-connected *OVF* links into two *OVF* sub-links. By applying *AF* propagation, we can reconnect the *OVF* sub-links as *OVF* link.

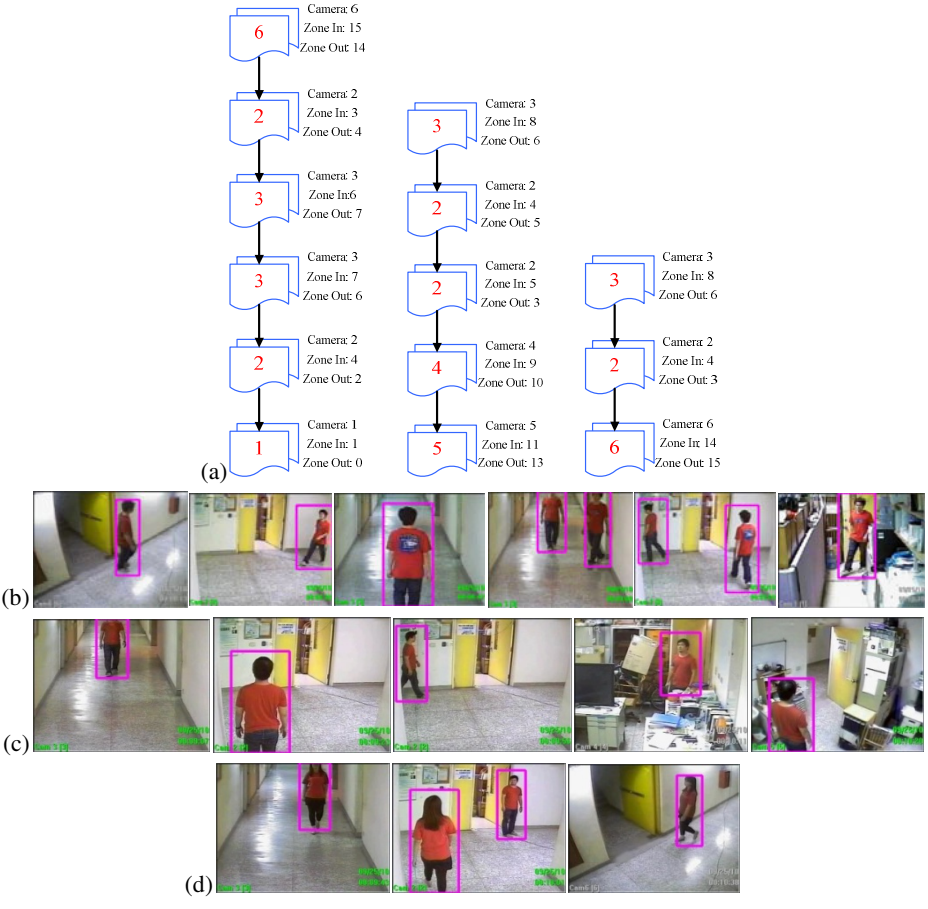

Fig. 12. The Experimental results of Experiment #3. (a) three different paths, (b)~(d) the frames of each OVF links.

6 Conclusions

The paper presents a tracking system for multiple cameras with non-overlapped views by exploiting the basic features, spatiotemporal features, and appearance features to determine human's tracks across cameras. We have shown that our method can detect and correct the miss-connected *OVF*s, and then reconnect the *OVF*s of the same object moving across cameras.

References

[1] Black, J., et al.: Wide Area Surveillance with a Multi-Camera Network. In: Proc. of Intelligent Distributed Surveillance Systems (2003)

[2] Lee, L., et al.: Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frame. IEEE Trans. PAMI 22(8), 758–768 (2000)

[3] Khan, S., et al.: Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapped Fields of View. IEEE Trans. PAMI (2003)
[4] Javed, O., et al.: KNIGHTM: a real time surveillance system for multiple overlapped and non-overlapped cameras. In: ICME (2003)
[5] Zhu, L.-J., et al.: Tracking of multiple objects across multiple cameras with overlapped and non-overlapped views. In: IEEE ISCAS (2009)
[6] Kettnaker, V., et al.: Bayesian Multi-camera Surveillance. In: IEEE CVPR (1999)
[7] Porikli, F., et al.: Multi-Camera Calibration, Object Tracking and Query Generation. In: IEEE ICME (2003)
[8] Javed, O., et al.: Tracking across Multiple Cameras with Disjoint Views. In: 9th IEEE ICCV (October 2003)
[9] Javed, O., et al.: Appearance modeling for tracking in multiple non-overlapped cameras. In: IEEE CVPR 2005, vol. 2, pp. 26–33 (June 2005)
[10] Javed, O., et al.: Modeling inter-camera space–time and appearance relationships for tracking across non-overlapped views. Computer Vision and Image Understanding, 146–162 (2008)
[11] D'Orazio, T., et al.: Color Brightness Transfer Function Evaluation for Non overlapped Multi Camera Tracking. In: ICDSC (2009)
[12] Chen, K.W., et al.: An Adaptive Learning Method for Target Tracking across Multiple Cameras. In: IEEE CVPR 2008, pp. 1–8 (June 2008)
[13] Dick, A., et al.: A Stochastic Approach to Tracking Objects across Multiple Cameras. In: Australian Conf. on Artificial Intelligence, pp.160–170 (2004)
[14] Ellis, T.J., et al.: Learning a Multi-Camera Topology. In: IEEE Workshop on VS-PETS (2003)
[15] Stauffer, C.: Learning to Track Objects through Unobserved Regions. In: IEEE Workshop on Motion and Video Computing, pp. 96–102 (January 2005)
[16] Mehmood, M.O.: Multi-camera based Human Tracking with Non- Overlapped Fields of View. In: Int. Conf. on AICT 2009, pp. 1–6 (October 2009)
[17] Cheng, E.D., et al.: Mitigating the Effects of Variable Illumination for Tracking across Disjoint Camera Views. In: IEEE AVSS 2006 (November 2006)
[18] Piccardi, M., et al.: Track matching over disjoint camera views based on an incremental major color spectrum histogram. In: IEEE AVSS (2005)
[19] Song, B., et al.: Robust Tracking in A Camera Network A Multi-Objective Optimization Framework. IEEE J. on Selected Topics in Signal Processing, 582–596 (2008)
[20] Comaniciu, D., Meer, P.: Mean-Shift: A robust Approach toward feature space analysis. IEEE Trans. on PAMI 24(5), 603–619 (2002)