

# Human tracking in camera network with non-overlapping FOVs

Lin Guoyu      Yang Biao      Zhang Weigong

(School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China)

**Abstract:** An adaptive human tracking method across spatially separated surveillance cameras with non-overlapping fields of views (FOVs) is proposed. The method relies on the two cues of the human appearance model and spatio-temporal information between cameras. For the human appearance model, an HSV color histogram is extracted from different human body parts (head, torso, and legs), then a weighted algorithm is used to compute the similarity distance of two people. Finally, a similarity sorting algorithm with two thresholds is exploited to find the correspondence. The spatio-temporal information is established in the learning phase and is updated incrementally according to the latest correspondence. The experimental results prove that the proposed human tracking method is effective without requiring camera calibration and it becomes more accurate over time as new observations are accumulated.

**Key words:** multiple camera tracking; non-overlapping FOVs; spatio-temporal information; human appearance model; incremental learning

**doi:** 10.3969/j.issn.1003-7985.2012.02.005

Surveillance cameras are increasingly being used as a tool to monitor and detect crime. As a result, there are large numbers of cameras which lack effective continuous monitoring due to the limitations of humans in managing large scale systems. Therefore, tools to assist and aid the surveillance operator's decisions are very essential and necessary. The goal of surveillance cameras is to locate targets, track their trajectories, and maintain their identities when they travel within or across cameras. Such a surveillance system consists of two parts: 1) Intra-camera tracking, i. e., to track targets within a camera; 2) Inter-camera tracking, i. e., to solve the "target handover" problem of tracked targets across cameras. Much research has been done on intra-camera tracking, and now the problem is solved very well. However, inter-camera tracking is more challenging than intra-camera tracking, because the appearance of a target in different cameras may be not con-

sistent due to various factors, such as camera characteristics, lighting conditions, and view angles. Besides, the open blind areas significantly increase the complexity of the inter-camera tracking association problem<sup>[1]</sup>.

Much work has been done<sup>[2-4]</sup> on multi-camera tracking with overlapping FOVs. These methods usually require camera calibration and environmental models to obtain the homographic relationship between these cameras. In fact, the benefit of calibrated cameras or site models is unavailable in most real situations. Maintaining calibration among a large camera network is a discouraging task, and it is very difficult to recalibrate every camera when slight changes in its position occur. However, in most surveillance situations this is unrealistic. Refs. [5-7] represent some early work on multi-camera tracking with non-overlapping FOVs. These references proposed that to solve the handover problem, it is necessary to establish the correspondence between objects in different cameras, and for which the spatio-temporal information and appearance information are two important cues.

For the spatio-temporal cue, various methods are proposed in Refs. [8-11]. Javed<sup>[8]</sup> proposed a general method to learn the camera topology and path probabilities of objects using Parzen windows. This is a supervised learning technique where the spatio-temporal information is learnt during the learning phase using a small number of manually labeled trajectories. Dick and Brooks<sup>[9]</sup> used a stochastic transition matrix to describe people's observed patterns of motion both within and between FOVs. The method required an online learning phase where a marker was carried around the environment. Makris et al.<sup>[10]</sup> investigated the unsupervised learning of recovering the network topology to facilitate tracking between spatially adjacent cameras using the cross-correlation method without hand-labeled correspondence. Wang et al.<sup>[11]</sup> learned the spatio-temporal cues by modeling objects' trajectories according to the activities information and considered that if the trajectories belonged to the same activity then they were likely to correspond to the same object. Besides, there was some work addressing the optimization framework of multiple targets correspondence. Kettner and Zabih<sup>[7]</sup> used a Bayesian formulation to reconstruct the paths of targets across multiple cameras. Javed<sup>[8]</sup> dealt with this problem by maximizing a posteriori probability using a graph-theoretic framework. Song and Roy-Chowdhury<sup>[12]</sup> proposed a multi-objective optimization framework by combining short-term feature correspondences across the cameras with long-term feature depend-

**Received** 2012-01-18.

**Biography:** Lin Guoyu (1979—), male, doctor, lecturer, Andrew\_Lin@seu.edu.cn.

**Foundation items:** The National Natural Science Foundation of China (No. 60972001), the Science and Technology Plan of Suzhou City (No. SG201076).

**Citation:** Lin Guoyu, Yang Biao, Zhang Weigong. Human tracking in camera network with non-overlapping FOVs[J]. Journal of Southeast University (English Edition), 2012, 28(2): 156 – 163. [doi: 10.3969/j.issn.1003-7985.2012.02.005]

ency models. The optimization framework mentioned above is to find a set of correspondences where the observations of different cameras correspond to the consecutive tracks of the same objects in the environment. But the fact is that some targets may disappear from one view field and reappear in another sometime later; some targets may disappear from the surveillance network forever, and some new targets which never appeared before may enter the surveillance network. So the spatio-temporal information obtained from the one-to-one correspondence is not quite right. And if the environment changes or some cameras move, the learning procedure will be rebooted to obtain the solution.

For the appearance cues, Porikli<sup>[13]</sup> derived a non-parametric function to model color distortion for pair-wise camera combinations using correlation matrix analysis and dynamic programming. Madden et al.<sup>[14]</sup> introduced an incremental major color spectrum histogram representation (IMCSHR) over a short window of successive frames to describe the object's main colors and compensate for small, short-term changes in the object's pose. Lian et al.<sup>[15]</sup> proposed a competitive major color spectrum histogram representation (CMCSHR) for appearance matching between two objects to improve the performance of major color spectrum histogram representation (MCSHR). The brightness transfer functions (BTFs) are widely used to map an observed brightness value in one camera to the corresponding observation in another camera<sup>[16-20]</sup>. For example, Javed et al.<sup>[16]</sup> demonstrated that the BTFs from a given camera to another camera lie in a low dimensional subspace and demonstrated that this subspace can be used to compute appearance similarity. Gilbert and Bowden<sup>[17]</sup> proposed an algorithm to learn the BTFs incrementally based on consensus-color conversion of Munsell color space. Mazzeo et al.<sup>[20]</sup> compared different methods to evaluate the color BTFs between non overlapping cameras. Actually persons are often characterized by clothing, hairstyles and makeup, and each part has quite different color features, which is important for inter-camera tracking, especially, when the scene contains multiple similar targets. So our method incorporates the above discriminative information to distinguish different persons.

## 1 Inter-Camera Tracking Based on Incremental Learning

The flowchart of human tracking across disjointed cameras based on incremental learning is shown in Fig. 1. The particle filter is exploited in intra-camera tracking, and if a person exits from the camera's FOV, the person's appearance mode is broadcast via the Ethernet to all the other camera units. When a new person is entering a camera's FOV, his/her appearance descriptor is compared with other persons that have exited other cameras before, based on the appearance cue and the min-max time constraint (in the learning phase) or the appearance cue and the spatio-temporal cue (in the tracking phase). The learning phase will

transform into the tracking phase when sufficient evidence has been collected (see section 1.3 for detail).

### 1.1 Learning phase and tracking phase

From Fig. 1, we can see that there are two phases in the flowchart of human tracking across disjoint camera views. When the human tracking system starts to run, there is no prior information. So the learning phase is performed in each camera unit to learn the spatio-temporal information only using the min-max time constraint. In the learning phase, the correspondence of an object appearing in camera  $C_j$  is used to learn the spatio-temporal information in the hypothesis space  $\Gamma$  which is constructed according to the min-max time constraint. Over a period of time, if sufficient observations have been collected in camera  $C_i$ , and the link between camera  $C_i$  and camera  $C_j$  is established (see section 1.4), it means that the spatio-temporal information of camera  $C_i$  and camera  $C_j$  has been learned, then the learning phase stops running and the tracking phase starts to run between camera  $C_i$  and camera  $C_j$ . In the tracking phase, the hypothesis space  $\Gamma$  of an object in camera  $C_j$  is formed according to the learnt spatio-temporal information; then the correspondences in the hypothesis space  $\Gamma$  can be found and the spatio-temporal information can be updated.

It is noticed that in the tracking phase, for some reasons, such as the changing environment, the link between camera  $C_i$  and camera  $C_j$  become ambiguous; then it turns into the learning phase again.

### 1.2 Min-max time constraint

Suppose that we have a surveillance system consisting of  $N$  cameras  $C_1, C_2, \dots, C_N$  with non-overlapping FOVs, and a topological network of the cameras is built to connect all the cameras based on their positions. Let  $t_{ex}^i$  and  $t_{en}^i$  be the exit time and entry time of a person moving from camera  $C_i$  to camera  $C_j$ , respectively, and  $T_{ij\_min}$  and  $T_{ij\_max}$  be positive temporal thresholds which are roughly the minimum time interval and the maximum time interval from cameras  $C_i$  to camera  $C_j$  for one person. Suppose that time  $t_{ex}^i$  and  $t_{en}^j$  satisfy

$$t_{ex}^i + T_{ij\_min} < t_{en}^j < t_{ex}^i + T_{ij\_max} \quad (1)$$

Then camera  $C_i$  and camera  $C_j$  may be adjacent in the topological network. In most situations, this constraint is reasonable because a person's walking speed and the distance between two cameras can be estimated in advance.

Let  $O_{new}$  be a new object that appears in the FOV of camera  $C_j$  and  $O_{ex}$  be an object that exits from camera  $C_i$ . If  $O_{new}$  and  $O_{ex}$  satisfy the temporal constraint in Eq. (1), we consider that object  $O_{ex}$  may be a correspondence candidate of object  $O_{new}$ . All these candidates are combined to form the hypothesis space  $\Gamma$  of  $O_{new}$ . It means that one or more examples in the hypothesis space  $\Gamma$  possibly corresponds to the new object  $O_{new}$ .

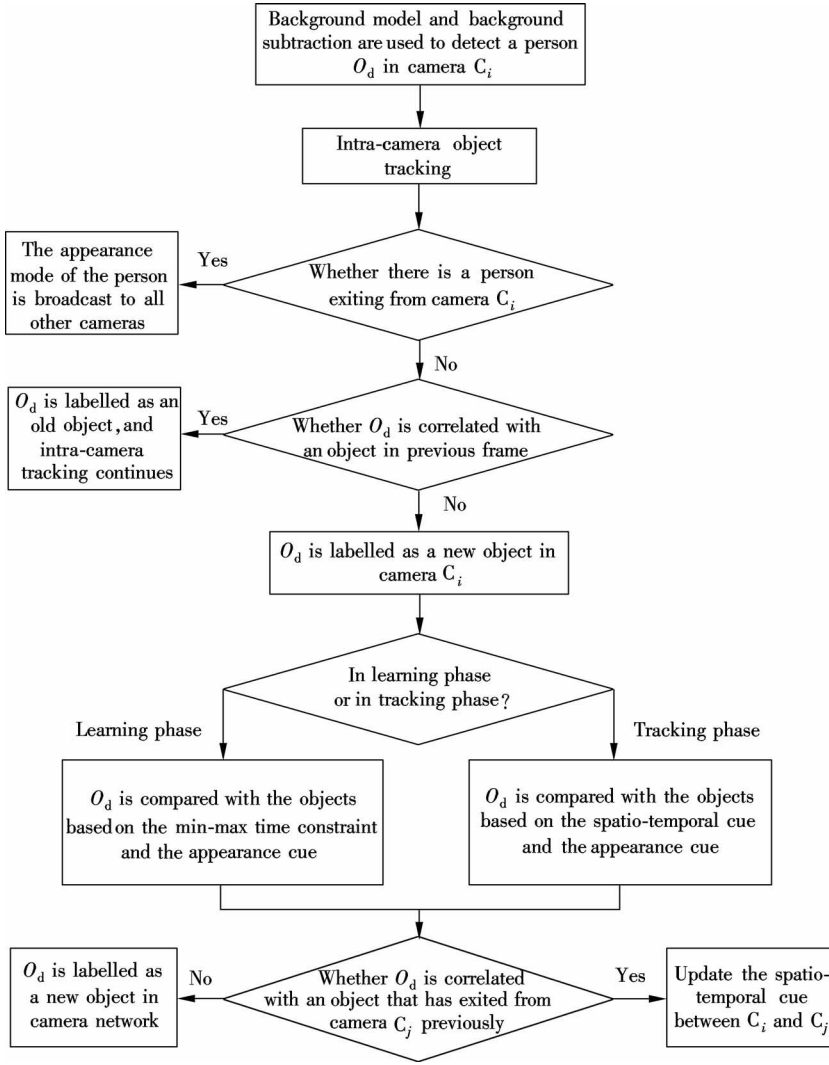


Fig. 1 Flowchart of human tracking across disjoint camera

### 1.3 Appearance model and correspondence

#### 1.3.1 Robust appearance model based on HSV

Our work will be addressed on the study for more reliable appearance matches. The visual cue is more reliable for distinguishing different persons than other cues, especially, in the cases where FOVs are disjointed. However, it is very challenging to design the visual cue because the person's appearance is complex and dynamic in general. A robust appearance model should be adaptive to various camera configurations, tracking targets and environments.

For the person's appearance, it is similar in shape but widely different in color and the person's appearance is dominated by his/her clothes, so color features are more suitable for his/her description. The RGB color space, used directly by most computer devices, expresses colors as the combination of three additive primary colors of light: red, green, and blue. But a commonly used color space that corresponds more naturally to human perception is the HSV color space, whose three components are hue, saturation, and value. Because the HSV color space

is relatively invariable to illumination changes and the histogram is invariant to scale, rotation and translation, we convert the target image from RGB to HSV and choose HSV color histograms to describe the person's appearance. In this paper, the definition of the HSV feature descriptor is defined as

$$H = \begin{cases} 0 & H \in [316, 360] \cup [0, 20] \\ 1 & H \in [21, 40] \\ 2 & H \in [41, 75] \\ 3 & H \in [76, 155] \\ 4 & H \in [156, 190] \\ 5 & H \in [191, 270] \\ 6 & H \in [271, 295] \\ 7 & H \in [296, 315] \end{cases}$$

$$S = \begin{cases} 0 & S \in [0, 0.2] \\ 1 & S \in (0.2, 0.7) \\ 2 & S \in [0.7, 1] \end{cases}, \quad V = \begin{cases} 0 & V \in [0, 0.2] \\ 1 & V \in (0.2, 0.7) \\ 2 & V \in [0.7, 1] \end{cases}$$

Generally, the height and the width of a human should be in proportion to some extent. From anthropometry, the statistical characteristics of the body proportions is ac-

quired. When the height of a body is  $H$ , then the height of the shoulder is  $0.84H$ , and the height of the buttocks is  $0.46H$ . After background subtraction, the tracked person blob is divided into three human body parts (head:  $0.16H$  from the top; leg:  $0.46H$  from the bottom; torso: the rest  $0.38H$ ) according to human body proportions. A 42 dimensional (14 dimensions for  $H$ ,  $S$  and  $V$ , respectively) HSV feature  $f_{\text{HSV}}$  is extracted from every body part in our design. In summary, the appearance mode of a person can be written as  $\text{Person} = (\{f_{\text{Head-HSV}}\}, \{f_{\text{Torso-HSV}}\}, \{f_{\text{Leg-HSV}}\})$ , where  $f_{\text{Head-HSV}}$ ,  $f_{\text{Torso-HSV}}$  and  $f_{\text{Leg-HSV}}$  are the HSV features extracted from head part, torso part and leg part, respectively.

The HSV histogram is used to describe the person's appearance because it is invariant to scale and the spatial distribution of the pixels which change among different cameras. Here we use color distance based on the Euclidean distance to compute the similarity between two targets in the HSV space, which is defined as

$$d(f_1, f_2) = \frac{p(H_1 - H_2) + p(S_1 - S_2) + p(V_1 - V_2)}{3} \quad (2)$$

where  $f_1$  and  $f_2$  represent HSV features of two targets, and  $p(*)$  is the histogram intersection operator.

Due to different view angles, even the HSV features extracted from the same person may not be identical. So a weighted algorithm for similarity measurement based on appearance characteristics of different human parts is proposed. For the head part, because of different colors between hair and face, the histogram of the head is quite different. So a small weight  $W_H$  is assigned to the head's HSV feature  $f_{\text{Head-HSV}}$ . For the torso part, the clothes color is dominant, and sometimes there are differences in pattern and color. For example, the torso's histogram of channel H and channel S from three view angles is similar, but the histogram of channel V is different. So a medium weight  $W_T$  is assigned to  $f_{\text{Torso-HSV}}$ . And for the leg, most pants' colors are unique, and the dominant color is usually distributed evenly, so the HSV feature  $f_{\text{Leg-HSV}}$  of this part is more credible. Then a larger weight  $W_L$  is assigned to  $f_{\text{Leg-HSV}}$ .  $W_H$ ,  $W_T$  and  $W_L$  are set as 0.10, 0.35 and 0.55, respectively in our implementation. Finally, the similarity distance between person  $P_i$  and person  $P_j$  can be defined as

$$\begin{aligned} d(P_i, P_j) = & w_H d(f_{\text{Head-HSV}_i} - f_{\text{Head-HSV}_j}) + \\ & w_T d(f_{\text{Torso-HSV}_i} - f_{\text{Torso-HSV}_j}) + \\ & w_L d(f_{\text{Leg-HSV}_i} - f_{\text{Leg-HSV}_j}) \end{aligned} \quad (3)$$

### 1.3.2 Person correspondence

In order to decrease the wrong correspondence, a matching strategy with two thresholds is proposed. When a person called  $O_i$  appears in one camera, the hypothesis space

$\Gamma$  is formed by using the min-max time constraint (see section 1.1). The similarity distance between  $O_i$  and each example in the hypothesis space  $\Gamma$  is calculated (see Eq. (3)), then distances are sorted in ascending order, namely  $d_{\min}, d_2, \dots, d_{\max}$ . The smaller distance between two objects means that they are more similar. In order to increase the accuracy of matching, two thresholds,  $M_{\text{Th1}}$  and  $M_{\text{Th2}}$ , are involved in the matching algorithm. If  $d_{\min} \geq M_{\text{Th1}}$ , it means that the appearance of person  $O_i$  is not similar to the appearance of others. So person  $O_i$  is likely to be a new one; otherwise,  $O_i$  may be matched with one example in hypothesis space  $\Gamma$ . If there is only one distance less than  $M_{\text{Th1}}$ , the correspondence is the example whose distance with  $O_i$  is  $d_{\min}$ . But the case that there are two or more examples whose distances are less than threshold  $M_{\text{Th1}}$  should be considered. Here a second constraint is added to avoid a possible mismatching. Suppose that these  $M$  examples whose distances are less than threshold  $M_{\text{Th1}}$  from a subset  $\{d_{\min}, d_2, \dots, d_M\}$  named  $E$ . If  $d_2 - d_{\min} > M_{\text{Th2}}$ , it is believed that the correspondence is the example whose distance with  $O_i$  is  $d_{\min}$ . If not, the exit time of the example and the entry time of  $O_i$  are considered and we have good grounds to believe that the correspondence is the example whose exit time is closest to the entry time of  $O_i$  in subset  $E$ .  $M_{\text{Th1}}$  and  $M_{\text{Th2}}$  are set as 0.35 and 0.2 in our implementation. The detail is shown in Algorithm 1.

#### Algorithm 1 Matching algorithm for one person

Compute the hypothesis space  $\Gamma$ ;

Compute the HSV feature vector of person  $O_i$ ;

Compute the HSV feature vector of each Example $_i^k$  in  $\Gamma$ ;

Compute  $d(O_i, \text{Example}_i^k)$ , and sort them by ascending order, namely  $d_{\min}, d_2, \dots, d_{\max}$ ;

if  $d_{\min} \geq M_{\text{Th1}}$  Not match;  
else

$M$  is the number of examples whose distance is less than  $M_{\text{Th1}}$ ;

if  $M = 1$  Select the example with distance  $d_{\min}$  as the correspondence;

else

if  $d_{\min}, d_2, \dots, d_M < M_{\text{Th1}}$

if  $d_2 - d_{\min} > M_{\text{Th2}}$  Select the example with distance  $d_{\min}$  as the correspondence;

else Select the example with the shortest time interval as the correspondence.

### 1.4 Update of spatio-temporal relationship

To learn the spatio-temporal relationship between any two cameras, we make use of the key assumption that persons will follow similar routes in the surveillance network as time elapses, and the repetition of the routes will form marked and consistent trends in the overall data. These trends among camera networks can be used to learn

the spatio-temporal relationship.

#### 1.4.1 Spatial information updating

In this paper, the spatial information does not refer to the existing probability of a path between cameras proposed in some work, but refers to whether there is a path existing from one camera to the others. Here the link rate  $L_{ij}$  of objects moving from camera  $C_i$  to camera  $C_j$  is introduced, which is calculated from the ratio of  $N^{ij}$  and  $T_{\text{elap}}^j$ , where  $N^{ij}$  is the number of correspondences from camera  $C_i$  to camera  $C_j$  and  $T_{\text{elap}}^j$  is the elapsed time of the camera  $C_i$  running. If an object entering camera  $C_j$  corresponds with an object that has exited camera  $C_i$  previously, it is  $N^{ij}$  plus one. Meanwhile, two proper thresholds,  $M_{\text{Th}}$  and  $N_{\text{max}}$ , are involved to prevent an improper link. If the link rate  $L_{ij}$  is smaller than the threshold  $M_{\text{Th}}$  or if  $N^{ij}$  is less than  $N_{\text{max}}$ , it is not sure that there is a path from camera  $C_i$  to camera  $C_j$ . Otherwise, there is a path from camera  $C_i$  to camera  $C_j$ .  $M_{\text{Th}}$  and  $N_{\text{max}}$  should be set according to different scenarios.

#### 1.4.2 Temporal information updating

The temporal information refers the statistical time range of a person walking from camera  $C_i$  to camera  $C_j$ , including the maximum time interval  $t_{ij\_max}$  and the minimum time interval  $t_{ij\_min}$ . Fig. 2 shows a time interval distribution for a person walking from camera  $C_2$  to camera  $C_4$ . It shows that the time interval distribution from camera  $C_2$  to camera  $C_4$  is almost between 8 s ( $t_{ij\_min}$ ) and 20 s ( $t_{ij\_max}$ ), and a distinct peak around 16 s indicates that most people take about 16 s to walk through the distance. It is worth noting that  $t_{ij\_max}$  is smaller than  $T_{ij\_max}$  and  $t_{ij\_min}$  is greater than  $T_{ij\_min}$ .

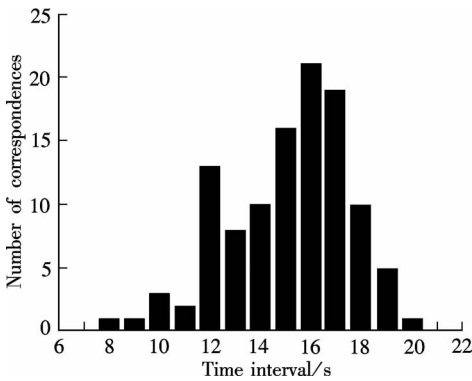


Fig. 2 Time interval distribution from camera  $C_2$  to camera  $C_4$

#### 1.4.3 Correspondence FIFO

The motion trends of the people may change with time. For example, as shown in Fig. 3, there is a path from camera  $C_1$  to camera  $C_4$  when LAB3 is open. However, suppose that LAB3 is closed due to some reasons, then the path from camera  $C_1$  to camera  $C_4$  should be removed. So it is necessary to update the spatio-temporal information between cameras during the lifetime of the system. Suppose that  $A_{ij}$  is a correspondence FIFO to

store the maximum  $N$  correspondences of people from camera  $C_i$  to camera  $C_j$  in the recent  $T_{\text{FIFO}}$  period, in which  $N$  and  $T_{\text{FIFO}}$  are both constants. Actually  $N$  should be large enough to guarantee that  $A_{ij}$  can store enough numbers of correspondences in the  $T_{\text{FIFO}}$  period. The detailed process is shown in Algorithm 2.

**Algorithm 2** Correspondence FIFO operation process

```

 $t_{ij\_oldest}$  is the established time of the oldest correspondence in  $A_{ij}$ ;
 $t_{ij\_newest}$  is the established time of the newest correspondence in  $A_{ij}$ ;
while  $t_{ij\_newest} - t_{ij\_oldest} > T_{\text{FIFO}}$  Delete the oldest match from  $A_{ij}$ ;
if a new correspondence is established
  if  $A_{ij}$  is not full Add the current match into  $A_{ij}$ ;
  else Delete the oldest match from  $A_{ij}$  and add the current match into  $A_{ij}$ ;

```

Using the correspondence in  $A_{ij}$  to update the spatio-temporal information.

## 2 Human Tracking Experiments

Different experiments are carried out to test the multi-camera human tracking algorithm in a disjoint multi-camera view scenario. This scenario environment is shown in Fig. 3. It is a laboratory room which contains four indoor cameras with non-overlapping FOV, namely camera  $C_1$ , camera  $C_2$ , camera  $C_3$ , and camera  $C_4$ .

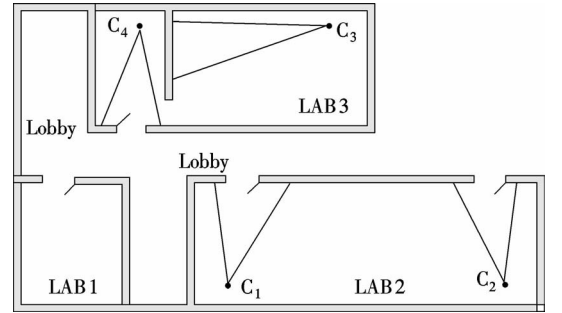


Fig. 3 Camera network topology in experiments

In the two scenarios, all the surveillance cameras are DS-2CD853F with  $1600 \times 1200$  high resolution and an acquisition frame rate of 12 frame/s. Each surveillance camera is connected to a computer which serves as a process module and communicates with the data server via the Ethernet network.

### 2.1 Appearance matching experiment

Three test cases are designed in the experiments to test the performance of the proposed appearance model with different view angles and a little illumination change. The configuration of three test cases is shown in Tab. 1. In the experiment, the persons in the image are segmented automatically, and Tab. 1 shows the results of automated appearance matching. The results indicate that within the

disjoint camera, disjoint tracks of the same person tend to be correctly matched with high accuracy despite different chromatic responses in the two cameras, the change of the view angles and the environmental factors.

**Tab. 1** Results of automated appearance matching

Test case	Number of people	Cameras used	Matched people	Unmatched people
1	30	C <sub>3</sub> and C <sub>4</sub>	25	5
2	30	C <sub>2</sub> and C <sub>1</sub>	26	4
3	30	C <sub>3</sub> and C <sub>2</sub>	26	4

Furthermore, to demonstrate the superiority of the proposed appearance model, two appearance models based on the whole body part (whole-body model) and three unweighted body parts (unweighted model) are introduced for comparison. The whole-body model is based on

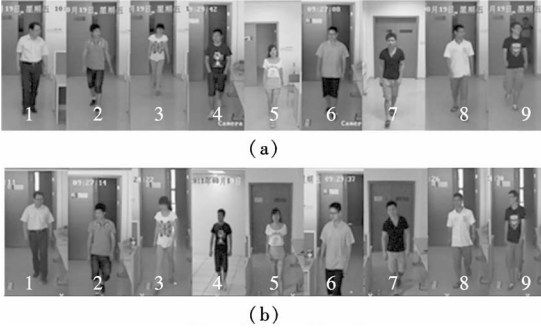
the HSV feature extracted from the whole person target, and the unweighted model is based on the HSV feature extracted from three body parts of people with equal ratios. Several pairs of observations are used to test the discrimination of the three methods. Tab. 2 and Tab. 3 show the similarity distance of two observations belonging to the same person (see Fig. 4) and the similarity distance of two observations belonging to different persons (see Fig. 5) with three methods, respectively. And a visual summary of matching comparison results is shown in Fig. 6. It is shown that the distinguished capability of the proposed appearance model is better than that of the other two models and the matching ratio of the proposed appearance model is higher than that of the other two models.

**Tab. 2** Similarity distance of the same observations with three methods

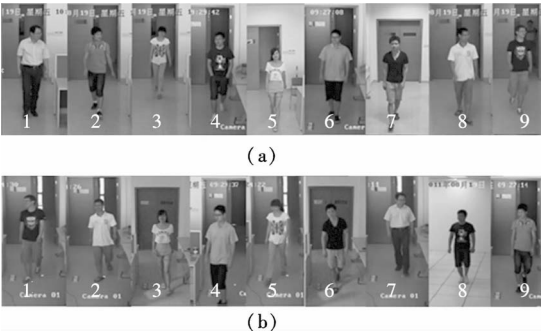
Method	Observation								
	1	2	3	4	5	6	7	8	9
Proposed method	0.10	0.11	0.20	0.13	0.14	0.13	0.18	0.15	0.16
Whole-body model	0.17	0.27	0.34	0.22	0.21	0.27	0.30	0.25	0.27
Unweighted model	0.11	0.21	0.23	0.18	0.17	0.19	0.22	0.17	0.24

**Tab. 3** Similarity distance of different observations with three methods

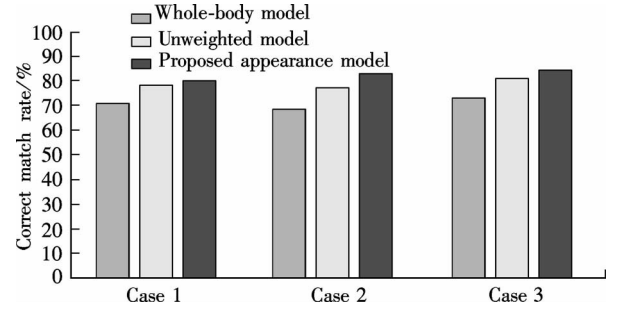
Method	Observation								
	1	2	3	4	5	6	7	8	9
Proposed method	0.92	0.87	0.75	0.85	0.73	0.87	0.88	0.90	0.84
Whole-body model	0.83	0.68	0.59	0.75	0.66	0.82	0.81	0.81	0.72
Unweighted model	0.85	0.76	0.71	0.82	0.70	0.86	0.82	0.85	0.77



**Fig. 4** The same observations. (a) Observations A; (b) Observations B



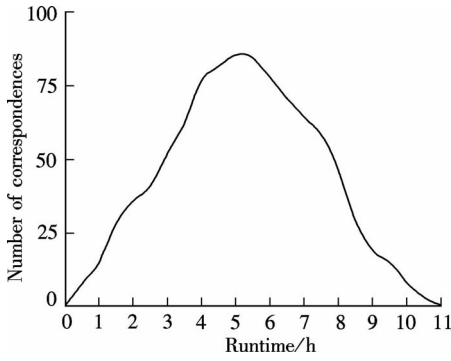
**Fig. 5** Different observations. (a) Observations A; (b) Observations B



**Fig. 6** A comparison of the correct match using three appearance models

## 2.2 Spatio-temporal information learning experiment

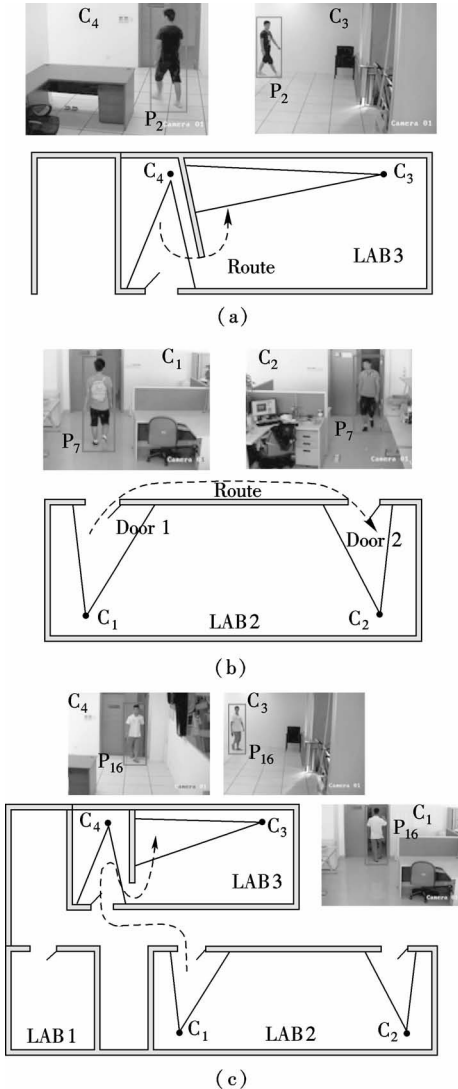
Fig. 7 shows the change trend of the spatio information from camera C<sub>1</sub> to camera C<sub>3</sub> when the door of LAB 3 closes after the system has run continuously for 6 h. When the door of LAB 3 closes, there is no person passing from camera C<sub>1</sub> to camera C<sub>3</sub> and no new correspondence added to the FIFO, so the number of correspondences decreases gradually toward zero according to the spatio-temporal information updating algorithm (see section 1.4).



**Fig. 7** Correspondence number distribution as the environment changes

### 2.3 Human tracking experiment

This section presents results obtained by automatically tracking human in a camera network whose layout is shown in Fig. 3. Fig. 8 shows the results of the human tracking in three scenarios respectively.



**Fig. 8** The result images. (a) Scenario 1; (b) Scenario 2; (c) Scenario 3

In scenario 1, about 30 persons exit from the FOV of  $C_4$  and enter the FOV of  $C_3$  in LAB3. The illumination condition and the view angle of the FOV of  $C_4$  are different from those of  $C_3$ . The learning phase has been performed for about one hour. One tracking result in scenario 1 is shown in Fig. 8 (a), where a person is assigned a correct label as he walks across two disjoint camera views. The left image of the first row shows someone is leaving the FOV of  $C_4$ , and the right image of the first row shows that the corresponding persons are entering the FOV of  $C_3$ . And the image of the second row shows the route of a person walking in the camera network topology.

In scenario 2, several persons walk out of the FOV of  $C_1$  in LAB 2, and a little while later, a person reenters LAB 2 and appears in the FOV of  $C_2$ . One tracking result in scenario 2 is shown in Fig. 8(b), where a person is assigned a correct label as he walks across two disjoint camera views.

A more complicated scenario is constructed in scenario 3, where a person exits from the FOV of  $C_1$  in LAB 2, then the person enters the FOV of  $C_4$  in LAB 3 and, for a little while, the person appears in the FOV of  $C_3$ . Fig. 8 (c) shows one tracking result in scenario 3, where a person is assigned a correct label as he walks across three disjoint camera views.

From the above, it is shown that even in the real environment with different illumination conditions and different view angles, the proposed method can achieve good tracking results.

### 3 Conclusion

In this paper, we present a robust method for human tracking. The contribution of our work focuses on building a discriminative appearance feature extracted from different human parts based on the HSV histogram to describe the tracked person and constructing an incremental learning method to acquire the spatio-temporal information with the time constraint and the weighted algorithm. To match two people, a weighted algorithm is used to compute the similarity distance with the time constraint, and then a similarity sort algorithm with two thresholds is used to decide the correspondence. The experimental results show that the tracking accuracy among spatially separated un-calibrated cameras increases with time and can adapt to the environmental changes, using no a priori information in a completely unsupervised fashion.

### References

- [1] Kuo C H, Huang C, Nevatia R. Inter-camera association of multi-target tracks by on-line learned appearance affinity models[C]//*Proceedings of the 11th European Conference on Computer Vision*. Berlin, Germany, 2010: 383 – 396.

- [2] Cai Q, Aggarwal J K. Tracking human motion in structured environments using a distributed-camera system[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999, **21**(11): 1241 – 1247.
- [3] Collins R T. Algorithms for cooperative multisensor surveillance [J]. *Proceedings of the IEEE*, 2001, **89**(10): 1456 – 1477.
- [4] Khan S, Shah M. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, **25**(10): 1355 – 1360.
- [5] Huang T, Russell S. Object identification in a Bayesian context [C]//*International Joint Conferences on Artificial Intelligence*. San Francisco, USA, 1997: 1276 – 1283.
- [6] Pasula H, Russell S, Ostland M, et al. Tracking many objects with many sensors [C]//*Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. San Francisco, USA, 1999: 1160 – 1171.
- [7] Kettner V, Zabih R. Bayesian multi-camera surveillance [C]//*Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Fort Collins, CO, USA, 1999: 253 – 259.
- [8] Javed O. Tracking across multiple cameras with disjoint views [C]//*Proceedings of the IEEE International Conference on Computer Vision*. Washington DC, USA, 2003: 952 – 957.
- [9] Dick A R, Brooks M J. A stochastic approach to tracking objects across multiple cameras [C]//*Proceedings of Australian Conference on Artificial Intelligence*. Berlin, Germany, 2004: 160 – 170.
- [10] Makris D, Ellis T, Black J. Bridging the gaps between cameras [C]//*Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington DC, USA, 2004: 205 – 210.
- [11] Wang X, Tieu K, Grimson E. Correspondence-free activity analysis and scene modeling in multiple camera views [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(1): 32 – 17.
- [12] Song B, Roy-Chowdhury A K. Robust tracking in a camera network: a multi-objective optimization framework [J]. *IEEE Journal on Selected Topics in Signal Processing*, 2008, **2**(4): 582 – 596.
- [13] Porikli F. Inter-camera color calibration by correlation model function [C]//*IEEE International Conference on Image Processing*. Barcelona, Spain, 2003: 133 – 136.
- [14] Madden C, Cheng E D, Piccardi M. Tracking people across disjoint camera views by an illumination-tolerant appearance representation [J]. *Machine Vision and Applications*, 2007, **18**(3): 233 – 247.
- [15] Lian G, Lai J, Zheng W. Spatial-temporal consistent labeling of tracked pedestrians across non-overlapping camera views [J]. *Pattern Recognition*, 2011, **44**(5): 1121 – 1136.
- [16] Javed O, Shafique K, Shah M. Appearance modeling for tracking in multiple non-overlapping cameras [C]//*IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington DC, USA, 2005: 26 – 33.
- [17] Gilbert A, Bowden R. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity [C]//*Proceedings of the 9th European Conference on Computer Vision*. Berlin, Germany, 2006: 125 – 136.
- [18] Prosser B, Gong S, Xiang T. Multi-camera matching using bi-directional cumulative brightness transfer functions [C]//*British Machine Vision Conference*. London, 2008: 1 – 10.
- [19] Jeong K, Jaynes C. Object matching in disjoint cameras using a color transfer approach [J]. *Machine Vision and Application*, 2008, **19**(5): 443 – 455.
- [20] Mazzeo P, Spagnolo P, Orazio T. Object tracking by non-overlapping distributed camera network [C]//*Proceedings of the Advanced Concepts for Intelligent Vision Systems*. Bordeaux, France, 2009: 516 – 527.

## 在非重叠视域监控网络中的人体目标跟踪

林国余 杨 彪 张为公

(东南大学仪器科学与工程学院, 南京 210096)

**摘要:**针对存在非重叠视野的摄像机监控网络,提出了一种基于人体外观模型和摄像机间时空信息的人体目标自适应跟踪算法.对于人体外观模型,首先根据人体测量学理论将人体目标划分成头、躯干和腿3个部分,分别提取各部分的HSV颜色直方图特征用于构建人体外观模型,然后引入加权因子计算人体目标之间的相似度,最后采用一种基于双阈值的相似度排序算法确定人体目标的匹配关系.对于摄像机间的时空信息,通过增量学习,不断积累目标关联信息,经统计分析逐步更新摄像机间时空信息.实验结果验证了所提出的跟踪算法在无需摄像机标定的条件下能够实现人体目标的连续跟踪,且随着关联匹配信息的累加,算法的跟踪准确性也逐步提高.

**关键词:**多摄像机跟踪;非重叠视域;时空关系;人体外观模型;增量学习

**中图分类号:**TP391