Adaptive Human Tracking Across Non-overlapping Cameras in Depression Angles

Shao Quan(邵荃)[1*], *Liang Binbin*(梁斌斌)[1], *Zhu Yan*(朱燕)[1],
Zhang Haijiao(张海蛟)[1], *Chen Tao*(陈涛)[2]

1. College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, P. R. China;
2. Institute of Public Safety Research, Tsinghua University, Beijing 100084, P. R. China

(Received 13 November 2014; revised 7 January 2015; accepted 12 January 2015)

Abstract: To track human across non-overlapping cameras in depression angles for applications such as multi-airplane visual human tracking and urban multi-camera surveillance, an adaptive human tracking method is proposed, focusing on both feature representation and human tracking mechanism. Feature representation describes individual by using both improved local appearance descriptors and statistical geometric parameters. The improved feature descriptors can be extracted quickly and make the human feature more discriminative. Adaptive human tracking mechanism is based on feature representation and it arranges the human image blobs in field of view into matrix. Primary appearance models are created to include the maximum inter-camera appearance information captured from different visual angles. The persons appeared in camera are first filtered by statistical geometric parameters. Then the one among the filtered persons who has the maximum matching scale with the primary models is determined to be the target person. Subsequently, the image blobs of the target person are used to update and generate new primary appearance models for the next camera, thus being robust to visual angle changes. Experimental results prove the excellence of the feature representation and show the good generalization capability of tracking mechanism as well as its robustness to condition variables.

Key words: adaptive human tracking; appearance features; geometric features; non-overlapping camera; depression angle

CLC number: TP391.41 **Document code**: A **Article ID**: 1005-1120(2015)01-0048-13

0 Introduction

Multi-camera visual surveillance systems have currently been widely distributed in many areas for applications of continuously tracking interesting objects, early-warning of abnormal events and so on. Particularly, in some cases of multi-airplane visual human tracking, aerial photography and urban multi-camera surveillance from tall buildings, there exists a multi-camera visual surveillance system in which all cameras are installed at high places with wide visual range and have large visual depression angles. A fundamental task for these particular multi-camera surveillance systems with visual depression angles is to associate people across different camera views at different locations and time. This is known as the human tracking problem in visual depression angles.

Human tracking in visual depression angles faces an issue of visually matching a target person across different cameras distributed over disjoint scenes of distance and time differences. In this case, classical human tracking algorithms will fail since cameras do not overlap. Hence, non-overlapping camera human tracking in this paper describes algorithms that deal with human tracking across non-overlapping camera views. Non-overlapping camera human tracking techniques build upon single camera human tracking techniques,

* **Corresponding author**: Shao Quan, Associate Professor, E-mail: shaoquan@nuaa.edu.cn.

How to cite this article: Shao Quan, Liang Binbin, Zhu Yan, et al. Adaptive human tracking across non-overlapping cameras in depression angles[J]. Trans. Nanjing U. Aero. Astro., 2015,32(1):48-60.
http://dx.doi.org/10.16356/j.1005-1120.2015.01.048

for a person needs to be tracked within one camera field of view (FOV) before it can be tracked in that of another. Therefore, the problem of non-overlapping camera human tracking becomes how to match individual from one independent surveillance area to another. It has many challenges among which feature representation and tracking mechanism are the most difficult. The tracked people should be differentiated from numerous visually similar but different people in those views, which requires a sufficiently discriminative feature representation to distinguish the target person from those similar yet different candidates. Potentially, different views may be taken from various shooting angles, causing dissimilar backgrounds under diverse illumination conditions, or with other view variables. It thus requires a robust tracking mechanism that can resist inter-camera and intra-camera shooting angle changes, as well as illumination change.

Designing suitable feature representation for human tracking is a critical and challenging problem. Ideally, the features should be robust to illumination changes, visual angle altering, foreground errors, occlusion, and low image resolution. Contemporary approaches typically exploit low-level features such as appearance[1-3], spatial structure[4-5], or their combinations[6-12]. This is because these features can be relatively easily and reliably measured. Moreover, they provide a reasonable level of inter-person discrimination and then can distinguish different people clearly.

Gianfranco et al.[1] and Hyun-Uk et al.[2] used appearance to re-identify people, and proved that appearance feature had good performance in identifying individuals. However, they could not deal with illumination change very well. Generally, in single visual angle individuals can be discriminated based on their appearances. However, appearance feature will alter with the change of visual angle, which occurs frequently across non-overlapping cameras. In this case, appearance is quite limited to distinguish individuals. Related researches compensated the limitation by combining geometric features with appearance[6-10]. Madden et al.[6] focused on a framework based on robust shape and appearance features. However, his proposal solely employed height as shape feature without considering the limitation of height in distinguishing human beings due to the close resemblance of human height. A good way to amend the limitation of height is to combine gait feature with height to compose a shape feature. Takayuki et al.[8] proposed a method that tracked a walking human using gait features to calculate a robust motion signature. Despite the well-done performance of gait feature in his method, a high accuracy rate and low computational cost were still far from being achieved[9]. Moreover, gait features are tough to adjust to the visual angle change. In short, it needs more explorations to obtain a better feature representation.

Once a suitable feature representation has been obtained, previous literatures typically used the nearest-neighbor[4] or model-based matching algorithms such as support-vector ranking[11] for human tracking. In each case, a distance metric must be chosen to measure the similarity between two samples. In single camera, both the model-based matching approaches[12-13] and the nearest neighbor distance metrics[14-15] can be optimized to maximize tracking performance. However, despite their excellent performance in single camera, they are still limited in coping with those intractable challenges in non-overlapping cameras. The first challenge is to overcome the inter-camera and intra-camera variations. These variations include the change of appearance feature, spatial structure, illumination condition and some other parameters, which makes non-overlapping camera human tracking tough to work well in various scenes from different visual angles. Furthermore, such variations between non-overlapping cameras are in general complex and multi-modal, and therefore complex for an algorithm to learn. The second challenge is how to achieve a good generalization capability. Previously, once trained for a specific pair of cameras, most models could not generalize well to other cameras from different visual angles[16] because there was

no connection between them. Therefore, it is necessary to establish a tracking mechanism with good generalization that models can be established once and then adaptively applied to different camera configurations.

To solve the problems mentioned above, improved appearance feature and geometric feature are explored, and an adaptive tracking mechanism is also designed. The improved appearance and geometric features make up a discriminative and robust human feature. The appearance feature includes color and texture information. Color is analyzed in hue-saturation-value (HSV) space. The HSV space is evenly partitioned and generates less color histogram bins than previous literatures, thus cutting the computational cost. Texture histograms are generated through an improved direction coded local binary pattern descriptor and they can describe local texture distribution better. As for the geometric feature, the mean value and standard deviation of height estimates of multi-shot blobs are calculated. Superior to single-shot geometric analysis, these two statistic parameters are computed from multi-shot blobs and can suppress the disturbance from noise blobs. Besides, these two geometric parameters can easily reflect height and gait movement simultaneously. To our knowledge, it is original to extract such geometric features in a statistical way of this paper.

In human tracking process, an adaptive tracking mechanism is designed. It aims to automatically match and track individuals based on both retrospective and on-the-fly information. The image blobs are divided into two groups, namely, the gallery group gathered by source images and the probe group gathered by target images. The gallery group trains the computation parameters of the target person and then tests the image blobs in the probe group. On one hand, in the gallery group the appearance feature from each visual angle is described by a primary appearance model. This paper creates primary appearance model to represent unique appearance feature seen from unique visual angle. On the other, the probe blobs are divided into groups based on state-of-the-art single camera human tracking techniques. Each group corresponds to a unique person and includes the appearance information of him/her. The geometric features of the groups are first compared with those of the gallery group. The geometrically similar groups are subsequently described and arrayed in an appearance model matrix, and then matched with the gallery primary appearance models. The group that has the maximum label scale with gallery primary appearance models is determined to be the target person. After being targeted, the blobs of the person are automatically collected into source blob sequences and updated for obtaining new gallery primary appearance models. The mechanism obtains a powerful generalization capability among different cameras. Superior to the existed methods, it has an increasingly better performance over time in term of generalization.

1 Motion Detection

Fig. 1 shows a configuration of the proposed method. This paper employs ViBe to detect and segment moving objects[18]. After segmenting an object, an external bounding box is used to contain it. The object segmentation actually obtains a foreground blob. Then the height to weight (H/W) ratio of foreground blob is computed. If its H/W ratio is between 5 and 10, this foreground blob is recognized as a human foreground blob. Otherwise, it will be deleted.

2 Appearance Feature Extraction

2.1 Partial body analysis

Instead of focusing on global appearance features, this paper analyzes the appearance features from human body parts. The areas around the chest, thigh and foot are chosen as the three parts which describes the most critical information of individual, and thereby the feature extraction becomes faster whilst not losing important appearance information. In this paper, the chest corresponds to the 15%—40% region of the external bounding box , while the thigh corresponds to

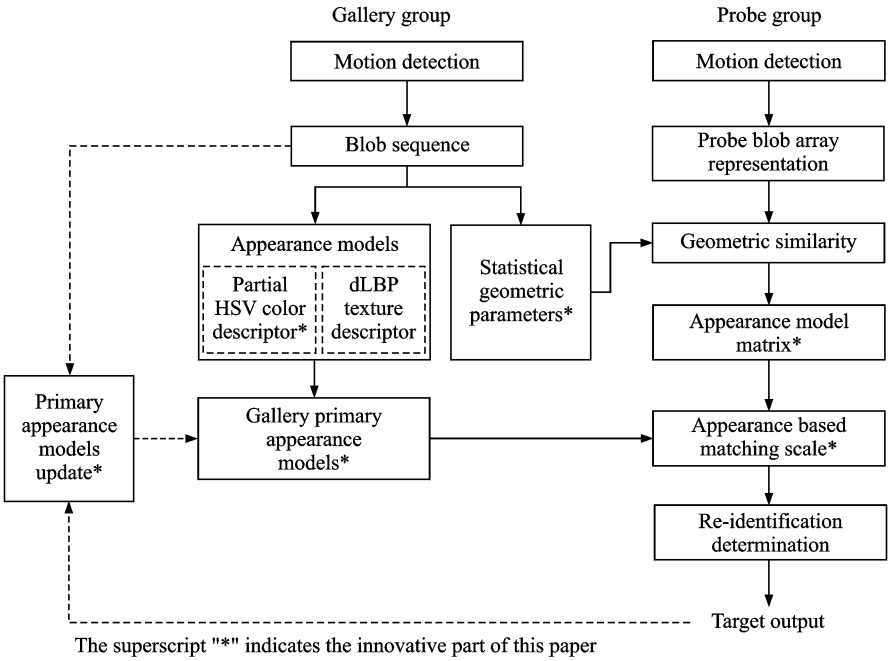

Fig. 1 Process flow diagram for the proposed method

50%—70% and the foot 85%—100%. The chest weights 60% in term of importance, the thigh 25% and the foot 15%.

2.2 Appearance model

Appearance model contains color and texture information. Color expresses chromatic information and texture gives spatial distribution information of pixels. The combination of color and texture enhances the discriminability to distinguish different people.

2.2.1 Color analysis

Color often varies with the illumination change. To cope with this, the HSV color space is employed. In the HSV color space, the effect from illumination change can be suppressed by decreasing Value characteristic. For an initial red-green-blue (RGB) image, it is first converted into HSV space[19]. After executing the convert, H has values from 0 to 360, S from 0 to 1, and V from 0 to 255.

In most previous appearance models, each characteristic in color space generated a histogram and thus produced massive bins which would be time-consuming. For saving time, this paper partitions HSV color space into rough segments. The Hue characteristic is partitioned evenly into Q_H segments, Saturation into Q_S segments and Value into Q_V segments. The values of H, S, V components are then converted into segment level H_C, S_C, V_C, respectively.

$$H_C=(H\times Q_H)/360,\ S_C=S\times Q_S$$
$$V_C=(V\times Q_V)/255 \tag{1}$$

Then H_C, S_C and V_C are integrated into a color descriptor γ_{HSV} as follows

$$\gamma_{HSV}=\mathrm{round}(\gamma_H H_C+\gamma_S S_C+\gamma_V V_C) \tag{2}$$

where γ_H, γ_S, γ_V denote the weight coefficients of H_C, S_C, V_C, round(·) represents a rounding function. γ_H, γ_S, γ_V are computed by

$$\begin{cases}\gamma_H=Q_S Q_V/(Q_S Q_V+Q_H+1)\\ \gamma_S=Q_H/(Q_S Q_V+Q_H+1)\\ \gamma_V=1/(Q_S Q_V+Q_H+1)\end{cases} \tag{3}$$

Since Q_S, Q_V and Q_H are all larger than 1, the weight of Value γ_V will be always less than γ_H and γ_S. It suppresses the impact of brightness and strengthens the robustness to illumination change. To avoid losing too much Saturation information, the segment numbers should be sorted in the descending order of $Q_H>Q_S>Q_V$. Here, the values of Q_H, Q_S and Q_V are 250, 50 and 5, respectively.

2.2.2 Texture analysis

Local binary pattern (LBP) descriptor is one

of the widely-used texture descriptors. Comparing with general LBP, the direction coded LBP (dLBP) descriptor presented in Ref. [20] considers the relations between center pixel and neighboring pixels, as well as the relations among border pixels along one direction, and thus can better describe the texture in interesting regions. Different from Ref. [20], we converts the eight-bit dLBP binary code into a decimal descriptor as

$$\mathrm{dLBP}_{NS,R}=\sum_{p=0}^{NS'-1}(s(v_p+v_{2NS'-p}-2v_c)2^{2p}+ s(|v_p-v_c|-|v_{2NS'-p}-v_c|)2^{2p+1}) \quad (4)$$

$$s(x)=\begin{cases}1 & x\geqslant 0\\ 0 & \text{Others}\end{cases} \quad (5)$$

where NS stands for the neighbor size and $NS=2NS'$. v_p is the value of pixel regularly spaced on circle and v_c the value of center pixel.

The dLBP in Eq. (4) involves not only comparison information between border pixels and center pixel but also comparison information among border pixels themselves along one direction, and it can rank the three pixel values along one direction. Therefore, it can better discriminate person's appearance.

2.2.3 Appearance modeling

Person's appearance feature is modeled by the pixels from chest, thigh and foot regions. Each pixel has its color and $\mathrm{dLBP}_{P,R}$ value, and each region will thus generate a color histogram and a dLBP texture histogram. The appearance model is constructed by concatenating six histograms from left to right in such an order: the color histogram of chest, the color histogram of thigh, the color histogram of foot, the texture histogram of chest, the texture histogram of thigh, and the texture histogram of foot.

3 Geometric Feature Extraction

3.1 Height estimation

Before calculating height estimate, apparent height should be measured at first. Apparent height is computed as the length from the middle top to the middle bottom of the bounding box. This paper calculates height estimate in the same way as Ref. [21].

3.2 Statistical geometric features

Statistical geometric features are obtained by computing the statistical parameters of height estimates. To suppress the impact from noise height estimates, the heights of blobs are sorted in a descending order, at the same time, the tallest and the shortest 5% will be deleted. The remaining height estimates compute the height estimate parameters. This paper uses mean and standard deviation of the remaining height estimates as the statistical geometric parameters.

Imagining that there are N blobs in training sample, and the height estimate of blob i is denoted as h_i. After deleting the tallest and shortest 5% blobs, the remaining blobs can be ranked as h_m, h_{m+1}, h_{m+2}, …, h_{M-1}, h_M, …, h_{n-2}, h_{n-1}, h_n in a descending order, where $n-m+1=90\%\cdot N$. The statistical geometric parameters are calculated by

$$\begin{cases}\mu_h=\sum_m^n h_i/(n-m+1)\\ \sigma_h=\sqrt{\left(\sum_m^n h_i-\mu_h\right)^2/(n-m+1)}\end{cases} \quad (6)$$

where μ_h measures the stature of a person and σ_h reflects the rhythm up and down displacement of the upper body. These two parameters connect the height feature with the gait feature in a simple way.

3.3 Geometric similarity

Once the statistical geometric parameters have been obtained, the similarity of geometric features of source sequences (μ_s, σ_s) and target sequences (μ_t, σ_t) is computed as

$$Sim_{s,t}=\omega_\mu Sim_\mu+\omega_\sigma Sim_\sigma \quad (7)$$

where $\omega_\mu=\omega_\sigma=0.5$, representing the weight of mean value and standard deviation respectively. Sim_μ and Sim_σ denote the similarity of (μ_s, μ_t) and (σ_s, σ_t) respectively.

$$Sim_\mu=(\mu_s-\mu_0-|\mu_s-\mu_t|)/(\mu_s-\mu_0) \quad (8)$$

$$Sim_\sigma=\begin{cases}(\sigma_s-|\sigma_s-\sigma_t|)/\sigma_s & 2\sigma_s>\sigma_t\\ 0 & \text{Others}\end{cases} \quad (9)$$

where μ_0 is a constant related to real situations. If $Sim_{s,t}$ is greater than threshold T_{geo}, it proves that source sequence and target sequence are similar.

4 Human Tracking

From different visual angles, the same person may appear strikingly different, especially when the colors and textures of the front, side and back clothing are totally diverse. But as long as one appears under the same visual angle, his or her appearance models will be pair-camera correlated. This correlation can help us identify the same person from frame to frame and track him/her in different views. Accordingly, authors collect the maximum appearance models from different visual angles, and compares them in the coming camera view with the collected models, so as to immediately target the person as soon as he/she appears under a collected visual angle.

When a person walks into a specified visual angle, he/she may produce a blob sequence, which mirrors the appearance feature in this specified visual angle. Ideally, each sequence corresponds to a visual angle, and the amount of sequences is equal to the amount of visual angles. However, since the noise blob sequences are unavoidable, thus leading to noise appearance models, primary appearance models are selected in an adaptive mechanism to prevent the noise appearance models.

4.1 Primary appearance models

For a person, each of his/her blobs has an appearance model and all the appearance models are constructed in the method described in Section 2.2.3. The appearance models are denoted as m_i ($i=1,2,3,\cdots,n$) and will be divided into classes. Those classes with over-threshold model population are selected as primary appearance model classes.

4.1.1 Appearance model classification

Appearance model classification is based on the pair-wise distance of appearance models. The appearance model m_i has six histograms, including three color histograms $H^i_{\text{color},j}$ and three dLBP histograms $H^i_{\text{dLBP},j}$ ($j=1,2,3$ corresponding to chest, thigh and foot). The distance of m_i and $m_{i'}$ is computed from the correlations of the six histograms in the following

$$Dis(m_i,m_{i'})=\sum_{j=1}^{3}\omega_{\text{color},j}Corre(H^i_{\text{color},j},H^{i'}_{\text{color},j})+\sum_{j=1}^{3}\omega_{\text{dLBP},j}Corre(H^i_{\text{dLBP},j},H^{i'}_{\text{dLBP},j}) \quad (10)$$

where $\omega_{\text{color},j}$, $\omega_{\text{dLBP},j}$ represent the weight coefficients of color correlation and dLBP correlation in region j respectively, satisfying $\sum_{j=1}^{3}\omega_{\text{color},j}+\sum_{j=1}^{3}\omega_{\text{dLBP},j}=1$.

The Bhattacharyya distance[10] is used to measure the correlation of histograms H_1 and H_2. Then $Dis(m_i, m_{i'})$ is compared with a set threshold T_{dis}. If $Dis(m_i,m_{i'})$ is less than T_{dis}, m_i, $m_{i'}$ are collected into the same class; otherwise, classified into different classes.

4.1.2 Primary appearance model class

All the appearance models are classified into N classes, denoted as MC_i. For MC_i, the amount of its inner models is S_i. Considering that noise blobs will not exceed 10% of total blobs in most cases, primary appearance model class is the one whose inner models are more than 10% of total models.

4.1.3 Primary appearance model

For each primary appearance model class, it has a primary appearance model, which represents the appearance feature of the model class. Primary appearance model comes from the arithmetic operations of all the inner models. As each appearance model has six histograms (three color histograms and three texture histograms), the primary appearance model is co-determined by the arithmetic operations of these six histograms. Each of the six histograms computes an arithmetic mean histogram, described quantitatively in Eq. (11).

$$\overline{v}(b)=\text{round}(\sum_{j=1}^{S_i}v^j(b)/S_i) \quad (11)$$

where $v^j(b)$ denotes the value of bin b in the corresponding histogram H_j, and S_i the amount of inner models. Then all the six mean histograms are concatenated and a primary appearance model

is constructed, which is denoted as m^P.

4.2 Person re-identification

Relying on the state-of-the-art tracking techniques based on spatial-temporal relations[23], the blobs in single camera could be classified into different groups. Each group corresponds to a unique person. According to temporal relation, an individual's blobs can be arranged as a sequence in time order. This sequence includes all the appearance information of an individual under all visual angles when one passes through a camera. Accordingly, each camera will have several blob sequences, and the amount of sequences is equal to the amount of individuals "seen" by the camera. Each blob sequence consists of some blobs and each blob has an appearance model. All the appearance models in camera C can be arrayed in a matrix $\boldsymbol{M}^C$

$$\boldsymbol{M}^C=\begin{bmatrix} m_{11}^C & m_{12}^C & \cdots & m_{1l_1}^C & \cdots & 0 \\ m_{21}^C & m_{22}^C & \cdots & m_{2l_2}^C & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ m_{i1}^C & m_{i2}^C & \cdots & m_{il_i}^C & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ m_{n1}^C & m_{n2}^C & \cdots & m_{nl_n}^C & \cdots & m_{nl_{\max}}^C \end{bmatrix} \tag{12}$$

Row i lists the blob sequence of person i composed of l_i blobs. m_{ij}^C stands for the appearance model of blob j in the blob sequence of person i. The column amount of $\boldsymbol{M}^C$ is equal to the length of the largest blob sequence. Those smaller rows are filled with zero.

The statistical geometric parameters of target person are initially computed from the gallery blobs. The sequence in matrix $\boldsymbol{M}^C$ which is similar to the sequence of target person will be geometrically eligible.

The geometrically eligible blob sequences are further tested by the primary appearance models denoted as $MC^P=\{MC_k^P MC_k^P=(m_k^P, S_k^P)\}$ in this paper, where m_k^P represents the primary appearance model of class MC_k^P, and S_k^P the amount of its inner models.

In order to determine whether m_{ij}^c in matrix $\boldsymbol{M}^C$ can be gathered into MC_k^P, sign function D_{sign} is designed as

$$D_{\text{sign}}(MC_k^P,\ m_{ij}^C)=\begin{cases}1 & Dis(m_k^P, m_{ij}^C)\leqslant T_{\text{dis}} \\ 0 & \text{Others}\end{cases} \tag{13}$$

$D_{\text{sign}}(MC_k^P, m_{ij}^C)=1$ indicates that m_{ij}^C will be gathered into MC_k^P. In this paper, if m_{ij}^C can be gathered into MC_k^P, it can be labeled as MC_k^P. On basis of sign function D_{sign}, the class label number of appearance models in matrix $\boldsymbol{M}^C$ can be determined through a class label determination matrix $\boldsymbol{DM}(MC^P,\ \boldsymbol{M}^C)$.

$$\boldsymbol{DM}(MC^P,\ \boldsymbol{M}^C)=\sum_{k=1}^{N} D_{\text{sign}}(MC_k^P,\ \boldsymbol{M}^C)=$$

$$\begin{bmatrix} \sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{11}^C) & \sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{12}^C) & \cdots & \sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{1l_1}^C) & \cdots & 0 \\ \sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{21}^C) & \sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{22}^C) & \cdots & \sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{2l_2}^C) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{i1}^C) & \sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{i2}^C) & \cdots & \sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{il_i}^C) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{n1}^C) & \sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{n2}^C) & \cdots & \sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{nl_n}^C) & \cdots & 0 \end{bmatrix} \tag{14}$$

The class label determination matrix reads how many class labels each blob has when matched with primary appearance model classes MC^P. The accumulation of row i, denoted as LS_i and calculated in Eq. (15), defines the label scale of blob sequence of person i when matched with

the primary appearance models.

$$LS_i = \sum_{j=1}^{l_i} D_{\text{sign}}(MC^P, m_{ij}^C) = \sum_{j=1}^{l_i}\sum_{k=1}^{N} D_{\text{sign}}(MC_k^P, m_{ij}^C) \quad (15)$$

where l_i is the length of blob sequence of person i. Sometimes, if m_{ij}^C has below-threshold distances with several different primary appearance models synchronously, it can be classified into several classes. Hence, the blob will have several different labels. In order to sign if a blob has at least one label, sign function lb_{sign} is introduced.

$$lb_{\text{sign}}(blob_{ij}^C) = \begin{cases} 1 & \sum_{k=1}^{N} D_{\text{sign}}(MC_k^F, m_{ij}^C) \geqslant 1 \\ 0 & \text{Others} \end{cases} \quad (16)$$

where $lb_{\text{sign}}(blob_{ij}^C) = 1$ indicates that $blob_{ij}^C$ has at least one label and can be gathered into at least one primary appearance model class. Then $blob_{ij}^C$ is marked as a labeled blob. The number of labeled blobs in blob sequence of person i is denoted as LBN_i.

$$LBN_i = \sum_{j=1}^{l_i} lb_{\text{sign}}(blob_{ij}^C) \quad (17)$$

A blob sequence with larger LBN means that more blobs in the sequence will be gathered into primary appearance model classes, but it does not mean it is more likely to be the target sequence. In fact, although those sequences containing more blobs tend to have larger $LBNs$, they are also likely to have more unlabeled blobs that do not belong to any primary appearance model class. In light of this, an appearance based blob sequence sign function ES_{sign} is designed to determine if a sequence is an appearance based blob sequence.

$$ES_{\text{sign}}(LBN_i) = \begin{cases} 1 & LBN_i / l_i \geqslant 30\% \\ 0 & \text{Others} \end{cases} \quad (18)$$

where $ES_{\text{sign}}(LBN_i) = 1$ indicates that the blob sequence of person i is an appearance eligible blob sequence.

The appearance eligible blob sequence of person t in matrix $\boldsymbol{M}^C$, which has the maximum label scale, is re-identified as the target sequence, and person t is re-identified as the target person, namely

$$t = \arg\max\{LS_i \times ES_{\text{sign}}(LBN_i)\} \quad (19)$$

4.3 Update of primary appearance models

The blobs of target person t captured from the newest camera have the newer information than those from foregoing cameras. Primary appearance models should be updated to ensure the accuracy of continuous disjoint tracking.

For a blob in sequence of target person t in new camera, if its minimum distance with old primary appearance models is shorter than the threshold, its appearance model will be collected into the closest primary appearance model class $MC_{i_{\min}}^P$. Otherwise, it will not be collected into any existed class but generate a new appearance model class itself. The new appearance model class is labeled as M_i^{new} and its inner model number is $S_i^{\text{new}}(i = 1, 2, \cdots, N^{\text{new}})$.

The appearance models in old primary model classes and new model classes renew the primary appearance models in the following sign functions

$$\text{sign}(M_i^P) = \begin{cases} 1 & S_i^P \Big/ \left(\sum_{i'=1}^{N} S_{i'}^P + \sum_{i''=1}^{N^{\text{new}}} S_{i''}^{\text{new}}\right) \geqslant 10\% \\ 0 & \text{Others} \end{cases} \quad (20)$$

$$\text{sign}(M_i^{\text{new}}) = \begin{cases} 1 & S_i^{\text{new}} \Big/ \left(\sum_{i'=1}^{N} S_{i'}^P + \sum_{i''=1}^{N^{\text{new}}} S_{i''}^{\text{new}}\right) \geqslant 10\% \\ 0 & \text{Others} \end{cases} \quad (21)$$

where $\text{sign}(M_i^P) = 1$ ($\text{sign}(M_i^{\text{new}}) = 1$) indicates that M_i^P (M_i^{new}) will be the new primary appearance model class. It will compute a new primary appearance model for the next camera.

5 Experimental Results

5.1 Experimental setup

The experiments are successively conducted in both well-known benchmark VIPeR dataset and real complex scenarios. Firstly, this paper utilizes the images in VIPeR to test the effectiveness of human re-identification based on the improved appearance feature descriptor. The images in the dataset are randomly splitted into two sets:

Camera A and Camera B. It is the most challenging dataset currently available for single-shot pedestrian re-identification. Secondly, the effectiveness of human tracking is tested in real complex scenarios composed of seven non-overlapping views, which may witness many vast angle changes and illumination alters. The scenarios reflect closer to real-life and reveal the toughness of disjoint video surveillance. Each view in complex scenarios captures many continuous images and enables a multi-shot analysis.

5.2 Experiments in VIPeR dataset

In the experiments of this paper dataset is splitted into 10 random sets and the average of the 10 experimental results are compared with those of other two excellent benchmark methods, namely Gray's ELF 200[22] and Chen's Adaboost classifier based on multiple features[23]. The 10 experimental results are concluded from 10 random sets of 200 pedestrians. The cumulative matching characteristic (CMC) curves are depicted in Fig. 2.

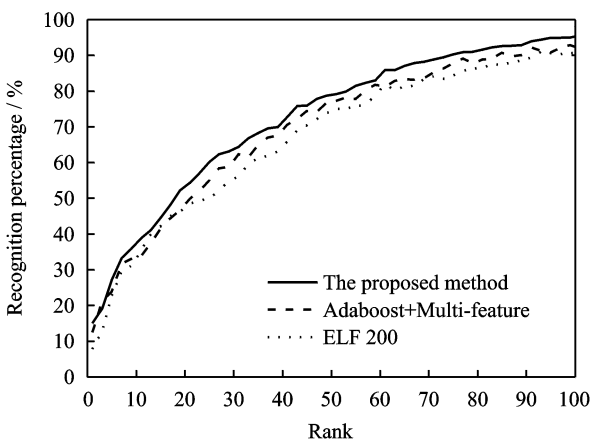

Fig. 2 CMC comparison of performance on VIPeR dataset

Fig. 2 demonstrates the proposed feature representation has the stronger discriminative power than other two state-of-the-art methods. It declares an excellent performance of single-shot people re-identification based on the improved local uniformly-partitioned HSV color descriptor and the improved dLBP texture descriptor.

Further experiments conducted on VIPeR dataset test the computation time of the proposed appearance features. The result is compared with other two human tracking approaches, namely, Chen's[23] and Hyun-Uk's[2]. All the normalized 632 images in Camera A are chosen to test the computation time of each approach. The comparison of computation time is illustrated in Fig. 3. It infers that the proposed method attracts appearance features faster than the other two excellent algorithms. The real-time feature extraction underpins and ensures the multi-shot human tracking in term of computational speed.

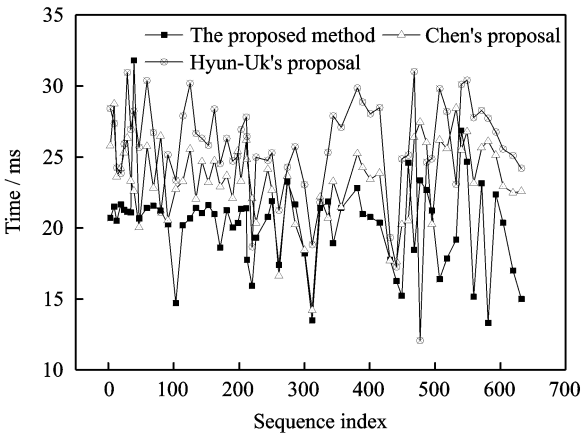

Fig. 3 Comparison of computation time for feature extraction

5.3 Experiments in complex scenarios

The complex scenario is chosen in the experiments. It aims to verify the continuous tracking across non-overlapping cameras. In the beginning of the track, the gallery group needs to be first initialized.

5.3.1 Gallery group initialization

The gallery group initialization is to choose initial source images to train the parameters of the target person. Here three experiments are conducted when initializing the gallery group. The first two use the images captured from single shooting angle, while the third one uses the images captured from multiple visual angles in the initialization. The three experiments intend to prove the importance of collecting feature information from multiple visual angles.

In the first two experiments, Camera 2 is the gallery camera and its first 168 images complete

the initialization of the gallery group with only one visual angle.

The first experiment selects Camera 3 as the probe camera, which has a similar visual angle to the gallery Camera 2. Persons 1, 2, 3 appear in Camera 3, whose blob sequences' statistical geometric parameters are listed in Table 1. The similarity of sequence of Person 1 with the gallery group is 48.8%, less than the threshold 75%, indicating that Person 1 differs from the gallery target in term of geometric features and it will be deleted. The label amount of each frame in the sequences of remaining Person 2 and Person 3 is illustrated in Fig. 4(a), where Person 2 has a larger label scale (the sum of the label amount of each frame) and its LBN is also manifestly larger than 30% of total sequence. Therefore, Person 2 is determined as the target person.

Table 1 Geometric parameters of blob sequences in Camera 3

Blob sequence	Mean/cm	Standard deviation/cm	Similarity/%
Person 1	181.9	4.3	48.8
Person 2	172.2	3	89.8
Person 3	173.1	2.8	83.2

Camera 6 is chosen as the probe camera in the second experiment, which monitors the same three persons as Camera 3 and has vast angle difference with the gallery Camera 2. Fig. 4(b) shows that when Person 2 and Person 3 come into Camera 6, they almost have no label scale. No one can be truly tracked. It means tracking failure appears in Camera 6.

In the third experiment Camera 4 is regarded as the gallery camera. It initializes the gallery group in multiple shooting angles using the first 245 images. Fig. 5 shows the FOV in Camera 4, displaying a wide view field. Fig. 5 also lists four primary appearance models of the target Person 2.

Just like the second experiment, Camera 6 is also selected as the probe camera in the third experiment. In contrast, the label scales in the third experiment (Fig. 6) dramatically outperform those of the second in Fig. 4(b).

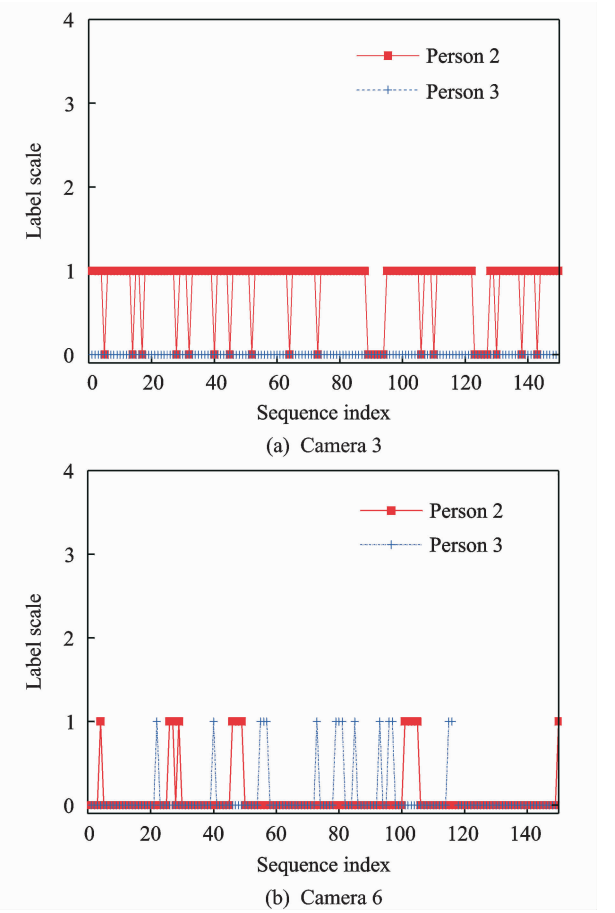

Fig. 4 Label scales of Persons 2, 3 in Cameras 3 and 6 when Camera 2 initializes the gallery group

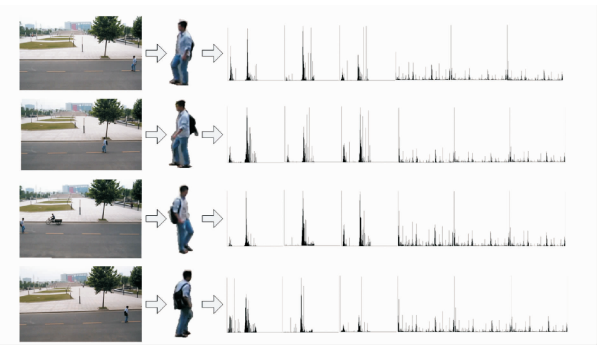

Fig. 5 FOV of Camera 4 with multiple visual angles and primary appearance models of Person 2

5.3.2 Continuous tracking

Another group of experiments is conducted in six non-overlapping cameras to validate continuous human tracking across non-overlapping cameras. Fig. 7 lists six FOVs in these experimental cameras. Each camera shoots in a unique angle and has different illumination conditions. Fig. 7 demonstrates the effectiveness of human tracking across six non-overlapping cameras.

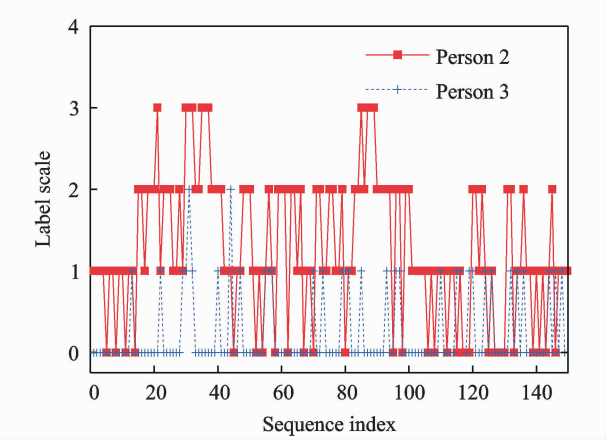

Fig. 6 Label scales of Persons 2, 3 in Camera 6 when Camera 4 initializes the gallery group

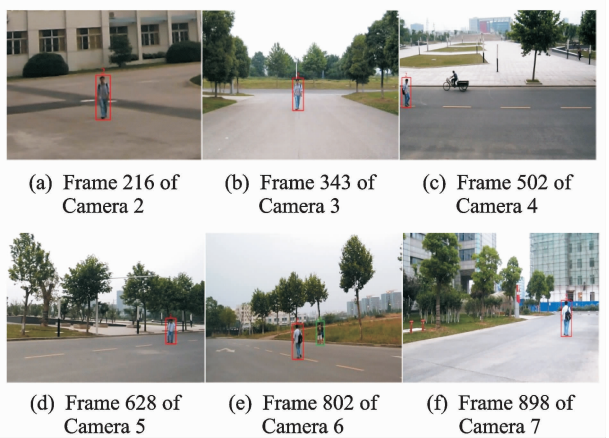

Fig. 7 Human tracking across six disjoint cameras

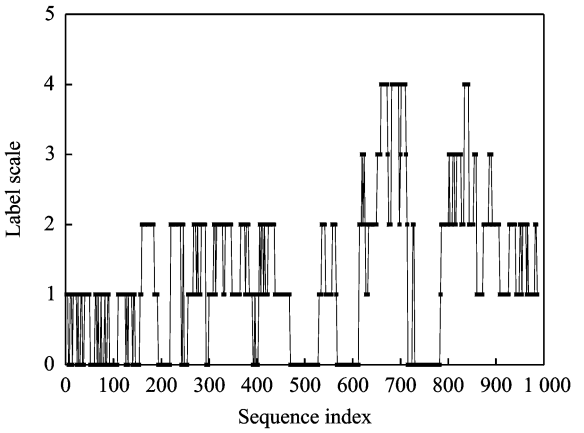

Fig. 8 Label scale of each frame from Camera 2 to Camera 7

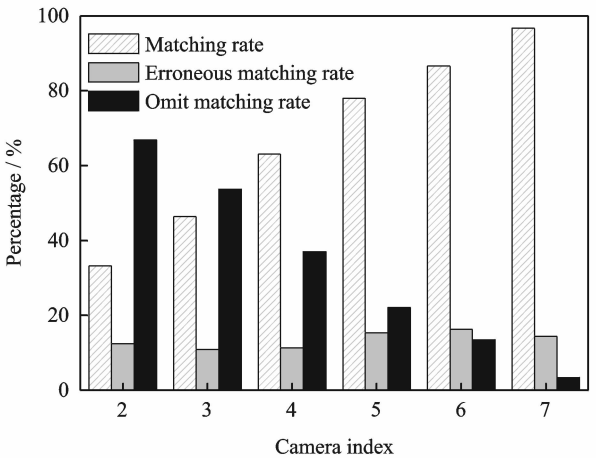

Fig. 9 Performance parameters of human tracking from Camera 2 to Camera 7

Fig. 8 shows the label amount of each frame from Camera 2 to Camera 7. The frames in Fig. 8 are arranged according to the time order. The label amount of each frame tends to increase from 0 to 4, which means more and more appearance features have been captured and stored in the gallery group.

Fig. 9 shows the performance parameters from Camera 2 to Camera 7, where matching rate represents the ratio of matched frames to all the analyzed images, and erroneous matching rate means the ratio of incorrectly matched frames to all the matched frames. The accurate matching rate increases over time, which infers that human tracking can accurately track the target people when one appears under different visual angles as well as in different view conditions. The up-and-up performance implys that the generalization capability generalizes across different cameras over time. The erroneous matching rate almost remains unchanged, and even gets worse over time. It is mainly due to the fact that incorrect blob is wrongly matched with primary appearance models and the increase of primary appearance models raises the risk of erroneously re-identifying a wrong blob. However, as the erroneous matching rate is just around 10%, it is still acceptable.

The experimental results in complex scenarios prove the feasibility of this adaptive human tracking mechanism across non-overlapping cameras. In the mechanism, the gallery group initialization is important. When the gallery group is initialized at multiple angles, it is more likely to successfully track the person in the probe camera since more appearance information is captured under different visual angles. Fig. 8 depicts the in-

crease of label amount of each frame over time. Essentially, it proves that the re-identification ability improves when more primary appearance models are stored in the system. The update of gallery group renews the primary appearance models camera by camera, thus establishing an adaptive human tracking mechanism in non-overlapping cameras.

6 Conclusions

An adaptive human tracking approach based on primary appearance model and statistical geometric features is proposed to track human across disjoint cameras in depression angles. All the extracted features manage to keep robust to illumination variations, foreground errors, as well as visual angle changes. The local uniformly-partitioned HSV color features are extracted in real-time. The combination of appearance and statistical geometric features produces a discriminative and robust feature representation. The human tracking mechanism is the main contribution. It uses both retrospective and on-the-fly information to collect the maximum appearance information captured from different visual angles. The update of primary appearance models enables the human tracking mechanism to renew adaptively. In this adaptive mechanism, the later camera will gain a higher hit rate and its generalization capability will become better. The experiments conducted in benchmark dataset show the excellent accuracy and real-time extraction of feature representation. The experiments conducted in complex scenario prove the good generalization capability of the proposed mechanism and also show good performance in resisting inter-camera and intra-camera variations.

Acknowledgements

This work was funded by the Natural Science Foundation of Jiangsu Province (No. BK2012389), the National Natural Science Foundation of China (Nos. 71303110, 91024024), and the Foundation of Graduate Innovation Center in NUAA (Nos. kfjj201471, kfjj201473).

References:

[1] Doretto G, SebastianT, Tu P, et al. Appearance-based person re-identification in camera networks: problem overview and current approaches[J]. J Ambient Intelligence and Humanized Computing, 2011 (2):127-151.

[2] Chae Hyun-Uk, Jo Kang-Hyun. Appearance feature based human correspondence under non-overlapping views[C]// Proceedings of 5th International Conference on Intelligent Computing, Emerging Intelligent Computing Technology and Applications. Ulsan: Springer, 2009:635-644.

[3] Zeng Fanxiang, Liu Xuan, Huang Zhitong, et al. Robust and efficient visual tracking under illumination changes based on maximum color difference histogram and min-max-ratio metric[J]. J Electron Imaging, 2013,22(4):043022.

[4] Lee Seok-Han. Real-time camera tracking using a particle filter combined with unscented Kalman filters [J]. J Electron Imaging, 2014,23(1):013029.

[5] Wu Yiquan, Zhu Li, Hao Yabing, et al. Edge detection of river in SAR image based on contourlet modulus maxima and improved mathematical morphology [J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2014,31(5):478-483.

[6] Madden C, Piccardi M. A framework for track matching across non-overlapping cameras using robust shape and appearance features[C]// Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance. London: IEEE, 2007:188-193.

[7] Zhang Chao, Wang Daobo, Farooq M. Real time tracking for fast moving object on complex background[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2010 (4):321-325.

[8] Takayuki Hori, Jun Ohya, Jun Kurumisawa. Identifying a walking human by a tensor decomposition based approach and tracking the human across discontinuous fields of views of multiple cameras[C]// Proceedings of Conference on Computational Imaging VIII. San Jose: SPIE, 2010:75330.

[9] LinYu-Chih , Yang Bing-Shiang, Lin Yu-Tzu, et al. Human recognition based on kinematics and kinetics of gait[J]. J Medical and Biological Engineering, 2010(31):255-263.

[10] Trevor Montcalm, Bubaker Boufama. Object inter-camera tracking with non-overlapping views: A new dynamic approach[C]// Proceedings of 2010 Canadian Conference Computer and Robot Vision. Ottawa:

IEEE, 2010:354-361.

[11] Prosser B, Zheng W, Gong S, et al. Person re-identification by support vector ranking[C]// Proceedings of British Machine Vision Conference. [S. l.]: BMVA Press, 2010:21.1-21.11.

[12] Bazzani L, Cristani M, Perina A, et al. Multiple-shot person re-identification by chromatic and epitomic analyses[J]. Pattern Recognition Letters, 2012 (33):898-903.

[13] Zheng W, Gong S, Xiang T. Person re-identification by probabilistic relative distance comparison[C]// Proceedings of IEEE Conference Computer Vision and Pattern Recognition. Colorado: IEEE, 2011: 649-656.

[14] Avraham T, Gurvich I, Lindenbaum M, et al. Learning implicit transfer for person re-identification [C]// Proceedings of European Conference on Computer Vision. Florence: Springer, 2012:381-390.

[15] Hirzer M, Beleznai C, Roth P, et al. Person re-identification by descriptive and discriminative classification[C]// Proceedings of 17th Scandinavian Conference on Image Analysis. Ystad: Springer, 2011:91-102.

[16] Zheng W, Gong S, Xiang T. Re-identification by relative distance comparison[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013 (35):653-668.

[17] Layne R, Hospedales T M, Gong S. Domain transfer for person re-identification[C]// Proceedings of ARTEMIS Workshop at ACM Multimedia. Barcelona: [s. n.], 2013:25-32.

[18] Yang Chenhui, Kuang Weixiang. Robust foreground detection based on improved vibe in dynamic background[J]. International Journal of Digital Content Technology and Its Applications (JDCTA), 2013 (7):754-763.

[19] Gonzalez R C, Woods R E. Digital image process [M]. 3rd Ed. Beijing: Publishing House of Electronics Industry, 2010.

[20] Jirí Trefny, Jirí Matas. Extended set of local binary patterns for rapid object detection[C]// Proceedings of the Computer Vision Winter Workshop. Nove Hrady: Czech Pattern Recognition Society, 2010: 37-43.

[21] Dai Xiaochen, Payandeh Shahram. Geometry-based object association and consistent labeling in multi-camera surveillance[J]. Emerging and Selected Topics in Circuits and Systems, IEEE Journal on, 2013, 3(2):175-184.

[22] Douglas G, Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features [C]// Proceedings of the 10th European Conference on Computer Vision. Marseille: Springer, 2008:62-275.

[23] Chen Xiaotang, Huang Kaiqi, Tan Tieniu. Object tracking across non-overlapping cameras using adaptive models[C]// Proceedings of ACCV 2012 International Workshops on Computer Vision. Daejeon: Springer, 2012:464-477.

(Executive editor: Zhang Tong)