DIRECTION-BASED STOCHASTIC MATCHING FOR PEDESTRIAN RECOGNITION IN NON-OVERLAPPING CAMERAS

Xiaotang Chen, Kaiqi Huang and Tieniu Tan

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{xtchen, kqhuang, tnt}@nlpr.ia.ac.cn

ABSTRACT

Pedestrian recognition is a challenging problem in non-overlapping multi-camera object tracking. In this paper, we present a novel approach for matching pedestrians across non-overlapping multiple cameras without the need of a training phase or spatio-temporal cues across cameras. To deal with viewpoint changes, we introduce the concept of directional angles estimated using the spatio-temporal continuity in the single camera tracking. To deal with pose changes, a stochastic matching strategy is performed, where the similarity of two blobs belonging to different viewpoints is calculated by a novel similarity measurement algorithm. The experiments are performed on different multi-view datasets. Experimental results demonstrate the effectiveness and robustness of the proposed method.

Index Terms— Pedestrian recognition, Directional cues, Stochastic matching, Non-overlapping camera views

1. INTRODUCTION

Object tracking is a problem of great interest for video surveillance. It is a challenging problem, especially when the number of cameras increases and the fields of cameras are non-overlapping. In the non-overlapping multi-camera object tracking, the space between adjacent cameras is not continuous, thus, pedestrian recognition is a critical process to keep objects continuously tracked across cameras. To solve this problem, many techniques with varying complexity have been proposed over the last few years.

Two kinds of cues are usually employed: spatio-temporal cues across cameras and appearance cues of objects. To get the spatio-temporal cues across cameras, K. Chen *et al*. [1] learns transition probability distributions. However, the transition probability distribution across cameras is not reliable, for it usually has worse performance when objects move inconstantly, stop or return while passing through the non-overlapping fields, which is common in real scenes.

For the appearance cues, methods generally use one or multiple kinds of features to represent the appearance of an object [3, 4]. However, the appearance is influenced by many factors, such as the illumination, camera properties, viewpoints, poses, and deformable properties of clothing, as shown in Figure 1. The differences in illumination can be compensated by using brightness transfer functions [5]. To match two objects with unknown viewpoint and pose, some approaches learn a similarity function [6] or a distance metric [7] based on a training procedure. However, their methods need to collect enough training samples, making them inconvenient to be used in real tracking systems.

Fig. 1. Some examples from the VIPeR dataset [2]. Each column is the same pedestrian from different viewpoints.

In this paper, we propose a direction-based stochastic matching (DSM) method to solve the problem of pedestrian recognition in non-overlapping multi-camera object tracking. Differently from previous methods, the proposed method does not require a training phase, or spatio-temporal cues across cameras. Instead, it uses the spatio-temporal cues in single camera tracking, which is much more reliable.

2. PROPOSED METHOD

The DSM method depends on directional cues to deal with changes in viewpoint and a stochastic matching strategy to compensate for small variances in pose. We assume that each object is seen from an arbitrary horizontal or nearly horizontal viewpoint. The assumption is satisfied in the general video surveillance scenes.

The flowchart of the proposed method is shown in Figure 2. Firstly, the directional angle of each object is estimated from its blob sequence. Then, the matching blobs are segmented into several patches according to the directional angles. Color-based features are extracted from each patch, and a similarity measurement is also proposed to measure the similarity of two patches. Finally, a stochastic matching strategy

which is robust to small pose changes is applied to measure the similarity of two blobs. The details of our method are described in the following sections.

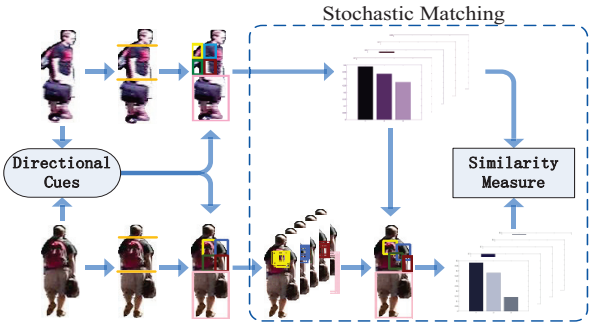


Fig. 2. The flowchart of the DSM method

2.1. Directional Cues

Instead of computing various viewpoints of cameras, we introduce the concept of directional angles to describe viewpoints. Under the assumption that each object is seen from an arbitrary horizontal or nearly horizontal viewpoint, the directional angle θ is defined according to the object's orientation, ranging from 0 to 2π. The definition is shown in Figure 3.

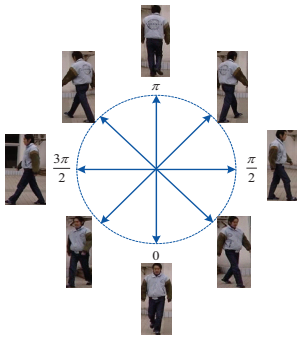
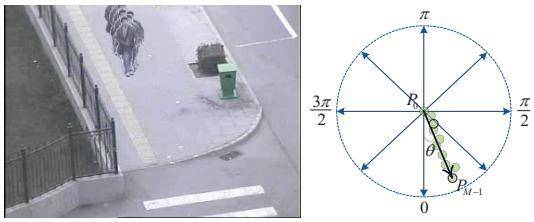


Fig. 3. The definition of the directional angle

We use the spatio-temporal cues in single camera tracking to estimate the directional angle of each object. Without loss of generality, we assume that each object moves towards its orientation. In the view of a single camera, it is easy to get a short sequence $\{P_0, P_1, \ldots, P_{M-1}\}$ of the pedestrian based on the spatio-temporal continuity, where M is the length of the sequence. The directional angle θ of the pedestrian is equal to the angle between the vector pointing to 0 and the vector $\overrightarrow{P_0P_{M-1}}$ from the earliest location P_0 to the latest location P_{M-1}. Figure 4 shows an example of estimating the directional angle.

2.2. Blob Segmentation Using Directional Cues

Given two blobs, a reference blob $blob_r$ and a candidate blob $blob_c$, both of them are divided into three regions using the method proposed in [8], corresponding to the head, the upper body and the lower body respectively. Due to different viewpoints, some visible regions in one blob become invisible in the other blob. The regions visible in both blobs are called valid regions. Using directional cues helps to find valid regions in both blobs. Generally, the appearance of the upper body is greatly influenced by viewpoint changes, while the appearance of the lower body is indistinctively varied with the viewpoint. Thus, we only deal with the valid region in the upper body instead of the whole upper body in the subsequent processes. The width of the whole upper body is defined as a unit, thus the width range of the valid region is a subset of $\{x|0{\leq}x{\leq}1\}$.



(a) (b)

Fig. 4. An example of estimating the directional angle. (a) A sequence of a pedestrian shown in an image; (b) The directional angle θ. The small dots represent the locations of the pedestrian in successive M frames.

The blob with smaller directional angle in $blob_r$ and $blob_c$ is denoted by $blob_1$, and the other is $blob_2$. The corresponding directional angle is θ_1 and θ_2 respectively. Thus, there are two conditions according to the difference between these two angles: $0{\leq}\theta_2 - \theta_1 < \pi$ and $\pi{\leq}\theta_2 - \theta_1 < 2\pi$.

Under the condition $0{\leq}\theta_2 - \theta_1 < \pi$, the width range of the valid region in the upper body is empirically calculated as follows:

$$\begin{aligned} \omega &= 0.75\pi\cos(|\theta_2 - \theta_1| + \pi) + 1.25\pi \\ \phi &= 0.125\cos(|\theta_2 - \theta_1| + \pi) + 0.125 \\ X_1 &= \{x|y = \sin(\omega(x + \phi)){\geq}0, 0{\leq}x{\leq}1\} \\ X_2 &= \{x|y = \sin(\omega(1 - x + \phi)){\geq}0, 0{\leq}x{\leq}1\} \end{aligned} \tag{1}$$

When $\pi{\leq}\theta_2 - \theta_1 < 2\pi$, let $\theta_1' = 2\theta_1 - \theta_2 + 2\pi$ and $\theta_2' = \theta_1$. The width range of the valid region in the upper body is empirically calculated as follows:

$$\begin{aligned} \omega &= 0.75\pi\cos(|\theta_2' - \theta_1'| + \pi) + 1.25\pi \\ \phi &= 0.125\cos(|\theta_2' - \theta_1'| + \pi) + 0.125 \\ X_1 &= \{x|y = \sin(\omega(1 - x + \phi)){\geq}0, 0{\leq}x{\leq}1\} \\ X_2 &= \{x|y = \sin(\omega(x + \phi)){\geq}0, 0{\leq}x{\leq}1\} \end{aligned} \tag{2}$$

where ω and ϕ controls the period and offset of the $y(x)$ curve respectively. X_1 and X_2 is the width range of the valid region of $blob_1$ and $blob_2$ respectively, equal to the intersection of $\{x|y{\geq}0\}$ and $\{0{\leq}x{\leq}1\}$. Figure 5 shows some examples of the width range of the valid region given different $\theta_2 - \theta_1$.

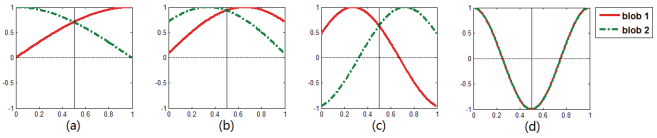
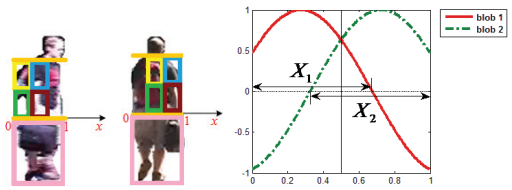


Fig. 5. Examples of the width range of valid regions. The value of $\theta_2 - \theta_1$ is 0, $\frac{\pi}{4}$, $\frac{\pi}{2}$, and π respectively from (a) to (d). The horizontal axis and vertical axis represents x and y respectively. In (a) and (b), the width ranges of valid regions for both blobs are $\{0 \leq x \leq 1\}$ according to Eq. 1, thus, the whole upper bodies for both blobs should be compared, corresponding to the fact that the observed regions of upper bodies are overlapped when the corresponding direction angles are close.

Once the width ranges of valid regions are determined, both blobs can be further divided into N_p patches, including $N_p - 1$ patches in the valid region of the upper body, and one patch of the lower body. There is a one-to-one correspondence between patches in $blob_1$ and patches in $blob_2$. Figure 6 shows an example of blob segmentation.



(a) $blob_1$ (b) $blob_2$ (c) $\theta_2 - \theta_1 = \frac{\pi}{2}$

Fig. 6. An example of blob segmentation. $blob_1$ and $blob_2$ is segmented into 5 patches in (a) and (b) respectively, according to the width ranges in (c).

2.3. Feature Extraction

Once the blobs are segmented into patches, a kind of color-based feature is extracted from each patch. In YCbCr color space, for each patch, all pixels are clustered into a fixed number of bins by using K-means. Each bin has two properties: the color and the frequency, denoted by $bin_b^i = \{C_b^i, P_b^i\}$, where i is the index for the patch and b is the index for the bin. The value of P_b^i is between 0 to 1, satisfying $\sum_{b=1}^{N_b} P_b^i = 1$. Bins in each patch are sorted in descending frequency.

We propose a novel similarity measurement method to measure the similarity between two patches. The similarity measurement is summarized in Algorithm 1, where $Dist(*,*)$ is measured by the Normalized Euclidean Distance.

2.4. Stochastic Matching

The center pixel of $Patch_c^i$ in $blob_c$ is denoted by (x_c^i, y_c^i). And the scale of the size of $Patch_c^i$ is denoted by s_c^i. Let x_c^i, y_c^i, and s_c^i follow the normal distributions, with the mean of $[x_c^i, y_c^i, 1]^T$ and the user-defined standard deviation. Then, transit each $Patch_c^i$ randomly to N_s patches, denoted by $Patch_c^{ij}$, where j is the index of transited patch.

The similarity between $blob_r$ and $blob_c$ is measured as:

$$Sim(blob_r, blob_c) = \frac{1}{N_p} \sum_{i=1}^{N_p} \max_j (Sim(Patch_r^i, Patch_c^{ij})) \quad (3)$$

where $Sim(Patch_r^i, Patch_c^{ij})$ is computed as Algorithm 1.

Algorithm 1 The Similarity Measurement

1: Extract F_r^i and F_c^i from $Patch_r^i$ and $Patch_c^i$ respectively. $F_r^i = \{\{C_{r_b}^i, P_{r_b}^i\}\}$, and $F_c^i = \{\{C_{c_b}^i, P_{c_b}^i\}\}$, where $r_b, c_b = 1, 2, \ldots N_b$
2: Initialize $Sim(Patch_r^i, Patch_c^i) = 0$
3: **for** Each $bin_{r_b}^i$ with non-zero $P_{r_b}^i$ in F_r^i **do**
4: find $bin_{c_b^*}^i$ with non-zero $P_{c_b^*}^i$ in F_c^i, satisfying: $c_b^* = \underset{c_b}{argmin} Dist(C_{r_b}^i, C_{c_b}^i)$
5: $Sim(Patch_r^i, Patch_c^i) += \min(P_{r_b}^i, P_{c_b^*}^i) \times (1 - Dist(C_{r_b}^i, C_{c_b^*}^i))$
6: $P_{r_b}^i \leftarrow P_{r_b}^i - \min(P_{r_b}^i, P_{c_b^*}^i)$; update the frequency $P_{r_b}^i$
$P_{c_b^*}^i \leftarrow P_{c_b^*}^i - \min(P_{r_b}^i, P_{c_b^*}^i)$; update the frequency $P_{c_b^*}^i$
7: **end for**

3. EXPERIMENTAL RESULTS

3.1. Pedestrian Recognition

The experiment is based on the VIPeR dataset [2]. All the pedestrians in this dataset have been labeled with angles which agree with the definition of the directional angle in Section 2.1. Some examples are shown in Figure 1.

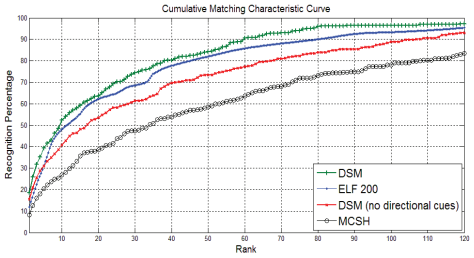


Fig. 7. CMC curves of different methods

The DSM method is compared with three different methods: ELF 200 [6], DSM without using directional cues, and a major color spectrum histogram (MCSH) which is similar to the work in [3]. The proposed DSM method uses N_b of 3 and N_s of 20. The number of patches N_p is 3 to 9, depending on the area of the valid region. For MCSH, the threshold of color distance is set to 0.06. For the DSM without using directional cues, the whole body of each pedestrian is segmented into several patches. In order to fairly compare with ELF 200, we show the average of the results on different random sets of 316 pedestrians for each of the other three methods. The

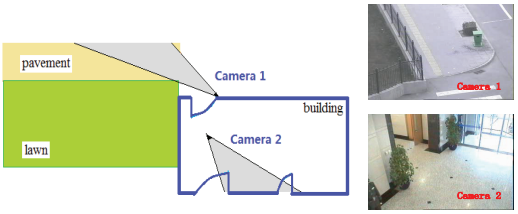
results are presented using cumulative matching characteristic (CMC) curves, shown in Figure 7. It indicates that the proposed DSM method performs best among these methods, and the rank 1 matching rate is nearly 20%. Furthermore, the DSM method outperforms the one that does not using directional cues, demonstrating the effectiveness of directional cues. Figure 8 shows some examples of the matching results using the DSM method.

(a) (b)

Fig. 8. Examples of the matching results. (a) Reference image; (b) Top 10 results (sorted left to right). The correct matches are circled by red dashed lines.

3.2. Pedestrian Tracking

The experimental setup consists of two non-overlapping cameras: one is outdoor, the other is indoor. The layout is shown in Figure 9. Both illumination and viewpoint are greatly different in the two views.



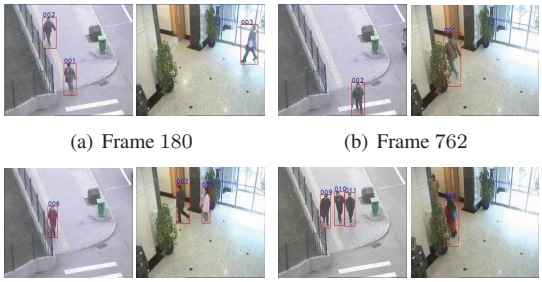
(a) (b)

Fig. 9. (a) The layout of the multi-camera system; (b) Views from Camera 1 and Camera 2.

In this experiment, the length of the sequence used to estimate the direction angle is set to 10. The video contains 37 pedestrians and 14 transfers. An object is correctly tracked if it retains a unique label through the whole video. The tracking accuracy rate is about 78.4% ($\frac{29}{37}$). Some examples of tracking results are shown in Figure 10. Both cameras overlook pedestrians from a distance, thus pedestrians are seen nearly horizontally. The proposed method can be applied to this condition, demonstrating its robustness.

4. CONCLUSIONS

In this paper, we have presented a novel solution to viewpoint invariant pedestrian recognition in non-overlapping multi-camera tracking. Without the need of a training phase or spatio-temporal cues across cameras, the proposed DSM method uses directional cues to deal with viewpoint changes, and a stochastic matching strategy to compensate for small changes in pose. Experimental results show that directional cues are efficient and robust when matching two objects with large changes in viewpoint. The tracking performance of the proposed method can be improved by incorporating color transfer functions and spatio-temporal cues across cameras.



(a) Frame 180 (b) Frame 762

(c) Frame 1124 (d) Frame 2017

Fig. 10. Examples of tracking results. Person 1 and Person 6 enter Camera 1 and Camera 2 successively. Note that all the persons retain unique labels.

Acknowledgement

This work is supported by National Natural Science Foundation of China (Grant No.60875021,60723005), NLPR 2008NLPRZY-2, National Hi-Tech Research and Development Program of China (2009AA01Z318), Key Project of Tsinghua National Laboratory for Information Science and Technology.

5. REFERENCES

[1] K. Chen, C. Lai, Y. Hung, and C. Chen, "An adaptive learning method for target tracking across multiple cameras," in *CVPR*, 2008, pp. 1–8.

[2] "Viper dataset," `http://vision.soe.ucsc.edu/?q=node/178`.

[3] E. D. Cheng and M. Piccardi, "Matching of objects moving across disjoint cameras," in *ICIP*, 2006, pp. 1769–1772.

[4] C. Kuo, C. Huang, and R. Nevatia, "Inter-camera association of multi-target tracks by on-line learned appearance affinity models," in *ECCV*, 2010, pp. 383–396.

[5] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *CVPR*, 2005, pp. 26–33.

[6] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008, pp. 262–275.

[7] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *ACCV*, 2010.

[8] M. Hu, W. Hu, and T. Tan, "Tracking people through occlusions," in *ICPR*, 2004, pp. 724 – 727.