MATCHING TRACKING SEQUENCES ACROSS WIDELY SEPARATED CAMERAS

Yinghao Cai, Kaiqi Huang and Tieniu Tan

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

ABSTRACT

In this paper, we present a new solution to the problem of matching tracking sequences across different cameras. Unlike snapshot-based appearance matching which matches objects by a single image, we focus on sequence matching to alleviate the uncertainties brought by segmentation errors and partial occlusions. By incorporating multiple snapshots of the same object, the influence of the variation is alleviated. At the training stage, given the sequence of a queried person under one camera, the appearance model is formulated by concatenating feature vectors with the majority of votes over the sequence. At the testing stage, Bayesian inference is incorporated into the identification framework to accumulate the temporal information in the sequence. Experimental results demonstrate the effectiveness of the proposed method.

***Index Terms*—** Hierarchical appearance matching, Dominant color representation, Bayesian inference

1. INTRODUCTION

Nowadays, multiple stationary cameras have been applied in surveillance applications to monitor activities over an extended area. In wide area surveillance, fusion of information from multiple cameras is required. The prerequisite for information fusion is to establish correspondences between observations across cameras. Establishing correspondences across multiple non-overlapping cameras is more challenging than single camera tracking since no spatial continuity can be exploited. Face recognition and gait recognition have been applied in recognizing a person at a distance in a non-intrusive, interaction-free manner. However, both are not very reliable in far-field scenarios. In this paper, we rely on appearance information, more specifically, color cues to identify moving objects across multiple non-overlapping cameras. We assume that the problem of single camera tracking is solved. The objective of this paper is to establish correspondences between video sequences as shown in Figure 1.

Since the fields of view of multiple cameras are non-overlapping, the appearances of moving objects under these cameras may exhibit significant differences due to variations of illumination conditions, poses and camera parameters. Therefore, the key problem of matching sequences across cameras is to build a representation for the appearance of objects which is robust to these variations. Many methods have been put forward to account for appearance changes across cameras, while the main focus was on snapshot based methods which match objects by a single image. However, establishing correspondence by a single image may bring uncertainties into the system because of imperfect segmentation and partial occlusions.

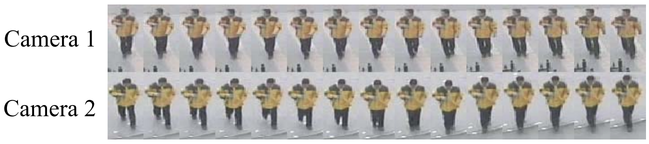


Fig. 1. Example Image Sequences

In this paper, we focus on sequence based matching. At the training stage, given the sequence of a queried person under one camera, the appearance model is formulated by concatenating feature vectors with the majority of the vote over the sequence. While at the testing stage, the matching score is calculated through sequentially accumulating the posterior likelihood by Bayesian inference. By incorporating multiple snapshots of the same object, the influence of variations is alleviated.

The paper is organized as follows. An overview of the related work is in Section 2. Section 3 describes appearance modeling and matching. Experimental results and conclusions are given in Section 4 and Section 5 respectively.

2. RELATED WORK

As we mentioned in Section 1, the appearance of moving objects can be easily influenced by variations in pose, illumination and camera parameters. To deal with illumination variations across cameras, color quantization [1, 2] and color constancy techniques have been proposed to build a representation which is robust to illumination changes. The quantization is designed so that perceptually similar colors are mapped to the same quantized value [2]. Color constancy is the ability of identifying the color of the object in spite of illumination variations and receiver characteristics [3]. Most of color constancy algorithms are under strong assumptions which can not directly be applied into applications such as ours. An alternative solution to the problem is to find a transformation matrix [4] or a mapping function [5] which map the appearance of one object to its appearance under another view.

Color spatial information is important in discriminating one object from another since objects may have similar color components with different layouts. Parametric approaches such as Gaussian mixture model are often used to estimate the distribution of a human body, [6] explicitly incorporates spatial position of the pixel into the feature representation. On the other hand, of non-parametric approaches, histogram is a widely used appearance descriptor. To incorporate spatial information into the histogram representation, an intuitive way is to divide the human body according to major color regions of the body such as head, torso and legs [7]. However, due to the partial occlusions and segmentation errors, accurately segmenting human body into subregions is not an easy task.

Our method is based on the assumption that each moving object is composed of a specific set of multiple colors arranged in a specific spatial layout. We partition the blob of moving object into regular patches for localization of color components. Each patch is represented by its dominant colors. Dominant color representation characterizes the appearance of each patch with a few “major” color which provides robustness to illumination changes to some extent.

Algorithm 1 Computation of Dominant Color Representation

```
M = 0; Initialize the number of dominant colors in the region. $I(x)$ is the RGB value at pixel $x$.
for Each pixel $x$ in the region do
    for Each dominant color $C_i$ do
        if dist$(I(x), C_i) \leq \alpha_1$ then
            $C_i \leftarrow (1-\frac{1}{W_i})C_i$ + $\frac{1}{W_i}I(x)$ ;update this dominant color
            $W_i \leftarrow W_i + 1$ ;update the frequency of this dominant color
        else
            $C_M \leftarrow I(x)$ ;assign to a new dominant color
            $M \leftarrow M + 1$
            $W_M \leftarrow 1$
        end if
    end for
end for
```

3. HIERARCHICAL APPEARANCE MATCHING

In this paper, we propose a novel method which matches objects by a set of images. Images of objects are sequentially acquired from each camera. The feature vector for moving object a from camera index c is defined as $f_a^c = \{f_{a,t}^c\}$, where t is time when the frame is obtained. Sequences of images are compared on three matching layers: matching between feature vectors of the object, matching between a class-specific appearance model and an unknown image, and matching over the sequence. In Layer 1, we obtain feature vectors for each frame individually, and assign a vote in favor of an occurring feature. The appearance model for the object is formulated by concatenating feature vectors with the majority of the vote over the sequence. In the second Layer, the similarity measure between the appearance model and an unknown image is given. In Layer 3, Bayesian inference is incorporated into the identification framework to accumulate the temporal information in the sequence. Experimental analysis demonstrates the effectiveness of the proposed hierarchical appearance matching.

3.1. Dominant Color Representation

Many methods have been proposed for appearance-based object tracking [4, 5, 6]. The purpose of target representation is to characterize the appearance of each object so as to be discriminable from other objects. Given a fixed camera, silhouettes of moving objects are obtained by background substraction techniques. We first normalize blobs of moving objects to a fixed size. Then, each blob is partitioned into regular patches according to the centroid of the blob in Figure 2(b). The size of the patch is chosen as a tradeoff between accuracy and efficiency. By employing a concept of color distance [8], we represent each patch by its dominant colors and frequencies of occurrence these colors appearing in the patch on the target. The computation of the dominant color representation is summarized in Algorithm 1.

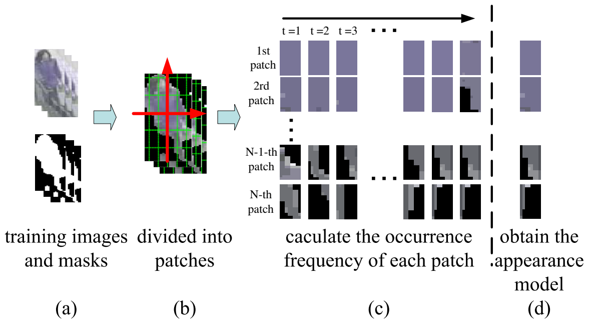
In Algorithm 1, colors within a distance threshold α_1 is regarded as a single color. The distance between two colors C_1 and C_2 is defined according to [8]:

$$\begin{aligned} dist(C_1, C_2) &= \frac{\|C_1 - C_2\|}{\|C_1\| + \|C_2\|} \\ &= \frac{\sqrt{(r_1 - r_2)^2 + (g_1 - g_2)^2 + (b_1 - b_2)^2}}{\sqrt{r_1^2 + g_1^2 + b_1^2} + \sqrt{r_2^2 + g_2^2 + b_2^2}} \end{aligned} \quad (1)$$

Similar to [8], colors in each patch are then sorted in descending frequency. Thus, the i-th patch of moving object a is represented by the first k dominant colors along with their frequency: $R_a^i = \{(C_1, W_1), ..., (C_k, W_k)\}$.

3.2. The Appearance Model

After obtaining the dominant color representation of each patch, the appearance model of one frame is represented as $f_{a,t} \equiv \{\{R_a^i, i = 1 \ldots N_a\}, t\}$, N_a is the number of the patches in the frame. Since the silhouettes of moving objects are obtained by background subtraction techniques which may result in local errors at boundaries of the silhouette. In addition, the appearance of the blob may be affected by partial occlusions. Instead of building the appearance model of the whole sequence by a single frame, we obtain the appearance model by accumulating consistent hypotheses over the



training images and masks (a) divided into patches (b) caculate the occurrence frequency of each patch (c) obtain the appearance model (d)

Fig. 2. The flowchart of the training process

sequence. Consistent hypotheses mean those color patches which repeatedly occur over the sequence. For all frames in the sequence, each patch in one frame is matched against its corresponding patch in another frame by a similarity measure:

$$Sim(R_a^i, R_{new}^i) = \min(P(R_a^i|R_{new}^i), P(R_{new}^i|R_a^i)) \quad (2)$$

where $P(R_{new}^i|R_a^i)$ is the probability of observing dominant color representation of R_{new}^i in R_a^i which is defined as:

$$P(R_{new}^i|R_a^i) = \frac{\sum_{n=1}^{M_a^i} \min\{W_{a,n}^i, \sum_{m=1}^{M_{new}^i} \delta(C_{a,n}^i, C_{new,m}^i) W_{new,m}^i\}}{|N_a^i|} \quad (3)$$

$|N_a^i|$ is the number of pixels in the i-th patch of moving object a. M_a^i and M_{new}^i are the numbers of dominant colors in each patch. $\delta(C_{a,n}^i, C_{new,m}^i)$ equals to 1 if two dominant colors are close enough according to (1). $P(R_a^i|R_{new}^i)$ can be defined similarly. For each frame in the sequence, we cast a vote for this specific dominant color representation if the similarity between two patches is larger than a threshold. We store the occurrence frequency of each feature vector. The appearance model of moving objects over the sequence is formulated by concatenating the feature vectors with the majority of the vote over the sequence. The whole process can be seen in Figure 2.

3.3. Matching Across Cameras

At the training stage, we obtain the appearance model for each object as the most frequently occurring color patches over the sequence. The sequence matching problem is now formalized as: given a class-specific appearance model $O_a^{c_1}$ under camera c_1, we need to find out observations under another camera c_2 which maximize the posterior probability $P(O_b^{c_2}|O_a^{c_1})$:

$$h = \underset{\forall O_b^{c_2}}{argmax} P(O_b^{c_2}|O_a^{c_1}) \quad (4)$$

At the testing stage, the recognition score is calculated by accumulating the posterior likelihood over the sequence. To incorporate temporal continuity, we model the sequence matching problem under the recursive Bayesian framework:

$$\begin{aligned} P(O_b^{c_2}|O_a^{c_1}) &= P(O_{b,t}^{c_2}, O_{b,1:t-1}^{c_2}|O_a^{c_1}) \\ &\propto P(O_{b,t}^{c_2}|O_a^{c_1}, O_{b,1:t-1}^{c_2}) P(O_{b,1:t-1}^{c_2}|O_a^{c_1}) \\ &\propto P(O_{b,t}^{c_2}|O_a^{c_1}) P(O_{b,t-1}^{c_2}, O_{b,1:t-2}^{c_2}|O_a^{c_1}) \quad (5) \\ &\propto P(O_{b,t}^{c_2}|O_a^{c_1}) P(O_{b,t-1}^{c_2}|O_a^{c_1}) P(O_{b,t-2}^{c_2} \\ &\quad , O_{b,1:t-3}^{c_2}|O_a^{c_1}) \propto \ldots \end{aligned}$$

where the similarity measure between the appearance model $O_a^{c_1}$ and an unknown object $O_{b,t}^{c_2}$ can be computed as:

$$P(O_{b,t}^{c_2}|O_a^{c_1}) = \frac{\sum_{i=1}^{\min(N_a, N_b)} |Sim(R_a^i, R_b^i) \geq \alpha_2|}{\min(N_a, N_b)} \quad (6)$$

By (5), we obtain the similarity score between two sequences. Compared with snapshot based matching, multiple frames and temporal continuity contained in the video facilitate human appearance matching and reduce the effect of noise to some extent.

4. EXPERIMENTAL RESULTS

The experimental setup consists of two outdoor cameras with non-overlapping fields of view. The layout is shown in Figure 3 (a). The fields of view of two cameras are about 100 meters apart. In single camera motion detection and tracking, Gaussian Mixture Model(GMM) and Kalman filter are applied, respectively. Some sample images can be seen from Figure 4.

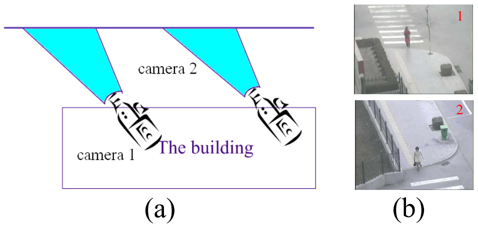


Fig. 3. (a) The layout of the camera system, (b) Views from two widely separated cameras.

We evaluate the effectiveness of the proposed method on a dataset of 42 people. In computing the dominant color representation of each patch, the color distance parameter α_1 is set to 0.01 and the similarity threshold α_2 is set to 0.9. There are mainly 2-3 dominant colors in each patch. Figure 5 shows an example of using dominant colors to represent an image. We can see from Figure 5 (c) that the color of the image is well preserved. Figure 6 shows the similarity matrix which records the similarity between video sequences under two cameras. Figure 7 (a) shows our rank matching performance. Rank i ($i = 1...10$) performance is the rate that the correct person is in the top i of the retrieved list. Different people with similar appearances bring uncertainties into the system which can explain the rank one accuracy of 78%. Figure 7 (b) shows that with the increase in the number of integration frames at the testing stage, the overall performance tends to become more stable.

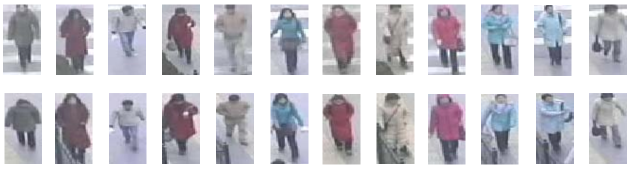
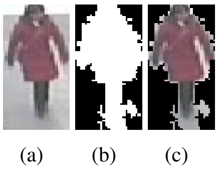


Fig. 4. Each column contains the same person under two disjoint views.

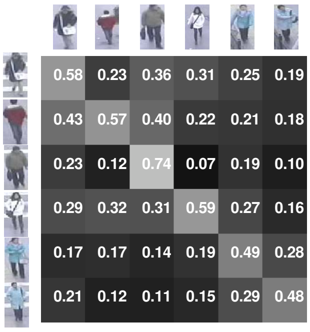


(a) (b) (c)

Fig. 5. An example of using dominant colors to represent an image. (a) Original blob, (b) Object mask, (c) Rendered image by dominant colors.

5. CONCLUSIONS

In this paper, we have presented a solution to the problem of matching sequences across different cameras. By partitioning the blobs into regular patches, the spatial information is preserved in its representation. By incorporating multiple frames of the same person, the variations caused by partial occlusion and pose changes are alleviated. The proposed method works well when cameras have similar color responses and illumination variations are decent. This work can be extended to the problem of camera handover and Content-Based Information Retrieval (CBIR) in surveillance applications.



0.58	0.23	0.36	0.31	0.25	0.19
0.43	0.57	0.40	0.22	0.21	0.18
0.23	0.12	0.74	0.07	0.19	0.10
0.29	0.32	0.31	0.59	0.27	0.16
0.17	0.17	0.14	0.19	0.49	0.28
0.21	0.12	0.11	0.15	0.29	0.48

Fig. 6. The Similarity Matrix: vertical columns show image samples under camera 1, horizontal lines are samples under camera 2. Each entry denotes the similarity between sequences under two cameras $P(O_b^{c_2}|O_a^{c_1})$.

Acknowledgement

This work is funded by research grants from the National Basic Research Program of China (2004CB318110), the National Science Foundation (60605014, 60332010, 60335010 and 2004DFA06900), and the CASIA Innovation Fund for Young Scientists. The authors also thank the anonymous reviewers for their valuable comments.

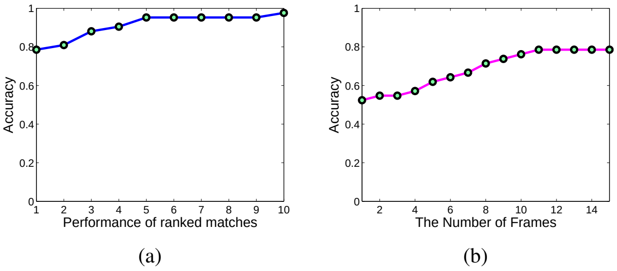


Fig. 7. (a) Rank Matching Performance. Rank i performance is the rate that the correct person is in the top i of the retrieved list. (b) Rank 1 Performance with different numbers of frames integrated at the testing stage.

6. REFERENCES

[1] Gang Wu, A. Rahimi, E.Y. Chang, Kingshy Goh, T. Tsai, Ankur Jain, and Yuan-Fang Wang, "Identifying color in motion in video sensors," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 561–569.

[2] D. Crandall and Luo Jiebo, "Robust color object detection using spatial-color joint probability functions," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004, pp. 379–385.

[3] Kobus Barnard, "Modeling scene illumination colour for computer vision and image reproduction: A survey of computational approaches," *SFU PhD depth paper*, 1998.

[4] Andrew Gilbert and Richard Bowden, "Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity.," in *ECCV (2)*, 2006, pp. 125–136.

[5] Omar Javed, Khurram Shafique, and Mubarak Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 26–33.

[6] Hanzi Wang, David Suter, Konrad Schindler, and Chunhua Shen, "Adaptive object tracking based on an effective appearance filter," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1661–1667, 2007.

[7] Ahmed M.Elgammal and Larry S.Davis, "Probabilistic framework for segmenting people under occlusion," in *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV)*, 2001, pp. 145–152.

[8] Massimo Piccardi and Eric Dahai Cheng, "Multi-frame moving object track matching based on an incremental major color spectrum histogram matching algorithm," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 19–27.