FACIAL FEATURE-INTEGRATED INTER-CAMERA HUMAN TRACKING

Young-Gun Lee, Jenq-Neng Hwang

Department of Electrical Engineering, University of Washington,
Box 352500, Seattle, WA 98195, USA
{lygstj, hwang}@uw.edu

ABSTRACT

This paper presents a new scheme to perform inter-camera human tracking in a surveillance camera network with high resolution cameras by taking advantage of all possible collected visual information. The proposed approach utilizes the tracked trajectory information of pedestrians within a camera to get accurate face positions and poses. To solve varied face pose problem under different cameras, we frontalize random posed face with a generic 2D-to-3D mapping matrix between facial feature points. Texture-based face descriptor is then exploited to extract useful features from facial components and combined with pose-invariant appearance feature, which models dominant color components in two partitioned body regions as GMM. The proposed algorithm shows promising performance by evaluating on the public benchmark Dana36 dataset.

Index Terms— human tracking, inter-camera tracking, facial feature, Dana36 dataset, disjoint camera view

1. INTRODUCTION

With incredible needs for security and safety purpose, surveillance camera systems have been popularly installed everywhere to densely monitor wide areas of city. Video analysis technologies are introduced to make an effort to reduce human operators' efforts due to the exponentially growing of deployed cameras. Especially, human tracking across cameras is one of long and tedious tasks for operators. Many researchers have suggested solutions to produce the trajectories of multi-target systematically and efficiently. A main problem in Inter-Camera Tracking (ICT) across non-overlapping cameras is appearance changes due to varied camera responses, viewpoints, illuminations and pedestrian poses under different cameras. To overcome these challenges, Chu *et al.* [1] estimate camera link model as an optimization problem to build the relationship between directly connected camera pairs based on an unsupervised manner. Chen *et al.* [2] formulate tracking problem as a global tracklet associate problem with piecewise major color spectrum histogram representation and the inter-camera similarity equalizer. To further improve the performance, context feature is integrated with appearance feature. Cai *et al.* [3] propose a relative appearance context model of the neighboring set. Chen *et al.* [4] integrate social grouping behavior of an elementary group with an online learned target-specific appearance model by using AdaBoost. However, to the best of our knowledge, there is no paper integrating facial feature into human tracking in a surveillance camera network. It was mostly impossible to get informative face region in the past due to the poor resolution of existing surveillance cameras. With increasing HD camera installation, facial information can be useful in the tracking to the practical use.

In this paper, we present a novel inter-camera human tracking scheme to achieve better performance by integrating facial feature with appearance feature. More specifically, we determine the face availability with motion information and extract robust facial feature by utilizing a face frontalization method in the random posed face. Further, we exploit pose-invariant feature to mitigate different poses across cameras. The proposed method shows the promising experimental results on the public benchmark, Dana36 dataset.

The rest of this paper is organized as follows. In Section 2, we present an overview of the overall tracking system. The algorithmic details of the proposed method are addressed in Section 3. The experimental results are shown in Section 4, followed by the conclusion in Section 5.

2. SYSTEM OVERVIEW

An overview of the proposed framework is presented in Fig. 1. Specifically, the system has Single-Camera Tracking (SCT) results as input, which contains a trajectory of a

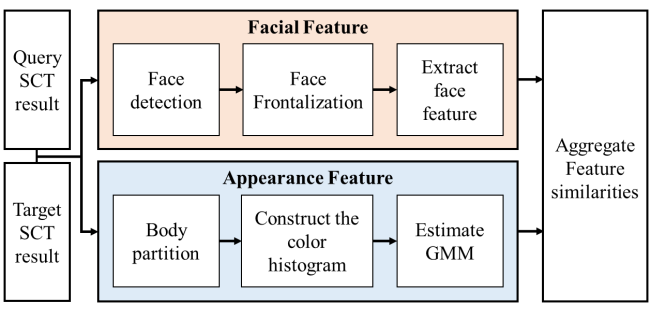


Figure 1. Facial and appearance feature integration.

tracked individual human in a camera. To evaluate whether they belong to the same identity, the proposed method extracts two features based on face and clothing. Face detector finds the face in the upper area of the bounding box based on the tracked human direction. Before extracting face feature, detected face is frontalized by utilizing a generic 2D-to-3D facial feature-points mapping matrix. For appearance feature, the pose-invariant Two-Way Gaussian Mixture Model Fitting (2WGMMF) feature [5] is exploited by GMM-represented dominant color histogram from partitioned human body (torso and legs). Two feature distances are effectively aggregated with systematically determined weights after min-max normalization.

3. PROPOSED ALGORITHMS

We propose a new ICT method to integrate facial and body appearance features into human tracking across multiple cameras.

3.1. Facial Feature

3.1.1. Face detection and feature localization

In a surveillance camera network, face area is relatively small and blurry as shown in Fig. 2(a) and 2(b) because of insufficient and unbalanced illumination. In addition, face is not available in case camera viewing angle is not appropriate. To solve this problem, we utilize motion trajectory information of the tracked person and the result of the SCT. A face detector searches face only in upper region of bounding box when people walk toward the camera because facing view of a head is usually similar as walking direction. Funnel-Structured cascade (FuSt) detection [6] is employed as the face detector. It provides a favorable solution for multi-view face detection and can detect faces with sizes larger than 20×20. The Supervised Descent Method (SDM) [7] is exploited for facial feature points detection. It localizes the 49 facial feature points as shown in Fig. 2(d).

3.1.2. Face frontalization

In ICT, face poses are not consistent due to different camera viewpoints, installation heights and varied pathways. We overcome these problems with face frontalization [8]. From the 2D coordinates of the extracted facial feature points and their corresponding 3D coordinates on the generic model, it estimates a projection matrix which represents face pose status. Then, frontalized face is synthesized by projecting extracted (query) facial feature points back onto the reference coordinate system by using the geometry of the 3D model as follows:

$$\mathbf{p}' \sim \mathbf{C}_M \mathbf{P}, \quad (1)$$

where $\mathbf{p}'$ denotes the 2D coordinate of pixels, $\mathbf{C}_M$ denotes a reference projection matrix, and $\mathbf{P}$ denotes the 3D point

Figure 2. (a) Frame7215 in CAM28 in Dana36 dataset. (b) Face detection results on the pedestrian bounding box. (c) Body partition results with ellipse shaped mask. (d) Facial feature points localization result. (e) Frontalization result.

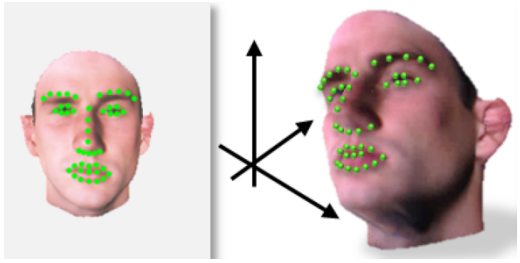


Figure 3. 49 Facial feature points on the 2D image and the corresponding 3D rendered model.

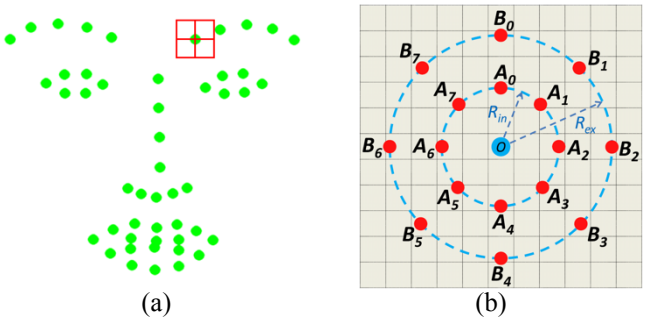
coordinates on the surface of the 3D model. Figure 3 shows the corresponding facial feature points between the 2D image and the 3D model and an example is shown in Fig. 2(e).

3.1.3. Face image descriptor

We describe face images based on 6 major facial components [9], 10 facial feature points on both eyebrows, 12 points on both eyes, 11 points on the left eye and eyebrow, 11 points on the right eye and eyebrow, 9 points on nose, and 18 points on mouth (see Fig. 4(a)). Around each facial feature point, a patch is located and further divided into 2×2 non-overlapping regions as presented in Fig. 4(a).

Each region is described with Dual-Cross Patterns (DCP) codes [9]. To quantize the texture information in each sampling direction as presented in Fig. 4(b), we assign each direction a unique decimal number as follows:

$$DCP_i = S(I_{A_i} - I_O) \times 2 + S(I_{B_i} - I_{A_i}), \quad 0 \le i \le 7, \quad (2)$$



(a) (b)

Figure 4. (a) 4 regions around a facial feature point. (b) Local sampling of Dual-Cross Patterns. Sixteen points are evenly sampled around the central pixel *O*.

where

$$S(x)=\begin{cases}1, & x\geq 0\\ 0, & x<0\end{cases}, \tag{3}$$

and I is the gray value of a pixel. The concatenated DCP feature of the 4 regions forms the description of the feature point. The similarity score between two feature vectors $\mathbf{y}^Q$ and $\mathbf{y}^T$ of each component j is measured by the cosine metric,

$$sim_{\text{facial}}(Q,T)=\sum_{j=1}^{6}\frac{\mathbf{y}_j^Q\cdot\mathbf{y}_j^T}{\|\mathbf{y}_j^Q\|_2\|\mathbf{y}_j^T\|_2}. \tag{4}$$

3.2. Appearance Feature

Although face area is available in surveillance video, human body carries more discriminative and richer information for re-identification. To mitigate the influence of illumination and camera response, color of target body region is first transferred into query's [10]. After isolating consistent clothing color regions by body partition on the ellipse shaped masked image (see Fig. 2(c)), dominant color components are modeled as a Gaussian Mixture Model (GMM) based on a 32-bin joint color histogram [5]. On both ways, query-to-target and target-to-query, one estimated GMM is fitted into another color histogram [5]. The similarity score is inversely proportional to the Negative Loglikelihood (NL):

$$d_{NL}\left(\mathbf{h}^Q,G\left(\mathbf{h}^T\right)\right)=-\ln\left(\sum_{k=1}^{K}\pi_k^T\mathcal{N}\left(\mathbf{h}^Q\mid\boldsymbol{\mu}_k^T,\boldsymbol{\Sigma}_k^T\right)\right), \tag{5}$$

where $\mathbf{h}$ denotes joint color histogram, $G(\cdot)$ denotes GMM from given color histogram, K is the number of Gaussian components, and $\mathcal{N}(\cdot)$ denotes the Gaussian distribution. π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ denote the mixing proportion, mean vector, and covariance matrix, respectively. The result from Eq. (5) is regarded as an one-way distance of a body part (torso or legs) and a small value indicates that they are likely to be the same identity. The 2WGMMF feature distance is represented as follows [5]:

$$\begin{aligned}d_{\text{2WGMMF}}(Q,T)=&\,d_{NL}\left(\mathbf{h}_{torso}^Q,G(\mathbf{h}_{torso}^T)\right)+d_{NL}\left(\mathbf{h}_{legs}^Q,G(\mathbf{h}_{legs}^T)\right)\\&+d_{NL}\left(\mathbf{h}_{torso}^T,G(\mathbf{h}_{torso}^Q)\right)+d_{NL}\left(\mathbf{h}_{legs}^T,G(\mathbf{h}_{legs}^Q)\right).\end{aligned} \tag{6}$$

3.3. Aggregation

Facial feature utilizes the cosine similarity and appearance feature exploits the negative loglikelihood. To effectively aggregate these two features, appearance feature distance, the result of Eq. (6), is transformed into similarity by inverse proportion,

$$sim_{\text{appearance}}(Q,T)=1/d_{\text{2WGMMF}}(Q,T). \tag{7}$$

Subsequently, both similarities are transformed into 0 to 1 by min-max normalization:

$$sim_i^{Norm}(Q,T)=\frac{sim_i(Q,T)-\min SIM_i}{\max SIM_i-\min SIM_i}, \tag{8}$$

where SIM denotes a set of similarity between a query and targets, $SIM=\{sim(Q,T_1),\ldots,sim(Q,T_N)\}$, max SIM and min SIM represent the smallest and the largest similarity values, respectively, and $i=\{\text{facial, appearance}\}$. Final aggregated similarity score is the summation of two feature similarity with the weighting factors:

$$sim_{\text{Final}}(Q,T)=w_{\text{facial}}\cdot sim_{\text{facial}}^{Norm}(Q,T)+w_{\text{appearance}}\cdot sim_{\text{appearance}}^{Norm}(Q,T), \tag{9}$$

where

$$w_i=\sigma_i^{\hat{sim}_i}\Big/\sum_i\sigma_i^{\hat{sim}_i}. \tag{10}$$

Discriminative ability of feature is reflected to the weights as proportion of standard deviation, which is computed in scale-normalized similarity distribution:

$$\hat{sim}_i(Q,T)=sim_i(Q,T)\big/\mu^{SIM_i}. \tag{11}$$

4. EXPERIMENTAL RESULTS

This section presents the experimental results of our approaches on the benchmark dataset, Dana36 [11], which is collected for evaluation of object matching and recognition methods in video surveillance scenarios.

4.1. Dataset and Evaluation Metric

Dana36 dataset consists of more than 23,000 images, depicting 15 persons and 9 vehicles. The dataset is acquired from 36 stationary surveillance cameras with resolutions ranging from standard VGA, 640×480, to three megapixels, 2048×1536. Among of 36 cameras, only CAM27 to 30 have 2048×1536 resolution, which is enough to detect faces in a full-frame. We exploit tracklet sets of persons, which are captured in these 4 cameras for evaluation. Figure 5 shows example frames of each camera view with the green bounding boxes representing the same identity.

The evaluation metric adopted is the Multi-Camera object Tracking Accuracy (MCTA) [12]:

$$\begin{aligned}\text{MCTA}&=Detection\times Tracking^{\text{SCT}}\times Tracking^{\text{ICT}}\\&=\left(\frac{2\times Precision\times Recall}{Precision+Recall}\right)\left(1-\frac{\sum_t mme_t^s}{\sum_t tp_t^s}\right)\left(1-\frac{\sum_t mme_t^c}{\sum_t tp_t^c}\right),\end{aligned} \tag{12}$$

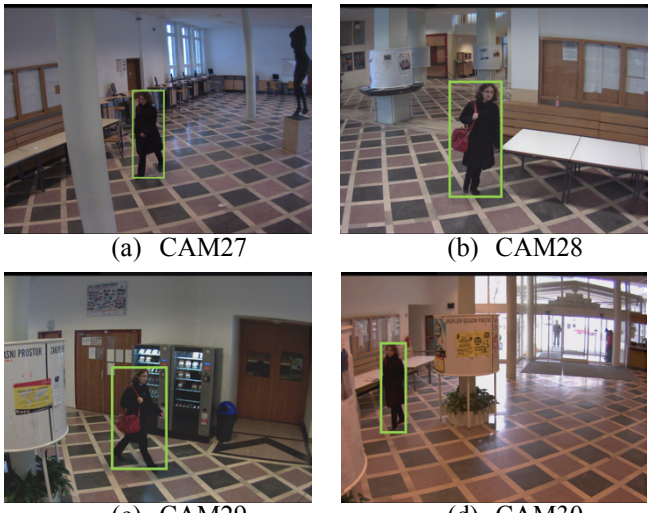
where mme_t and tp_t denote the number of mismatches and ground truth, respectively at time t. MCTA ranges from 0 to 1. The metric can be divided into three parts, Detection, SCT and ICT abilities, which separately correspond to the three brackets in Eq. (12). In this paper, the experiments focus on testing the ICT ability of the proposed method, so we use the ground truth of SCT as the inputs, resulting in *Precision* and *Recall* being 1. Thus, MCTA depends on tp_t^c and mme_t^c, which represent the number of true positive and mismatches for time t across cameras, respectively.

4.2. Tracking Accuracy

We have several experiments to compare the effectiveness of each proposed methods. Firstly, we compare the ICT performance of separate and combination of facial and body appearance features. In Table I, the proposed method achieves the best result, MCTA is 0.5785, and the appearance feature is the second one, with MCTA being 0.5651. Since face regions are not sharp in surveillance camera, the face image descriptor has difficulty in extracting discriminative texture information from them.

Secondly, effect of frontalization is also compared in ICT results. In the proposed method, frontalized faces are synthesized to mitigate the problem caused by different posed faces. In Table I, the first result obtained with extracting facial features on original face images, not on frontalized faces is the MCTA decreases about 0.05 compared to the result with frontalization in the second row.

Thirdly, face detection improvement with tracking motion information is shown in the experiment. In Table II, we show in the first column the total number of frames, which have pedestrian, in each camera from 27 to 30. The numbers on the FuSt detection [6] results are the number of detected faces based solely by FuSt algorithm. The proposed method combines FuSt method and tracked motion information. More specifically, the proposed method searches face only in case people walk toward cameras and face detector is applied on upper body region in bounding box of pedestrians. Since many frames capture the rear head or nearly rear head region, both method detects less number of faces than frame numbers. In Table II, FuSt method sometimes detects faces in another body parts in rear pedestrian appearance or background, *e.g.,* posters and snack bag, so that it has lots of False Positive (*FP*) cases. However, the proposed method has no *FP* face and more True Positive (*TP*) faces than FuSt method only. As a result, the accuracy, in terms of *Precision*, of face detection is improved with motion trajectory information.



(a) CAM27 (b) CAM28

(c) CAM29 (d) CAM30

Figure 5. Example frames of 4 cameras captured the same identity.

Table I. Experimental results of inter-camera tracking.

Method	mme^c	MCTA
Facial feature without frontalization	3759	0.2525
Facial feature with frontalization	3508	0.3025
Appearance feature (2WGMMF)	2187	0.5651
Proposed	2120	0.5785

Table II. Experimental results of face detection.

CAM# (frames)	FuSt [6]			Proposed		
	TP	***FP***	***Precision***	***TP***	***FP***	***Precision***
CAM27 (1446)	569	19	0.9677	583	0	1
CAM28 (1428)	597	11	0.9819	618	0	1
CAM29 (605)	185	102	0.6446	248	0	1
CAM30 (797)	186	220	0.4581	239	0	1
Total	1537	352	0.8137	1688	0	1

5. CONCLUSION

In this paper, we present a new tracking scheme to comprehensively integrate both face and appearance features. Based on the proposed scheme, unconstrained face and body poses are dealt effectively and systematically. By considering motion information in a camera, face detection accuracy is improved and feature becomes robust with the frontalization process. To demonstrate the efficiency of the proposed method, we perform experiments in surveillance scenarios videos. The experimental results show that the proposed scheme successfully incorporate facial feature and pose-invariant feature. Beyond providing a simple and effective means for human tracking, our work gives clue that can solve totally different problem in ICT, *e.g.*, the same people change his/her clothing before crossing cameras.

6. REFERENCES

[1] Chun-Te Chu and Jenq-Neng Hwang, "Fully unsupervised learning of camera link models for tracking humans across nonoverlapping cameras," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 979–994, 2014.

[2] Weihua Chen, Lijun Cao, Xiaotang Chen, and Kaiqi Huang, "A novel solution for multi-camera object tracking," in *IEEE International Conference on Image Processing (ICIP)*, pp. 2329–2333, 2014.

[3] Yinghao Cai and Gerard Medioni, "Exploring context information for inter-camera multiple target tracking," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 761–768, 2014.

[4] Xiaojing Chen and Bir Bhanu, "Integrating social grouping for multi-target tracking across cameras in a crf model," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[5] Young-Gun Lee, Shen-Chi Chen, Jenq-Neng Hwang, and Yi-Ping Hung, "An ensemble of invariant features for person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 470–483, 2017.

[6] Shuzhe Wu, Meina Kan, Zhenliang He, Shiguang Shan, and Xilin Chen, "Funnel-structured cascade for Multiview face detection with alignment-awareness," *Neurocomputing*, vol. 221, pp. 138–145, 2017.

[7] Xuehan Xiong and Fernando De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 532–539, 2013.

[8] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4295–4304, 2015.

[9] Changxing Ding, Jonghyun Choi, Dacheng Tao, and Larry S Davis, "Multi-directional multi-level dual-cross patterns for robust face recognition," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 518–531, 2016.

[10] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley, "Color transfer between images," *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, 2001.

[11] Vildana Suli´c Kenk, Stanislav Kova˘ci˘c, Matej Kristan, Melita Hajdinjak, Janez Per˘s, "Visual re-identification across large, distributed camera networks," *Image and Vision Computing*, vol. 34, pp. 11–26, 2015.

[12] Weihua Chen, Lijun Cao, Xiaotang Chen, and Kaiqi Huang, "An equalised global graphical model-based approach for multi-camera object tracking," *arXiv preprint arXiv*:1502.03532, 2015.