FACIAL FEATURE-INTEGRATED INTER-CAMERA HUMAN TRACKING

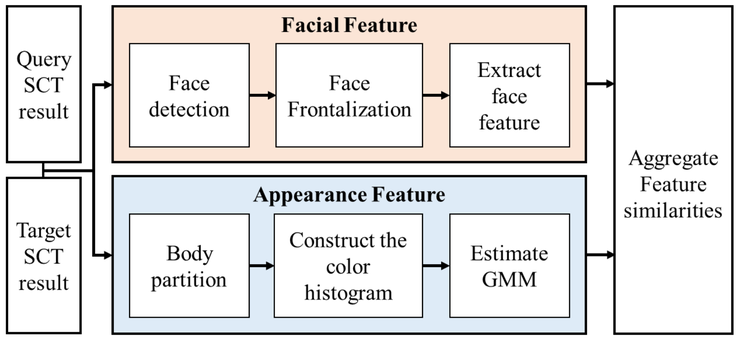
Young-Gun Lee, Jenq-Neng Hwang

University of Washington, Seattle, WA 98195

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.

Overall System



Face Detection and Feature Points Localization

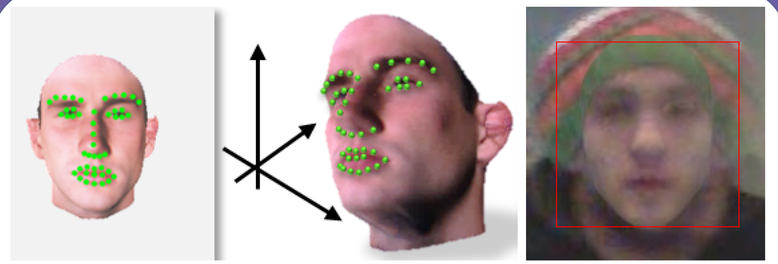
- Face area is relatively small and blurry because of insufficient and unbalanced illumination
- Motion trajectory information of tracked person and the result of SCT are utilized
- Funnel-Structured cascade detection (FuSt) [1] searches face only in upper region of bounding box when people walk toward the camera and can detect faces larger than 20×20
- Supervised Descent Method [2] localizes the 49 facial feature points

Face Frontalization

- Face poses are not consistent due to different camera viewpoints, installation heights and varied pathways
- From the 2D coordinates of the extracted facial feature points and their corresponding 3D coordinates on the generic model, it estimates a projection matrix
- Frontalized face is synthesized by projecting extracted facial feature points back [3]

$$\mathbf{p}' \sim \mathbf{C}_M \mathbf{P}$$

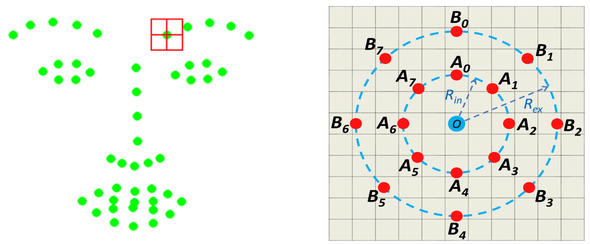
where $\mathbf{p}'$ denotes the 2D coordinate of pixels, $\mathbf{C}_M$ denotes a reference projection matrix, and $\mathbf{P}$ denotes the 3D point coordinates on the surface of the 3D model



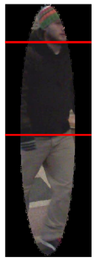
Face Image Descriptor

- 6 major facial components: 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, 18 points on mouth
- Around each facial feature point, a 2×2 non-overlapping region is located and described with Dual-Cross Patterns code [4]

$$DCP_i = S(I_{A_i} - I_O)\times 2 + S(I_{B_i} - I_{A_i}), \quad 0 \le i \le 7, \qquad \text{where} \quad S(x) = \begin{cases} 1, & x \ge 0 \\ 0, & x < 0 \end{cases}$$



Appearance Feature



- Although face area is available in surveillance video, human body carries more discriminative and richer information
- After isolating consistent by body partition on the ellipse shaped masked image, dominant color components are modeled as a Gaussian Mixture Model on a 32-bin joint color histogram [5]
- Feature distance:

$$d_{\text{2WGMMF}}(Q,T) = d_{NL}\left(\mathbf{h}^Q_{torso}, G(\mathbf{h}^T_{torso})\right) + d_{NL}\left(\mathbf{h}^Q_{legs}, G(\mathbf{h}^T_{legs})\right) + d_{NL}\left(\mathbf{h}^T_{torso}, G(\mathbf{h}^Q_{torso})\right) + d_{NL}\left(\mathbf{h}^T_{legs}, G(\mathbf{h}^Q_{legs})\right).$$

Features Aggregation

- Facial feature similarity score between query feature vector $\mathbf{y}^Q$ and target feature vector $\mathbf{y}^T$:

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

- Appearance feature similarity score: $sim_{\text{appearance}}(Q,T) = 1/d_{\text{2WGMMF}}(Q,T)$
- Final aggregated similarity score:

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

where $sim^{Norm}_i(Q,T) = \dfrac{sim_i(Q,T) - \min SIM_i}{\max SIM_i - \min SIM_i}$ and $w_i = \sigma_i^{sim_i} \Big/ \sum_i \sigma_i^{sim_i}$

- Discriminative ability of feature is reflected to the weights, w_i

Dataset and Evaluation Metric

- Dana36 dataset: 23,000 images depicting 15 persons and 9 vehicles
- Among of 36 cameras, only CAM27 to 30 have 2048×1536 resolution, which is enough to detect face in a full-frame
- Tracklet sets of persons captured in these 4 cameras are exploited for evaluation

(a) CAM27 (b) CAM28 (c) CAM29 (d) CAM30

- Evaluation metric for tracking: Multi-Camera object Tracking Accuracy (MCTA)

$$\text{MCTA} = Detection \times Tracking^{\text{SCT}} \times Tracking^{\text{ICT}} = \left(\frac{2\times Precision \times Recall}{Precision + Recall}\right)\left(1 - \frac{\sum_t mme^s_t}{\sum_t tp^s_t}\right)\left(1 - \frac{\sum_t mme^c_t}{\sum_t tp^c_t}\right)$$

- Evaluation metric for face detection: $Precision = \dfrac{TP}{TP + FP}$

Tracking Accuracy

- 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) [4]	2187	0.5651
Proposed	2120	0.5785

- Table II. Experimental results of face detection

CAM# (frames)	FuSt [1]			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

References

[1] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen, "Funnel-structured cascade for Multiview face detection with alignment-awareness," *Neurocomputing*, vol. 221, pp. 138-145, 2017.

[2] X. Xiong and F. D. Torre, "Supervised descent method and its applications to face alignment," *in Proceedings of the IEEE conference on CVPR*, pp. 532-539, 2013.

[3] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," *in Proceedings of the IEEE conference on CVPR*, pp. 4295-4304, 2015.

[4] C. Ding, J. Choi, D. Tao, and L. S. Davis, "Multi-directional multi-level dual-cross patterns for robust face recognition," *IEEE transactions on PAMI*, vol. 38, no. 3, pp. 518-531, 2016.

[5] Y.-G. Lee, S.-C. Chen, J.-N. Hwang, and Y.-P. Hung, "An ensemble of invariant features for person reidentification," *IEEE transactions on CSVT*, vol. 27, no. 3, pp. 470-483, 2017.

RESEARCH POSTER PRESENTATION DESIGN © 2015
www.PosterPresentations.com