Robust Real-Time Detection, Tracking, and Pose Estimation of Faces in Video Streams

Kohsia S. Huang and Mohan M. Trivedi

Computer Vision and Robotics Research Laboratory
University of California, San Diego
http://cvrr.ucsd.edu/

Abstract

Robust human face analysis has been recognized as a crucial part in intelligent systems. In this paper we present the development of a computational framework for robust detection, tracking, and pose estimation of faces captured by video arrays. We discuss the development of a multi-primitive skin-tone and edge-based detection module embedded in a tracking module for efficient and robust face detection and tracking. A continuous density HMM based pose estimation is developed for an accurate estimate of the face orientation motions. Experimental evaluations of these algorithms suggest the validity of the proposed framework and its computational modules.

1. Introduction

Human-computer interaction has been an active topic in the research community of computer vision and intelligent systems. These systems involve the recognition of human identities and activities in indoor, outdoor, and mobile environments, and among them face related analysis is the central focus. However, it is recognized that without an accurate, robust, and efficient face detection as the front-end module, successful face analysis cannot be realized. Robustness to background and illumination variations is known as a major challenge [5].

Figure 1 shows an intelligent environment system that captures and tracks people to automatically derive events with camera arrays. The 3D tracker runs on an omnidirectional camera array for rough locations and heights of people in a large area. A pan-tilt-zoom (PTZ) rectilinear camera then focuses on the head of a person. Within the video, the human face is detected and tracked with finer resolution. Face orientation is also estimated to select a suitable PTZ camera to capture the near frontal face for robust recognition. Note that the face detection, tracking, orientation, and recognition are video-based to accumulate and filter the image likelihoods over frames [6]. It would significantly enhance the accuracy and the robustness in real-world situations such as illumination variations, cluttered backgrounds, occlusions, noises, etc.

In this paper we focus on the algorithms of robust real-time face detection, tracking, and orientation estimation from video data. We then evaluate them using indoor, outdoor, and mobile video sequences.

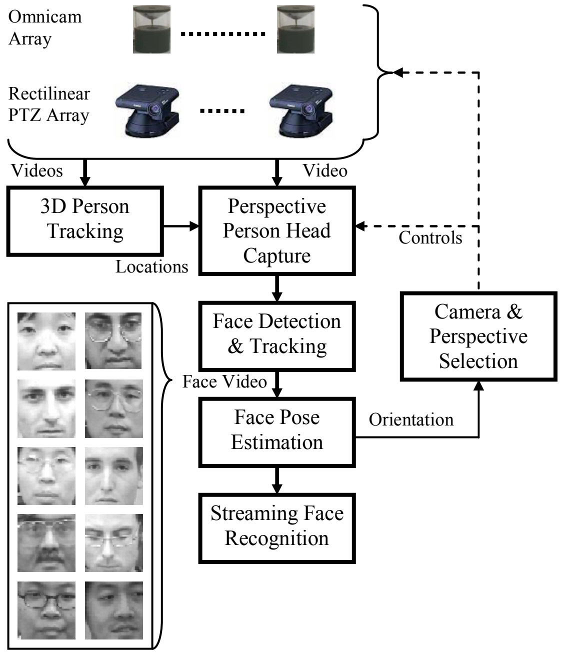

Figure 1: An integrated system for person tracking and identification. It uses video arrays for multi-person tracking and captures high resolution video using the most appropriate camera. Captured video is analyzed for person identification or verification.

2. Robust Real-Time Multi-Primitive Face Detection and Tracking

Robust face detection and tracking is crucial in the integrated face analysis performance in indoor, outdoor, and mobile environments [7]. We use skin color and elliptical edge features in this algorithm. Skin color allows rapid face candidate finding, yet it can be affected by other skin-tone objects and is sensitive to lighting spectrum and intensity changes. Elliptical edge detection is more robust in these cases, yet it needs more computation and is vulnerable to highly cluttered backgrounds. These two tend to complement each other [3]. The proposed closed-loop face detection and tracking is illustrated in Figure 2. The subject video is first subsampled to speed up processing. On the skin color track, skin blobs are detected [12] if the area is above a threshold and face-cropping windows are derived from the blob moments. On the edge track, face is detected by matching an ellipse to the face contour. Since elliptic regressions from edge pixels [4] are slow and do not always yield valid head ellipses, we use a combination of two template methods that find the best match in a set of pre-defined head ellipses to the edges [3][8]. Possible head top is first located by finding the horizontal edge links. Then an ellipse template is attached along the horizontal edge links at the top pixel of the ellipse. The matching is to find the maximum ratio $R = (1 + I_i)/(1 + I_e)$ in the ellipse set, where $I_i = (1/N_i)\sum w \cdot p$ is a weighted average of p over a ring just inside the ellipse with w=2 at the top quarter of the ring and 1 elsewhere, $I_e = (1/N_e)\sum p$ is the averaged p over a ring just outside the ellipse, and $p = |n \cdot g|$ is the absolute inner product of the normal vector on the ellipse with the image gradient at that point. This algorithm improves [8] by using the inner product p and accelerates [3] by the ellipse search scheme, thus making real-time full-frame ellipse matching possible. After this, a face-cropping window is derived for each ellipse.

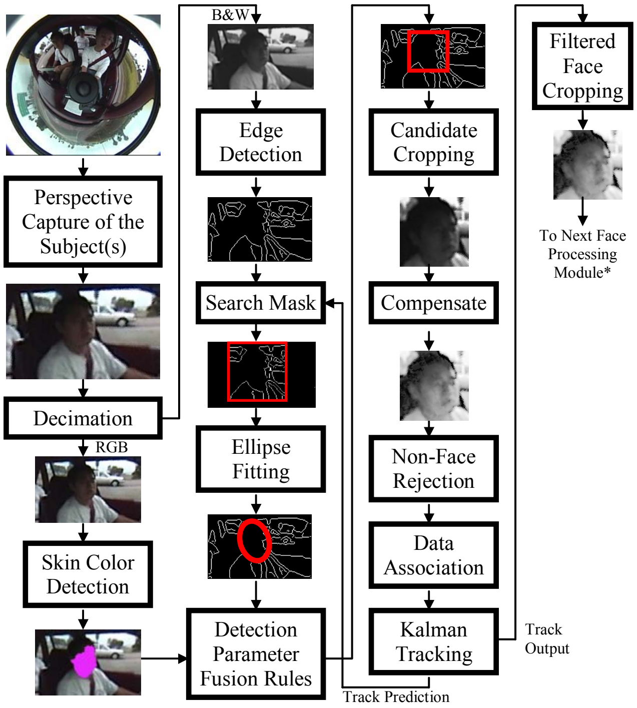

Figure 2: The integrated "closed-loop" face detection and tracking on a mobile omni-video example.

The skin-tone and elliptic face-cropping windows are then fused. For each skin window, we find a closest ellipse window and average their upper-left corner coordinates and window sizes to crop the face candidate. The weighting between them can be adjusted for best results. If there is no ellipse detected, skin windows are used solely, and vice versa for the ellipse windows.

The face candidates are scaled to 64×64 size and compensated for uneven illumination by subtracting a least-squares fitted intensity grade plane. Then they are verified by distance from feature space (DFFS) in PCA subspace [11] to reject non-face candidates. Each positive face window is associated to an existing face window track by nearest neighborhood and used to update the constant velocity Kalman filter [1] of the track. The Kalman filter interpolates detection gaps and predicts the face location in the next frame. For each track prediction, an ellipse search mask is derived for the next frame to speed up ellipse detection by minimizing the ellipse search area. A face track is initialized when a face is detected for some consecutive frames. The track is terminated if the predicted face window is non-face for some frames.

3. Robust Estimation of Face Orientation: A Multi-State Approach

Next the face orientations in the video can be estimated for active camera control and assessing the attentive direction of the person. We compare two face orientation estimation schemes as shown in Figure 3 and Figure 4. First the face frame is projected into a PCA subspace and only the first D dimensions are used since they carry most information [9]. The PCA subspace is constructed with the correlation matrix of training faces so that illumination variations can be reduced by projection vector normalization [6][2]. In Figure 3, face orientation of the frame is estimated by maximum likelihood (ML) and filtered by a Kalman filter across frames. Mean and covariance of the N Gaussian likelihood functions are estimated by Linde-Buzo-Gray vector quantization (LBG-VQ) on the training projections. Corresponding facing angle of a likelihood is found by averaging the ground truth angles of the training frames which are classified to this class by ML.

In Figure 4, we build a continuous density hidden Markov model (CDHMM) [10]. The Markov chain (state transition matrix A) is linear bi-directional with N states relating to certain facing angles in order to model a continuous face turning. The observation probability $b_j(\mathbf{x}_k)$ of state j is modeled by a mixture of M Gaussian densities, where $\mathbf{x}_k$ is the projection vector of face frame k. The state sequence $q(k)$ of the face video which relates to the face orientations can be estimated by maximum *a posteriori* (MAP) in real-time or optimally by Viterbi algorithm with minor delay caused by sequence framming.

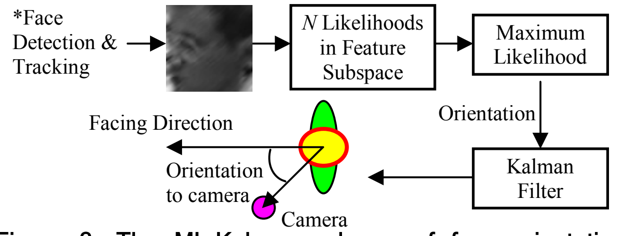

Figure 3: The ML-Kalman scheme of face orientation estimation.

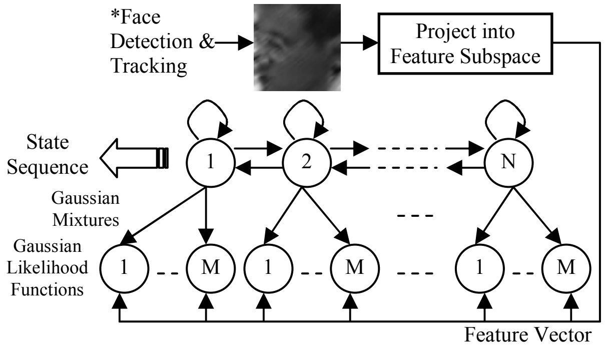

Figure 4: Face orientation estimation by CDHMM.

On CDHMM training, the initial probabilities π and the mixture coefficients C are randomly initialized. The $N\times M$ mean vectors μ's of the Gaussian densities are initialized by two rounds of LBG-VQ. First the training PCA projections are partitioned into N regions by VQ. The facing angle of a region is the averaged ground truth angle of the training video frames that fall into the region. By these angles the N regions are assigned to the N states in ascending order. Next the μ's of the M Gaussian densities in one of the N regions are found by LBG-VQ on the training vectors in that region. The covariance U's of the Gaussian densities are initialized as σI.

The next issue is to determine the facing angles of the states. The state sequence $q(k)=\left\{s_{k=1}^{T}\middle|s_k=1,\ldots,N\right\}$ of a training face video is first estimated. Given the ground truth sequence of the training video $t(k)=\left\{\theta_k\middle|k=1,\ldots,T\right\}$, we want to find the association

$$A(s)=\begin{cases}A_1\ , & s=1\\ A_2\ , & s=2\\ \vdots & \vdots\\ A_N\ , & s=N\end{cases} \tag{1}$$

by minimizing the mean squared error,

$$MMSE=\min_{A_1,A_2,\ldots,A_N}\left(\frac{1}{T}\right)\sum_{k=1}^{T}\left[A\left(q(k)\right)-t(k)\right]^2 \tag{2}$$

Then the least squares solution is proven to be

$$A_s=E_s\left[t(k)\right]=\frac{1}{T_s}\sum_{k=1}^{T_s}t(k) \tag{3}$$

for $s=1,\ldots,N$, where T_s indicates all the time indices in the training video when $q(k)=s$.

4. Experimental Evaluation and Analysis

Evaluation of the head tracking and face orientation estimation is accomplished using an extensive array of indoor, outdoor, and mobile videos, as shown in Figure 8. In these test clips, the camera and subjects are fixed, so person tracking in Figure 1 is not needed and the perspective view of a subject is manually selected. Figure 5 shows some indoor face detection and tracking results. It indicates that the multi-primitive face detection & tracking is very robust to extreme cases such as highly cluttered background, skin-tone object interference, and colored dim lighting in dark room. The standard deviation of face alignment within the 64×64 face video after Kalman tracking is approximately 8 pixels.

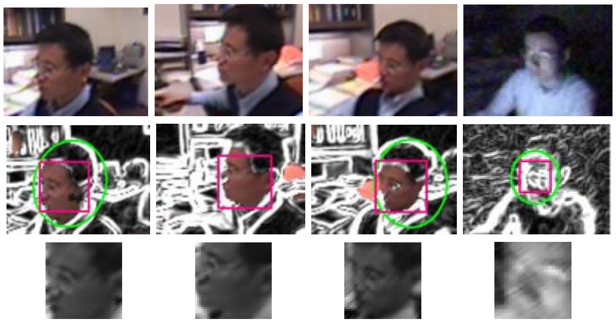

Figure 5: Some results of the multi-primitive face detection and tracking. Top row shows various backgrounds and lightings, middle row shows combined skin-color and edge detections, and lower row shows the cropped faces.

The two face orientation estimation schemes are compared using a mobile video of 2300 frames, where the ground truth facing angles are estimated manually frame by frame. The video is processed twice to extract the training and testing face videos of the same length but of different face alignments, due to current hardship of obtaining ground truth. We tried different combinations of N, M, D, and σ of the CDHMM for horizontal face orientation. The transition length *TrL* (nonzero terms from the diagonal elements in the state transition matrix A) is 2 and can be 3 to model faster face turn. After trial phase, N = 38, M = 1, D = 10, and σ = 1/2 seems to produce the minimum standard deviation of the estimates (12°) as in Figure 7. Comparing to Figure 6, the ML-Kalman estimate with N = 38 VQ steps and D = 10 is more noisy and the standard deviation (19°) is higher. Hence the CDHMM scheme is preferable. The reason that the CDHMM approach works better is that it is a *delayed decision* approach. In the ML-Kalman case, ML decision is made before Kalman filtering and blocks useful cues.

In the future we will collect more videos on diverse subjects and environments with a synchronized ground truth measuring device in order to enhance the CDHMM robustness. Vertical face orientation can also be estimated by another CDHMM. Finally, LDA subspace can be used since it is more robust to illumination variations [2].

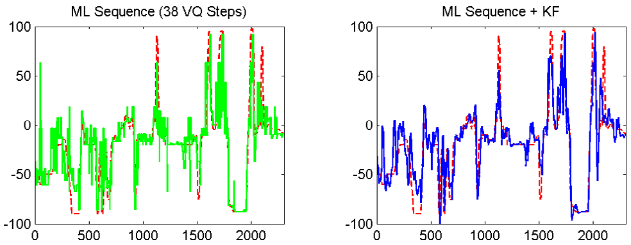

Figure 6: ML-Kalman based horizontal face orientation estimation. Left: Single-frame ML face orientation sequence. Right: Kalman filtered sequence. Solid line: Estimated face orientation; Dotted line: Ground truth value. Horizontal axis: Frame number; Vertical axis: Facing angle (-: facing right; 0: frontal; +: facing left). *N* = 38, *D* = 10.

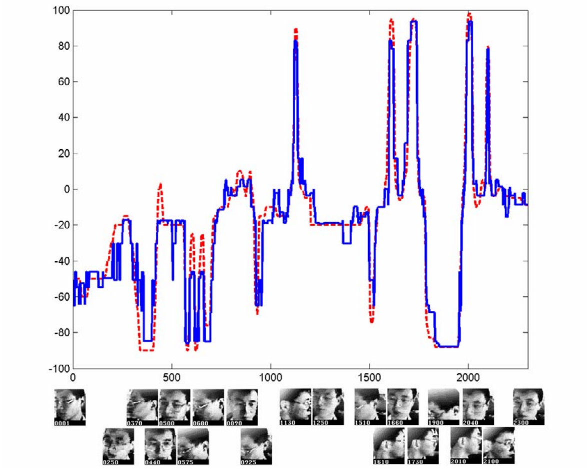

Figure 7: Horizontal face orientation estimation of the CDHMM scheme. Solid line: Estimated face orientation; Dotted line: Ground truth. CDHMM setup: *N* = 38, *M* = 1, *D* = 10, *TrL* = 2, σ = 1/2.

5. Concluding Remarks

In this paper we have presented an intelligent system for capturing humans to detect and track their faces. Real-time robust face detection and tracking is achieved by a multi-primitive closed-loop face analysis architecture. Novel algorithms to estimate face orientations using ML-Kalman filtering and multi-state CDHMM models have been evaluated using a series of experimental studies. These experiments support the basic feasibility and promise of the multi-state approach.

References

[1] Y. Bar-Shalom, X. Rong Li, T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*, John Wiley and Sons, 2001.

[2] P. Belhumeur, J. Hespanha, D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE. Trans. PAMI*, vol. 19, no. 7, pp. 711-720, Jul. 1997.

[3] S. Birchfield, "Elliptical Head Tracking Using Intensity Gradient and Color Histograms," *Proc. IEEE CVPR Conf.*, Jun. 1998.

[4] A. Fitzgibbon, M. Pilu, R. B. Fisher, "Direct Least Squares Fitting of Ellipse," *IEEE Trans. PAMI*, vol. 21, no. 5, pp. 476-480, May 1999.

[5] E. Hjelmas and B. K. Low, "Face Detection: A Survey," *Comp. Vis. Img. Understd.*, vol. 83, pp. 236-274, 2001.

[6] K. S. Huang, M. M. Trivedi, "Streaming Face Recognition using Multicamera Video Arrays," *Proc. Int'l. Conf. on Pattern Recognition*, vol. 4, pp. 213-216, Aug. 2002.

[7] K. S. Huang, M. M. Trivedi, T. Gandhi, "Driver's View and Vehicle Surround Estimation using Omnidirectional Video Stream," *Proc. IEEE Intelligent Vehicle Symp.*, pp. 444-449, Jun. 2003.

[8] A. Jacquin, A. Eleftheriadis, "Automatic Location Tracking of Faces and Facial Features in Video Sequences," *Proc. Int'l. Wksp. on Auto. Face Gesture Recog.*, June 1995.

[9] J. Ng, S. Gong, "Multi-View Face Detection and Pose Estimation Using A Composite Support Vector Machine Across the View Sphere," *Proc. Int'l. Wksp. on Recog., Ana, and Track. of Faces and Gestures in Real-Time Sys.*, pp. 14-21, 1999.

[10] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[11] M. Turk, A. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE CVPR Conf.*, pp. 586-591, Jun. 1991.

[12] J. Yang, A. Waibel, "A Real-Time Face Tracker," *Proc. WACV'96,* pp. 142-147, 1996.

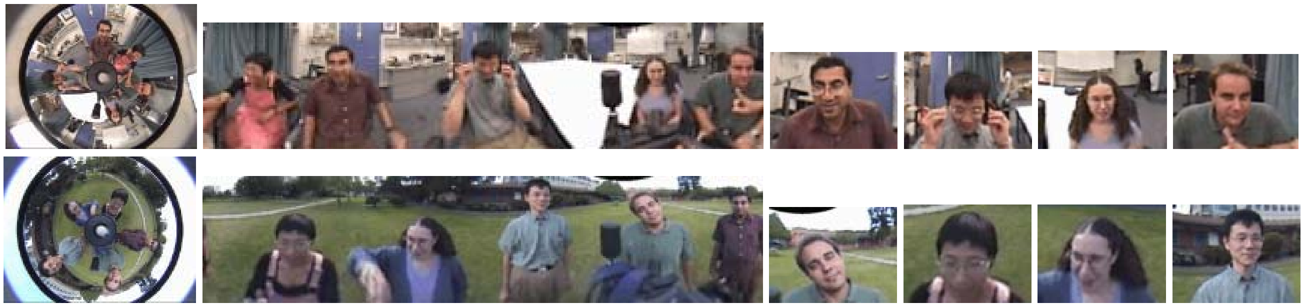

Figure 8: Sample indoor and outdoor images from the test and training video sequences for face detection, tracking, and face orientation estimation. Left to right: the source omni-videos, the unwarped panoramas, and human perspectives.