The 17th International Conference on Pattern Recognition (ICPR 2004)
Cambridge, United Kingdom, Aug. 23-26, 2004

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

Human face analysis has been recognized as a crucial part in intelligent systems. However, there are several challenges before robust and reliable face analysis systems can be deployed in real-world environments. One of the main difficulties is associated with the detection of faces with variations in illumination conditions and viewing perspectives. 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 development of a multi-primitive skin-tone and edge-based detection module integrated with a tracking module for efficient and robust detection and tracking. A multi-state continuous density Hidden Markov Model based pose estimation module is developed for providing an accurate estimate of the orientation of the face. A systematic evaluation of these algorithms and the overall framework is performed with an extensive set of experiments. Results of these experiments suggest the validity of the proposed framework and its computational modules.

1. Introduction

Human-computer interactions has been a very 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 [3]. However, it is recognized that without an accurate, robust, and efficient face detection as the front end module, successful face analysis can not be realized. Robustness with respect to background and illumination variations is recognized as a major challenge [1].

Figure 1 shows the architecture of an intelligent environment with camera arrays that capture and track people to automatically derive events in the environment. The 3D tracker operates on a broad area using an omnidirectional camera array and provides the rough locations and heights of people. These cues are used to actuate a pan-tilt-zoom (PTZ) rectilinear camera to focus on the head of a person. Within the perspective view, the human face is detected and tracked with finer resolution. The face orientation is then estimated to select a suitable camera to capture further details of the face for robust recognition. It is noted that for real-world situations, the face detection, tracking, orientation, and recognition modules need to be video-based to accumulate and interpolate the image likelihoods over time. This would significantly enhance the accuracy and the robustness to fluctuations such as illumination variations, cluttered backgrounds, occlusions, noises, etc.

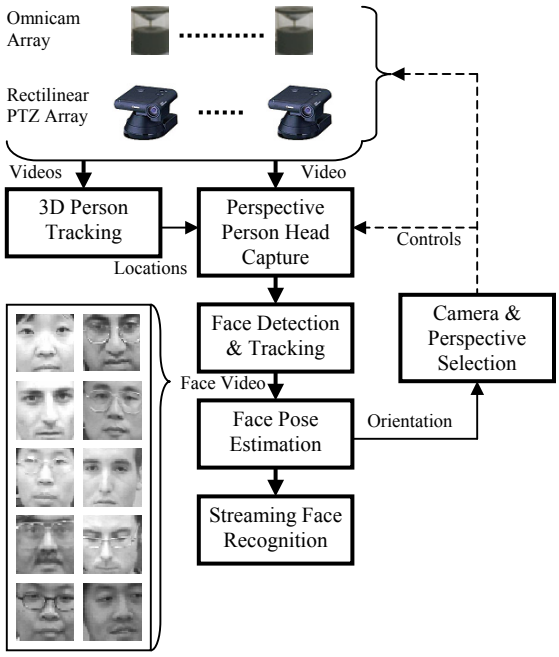

Figure 1. An integrated system for person tracking and identification. It uses video arrays for multiple person tracking and captures high resolution video using the most appropriate camera. Captured video is analyzed for person identification or verification.

Our research discussed in this paper is focused on the development of robust face detection, tracking, and face-orientation estimation algorithms from video data. These

modules are also evaluated experimentally using indoor, outdoor, and mobile video sequences. In the following sections, we will present the details of these three items.

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

Face detection and tracking play a crucial role in the integrated system performance for face orientation and recognition [1]. This module should be robust on indoor, outdoor, and mobile situations [3]. We considered skin color and elliptical edge features for the real-time face detection algorithm. Skin color allows rapid face candidate finding, yet it can be interfered by other skin-tone objects and is sensitive to illumination spectrum and intensity variations as in the mobile cases. Edge-based face detection is more robust in these cases, but it tends to err on highly cluttered backgrounds and needs more computation. These features tend to complement each other [4]. The proposed closed-loop face detection and tracking scheme is illustrated in Figure 2. The perspective view on the person subject is first generated from a camera video. Then the view is sub-sampled to speed up processing. On the skin color side, skin color blobs are detected [6] if its area is above a threshold. Then the parameters of face cropping window is evaluated from the moments of the blob. On the edge side, face is detected by matching an ellipse to the face contour. Direct ellipse fitting from edge detections such as randomized Hough transform [7] and least-squares [8] are not feasible here because the aspect ratio and pose of the detected ellipse is not constrained and improper ellipse detections for faces have to be discarded, thus disabling it from real-time processing. Other approaches match a set of pre-defined ellipses to the edge pixels [4][5]. Our method is a combination of their advantages. Possible head top is first found by merging the horizontal edge pixels if their distance is below a bound and the image intensity gradient is nearly vertical. Then a head-resembling ellipse template is attached along the horizontal edge links at the top pixel of the ellipse. The matching is done by finding the maximum ratio $R = (1+I_i)/(1+I_e)$ for all the ellipse-edge attachments, where $I_i = (1/N_i)\sum w \cdot p$ is a weighted average of p over a ring zone just inside the ellipse with higher weights w at the top portion of the zone, $I_e = (1/N_e)\sum p$ is the averaged p over a ring zone just outside the ellipse, and $p = |n \cdot g|$ is the absolute inner product of the normal vector on the ellipse with the image intensity gradient vector at that point. This inner product force the image intensity gradients to be parallel to the normal vectors on the ellipse template, thus reduces the false detections of using gradient magnitude alone as in [5]. This method also includes a measure for quicker ellipse search only on plausible points than that of extensive search as in [4]. It also makes real-time full-frame ellipse search possible.

After the skin blobs and face contour ellipses are detected, their parameters are fused to produce the face candidate cropping window. The square window is defined by the upper-left corner coordinates and the size. For each skin-tone blob window, we find a closest ellipse window of similar size and average their upper-left corner coordinates and window sizes for the face candidate cropping window. The weighting between the skin-tone blob and the ellipse is experimentally adjusted to yield the best detection accuracy. If there is no ellipse detected, skin-tone blobs are used solely, and vice versa for the ellipses.

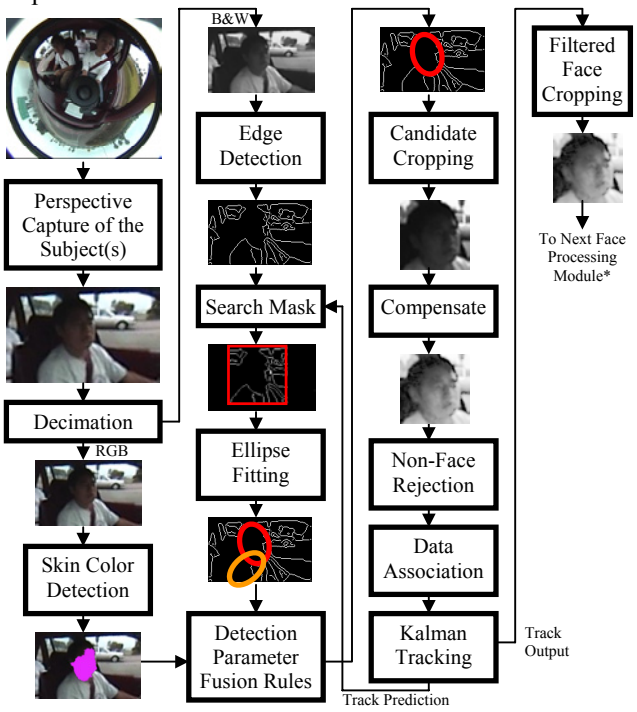

Figure 2. The computational flow chart for the integrated "closed-loop" head/face detection and tracking modules.

The detected face windows then crop the face candidates from the perspective image and scale them to 64×64 size. These image are compensated for uneven illumination by least squares fitting of a compensation plane to the image intensity. Then they are verified by distance from feature space (DFFS) [10] to reject non-face candidates. Illumination compensation is needed since PCA method is sensitive to illumination variations. The training face images of the PCA subspace are also compensated in the same way. Then the upper-left corner coordinates and the size of the verified face cropping window are associated to the existing tracks by nearest neighborhood, then used to update a constant velocity Kalman filters [11] for face tracking as

$$\begin{bmatrix}\mathbf{x}(k+1)\\ \dot{\mathbf{x}}(k+1)\end{bmatrix}=\begin{bmatrix}\mathbf{I} & T\cdot\mathbf{I}\\ 0 & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{x}(k)\\ \dot{\mathbf{x}}(k)\end{bmatrix}+\begin{bmatrix}T^2\cdot\mathbf{I}/2\\ T\cdot\mathbf{I}\end{bmatrix}\nu(k)$$
$$\mathbf{y}(k)=\begin{bmatrix}\mathbf{I} & 0\end{bmatrix}\begin{bmatrix}\mathbf{x}(k)\\ \dot{\mathbf{x}}(k)\end{bmatrix}+\omega(k) \qquad (1)$$

where the state **x** and measurement **y** are 3 by 1 and **I** is a 3 by 3 identity matrix. T is the sampling interval or frame duration that is experimentally measured. The covariance of measurement noise $\omega(k)$ and the covariance of random maneuver $\nu(k)$ are empirically chosen for a smooth but agile tracking. The states are used to interpolate detection gaps and predict the face location in the next frame. For each track, an ellipse search mask is derived from the prediction and fed back to ellipse detection for the next frame. This search mask speeds up the ellipse detection by minimizing the ellipse search area. It also helps reducing false-positives as illustrated in Figure 2.

A face track is initialized when a single-frame face is detected for some consecutive frames. Once the face is found and under tracking, the ellipse search window is narrowed down from full-frame search. The track is terminated when the predicted face location is classified as non-face for some consecutive frames.

3. Robust Estimation of Face Orientation: A Multistate Approach

From the detected face video, face orientation need to be estimated to assess the attentive direction of the subject or to determine a camera that is closest to the frontal view of the subject's face for face recognition. We first propose a simple scheme for estimating face orientation as illustrated in Figure 3. First the face image is compared to the view-based PCA templates to assess the face orientation. In the training stage, we first collect a set of equalized training faces of multiple people with multiple horizontal face orientations. The orientations of the training faces are approximately –60, –30, 0, 30, and 60 degrees. Note that the PCA subspace is constructed from the correlation matrix of the training faces instead of covariance matrix [10], since the projection vectors are normalized to reduce the effect of mean illumination variations [2][13]. Mean and covariance of the normalized training projections are estimated for each face orientation category to build a multidimensional Gaussian likelihood function. In the estimation stage, the face image is projected into the PCA subspace to estimate the face orientation by maximum likelihood (ML) decision. The estimated face orientation is then filtered by another Kalman filter as in equation (1) across frames.

There are several defects of this simple scheme of face orientation estimation. First, the ML estimator gives high quantization steps of facial poses. Second, the output of Kalman filter would have some dynamic delay to the actual face orientation. Third, the Gaussian likelihood information in Figure 3 is discarded by the ML decision, which would be very useful in making accurate estimations. To improve these problems, we construct a continuous density hidden Markov model (CDHMM) to fully utilize the information as shown in Figure 4 [12]. The Markov chain is linear bi-directional with N states which correspond to certain facing angles in order to model a continuous face turning. The observation probability of the j-th states $b_j(O)$ is modeled by a mixture of M Gaussian distributions, where O is the projection vector of the face image in the PCA subspace. The state sequence $q(k)$ of the face video can be estimated by maximum *a posteriori* (MAP) estimation in real-time or optimally estimated by Viterbi algorithm with some delay caused by sequence framming. The estimated state sequence represents the face orientation movement of the subject.

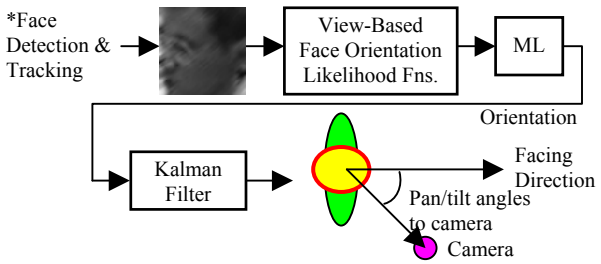

Figure 3. The first scheme of face orientation estimation.

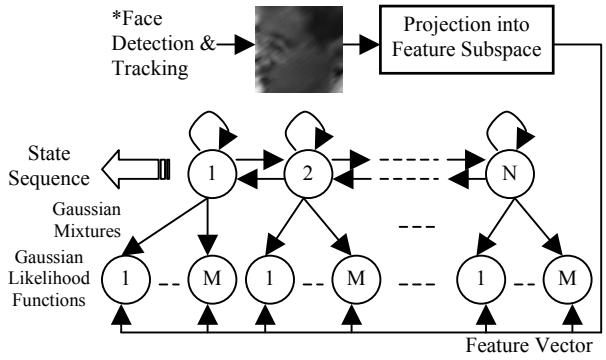

Figure 4. Modified face orientation estimation by continuous density HMM. Face video is projected into feature subspace and generates *M* Gaussian likelihood values.

There are two issues on training the CDHMM. The first is on initializing the model. The initial probabilities π, state transition matrix A, and mixture coefficients C can be initialized randomly or uniformly, but the mean vectors μ's and covariance matrices U's of the Gaussian likelihood functions should be carefully initialized [12]. We collect the ground truth facing angles of every frame in the training face sequences. Then we use Linde-Buzo-

Gray vector quantization (LBG-VQ) algorithm to partition the normalized training PCA projection vectors of the face videos into N regions. The facing angle of a region is estimated by averaging the ground truth facing angles of the training projection vectors that belong to the region. Thus we assign the N regions to the N states according to the averaged facing angles, from small to large. Finally the μ's of the M Gaussian functions of a state are determined by a second round LBG-VQ on the training projection vectors in the assigned region of that state. The U's of the Gaussian functions are initialized as σI, where I is the identity matrix.

The second issue is on determining the facing angle of a certain state. After training the CDHMM, we estimate the state sequence $s[n] = \left\{k_{n=1}^{T} \middle| k = 1,2,\ldots,N\right\}$ of the training face video. Let the ground truth sequence of the training face video be $t[n] = \left\{\theta_n \middle| n = 1,2,\ldots,T\right\}$. We want to find the association function:

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

We do this by minimizing the mean squared error to find the optimal $A(s)$,

$$MMSE = \min_{A_1, A_2, \ldots, A_N} \left(\frac{1}{T}\right) \sum_{n=1}^{T} \left[A(s[n]) - t[n]\right]^2 \tag{3}$$

Since the angles $A_1, A_2, \ldots, A_N$ are mutually independent, the *MMSE* in equation (3) can be partitioned into N sub-problems,

$$MMSE = \left(\frac{1}{T}\right) \sum_{i=1}^{N} \min_{A_i} \sum_{n=1}^{T_i} \left[\underbrace{A(s[n])}_{=A_i} - t[n]\right]^2 \tag{4}$$

where T_i indicates all the time indices when $s[n] = i$. Then for each sub-problem in equation (4), the least squares solution

$$A_i = E\left[t[n]\right] = \frac{1}{T_i} \sum_{n=1}^{T_i} t[n] \tag{5}$$

for $i = 1,2,\ldots,N$ leads to the global minimization of the *MSE* in (3). Therefore the optimal state-angle assignment of a certain state is the averaged ground truth facing angles of the frames in which the state sequence of the training face video equals to that certain state.

4. Experimental Evaluation and Analysis

Evaluation of the head tracking and face orientation estimation is accomplished using an extensive array of experimental data. We tested many video clips taken on different setups and environments, including indoor, outdoor, and mobile, as shown in Figure 9. In these test clips, the omnicamera and subjects were fixed and no person tracking was involved (cf. Figure 1). For testing the accuracy of face detection and tracking in Figure 2, the perspective view of a subject was manually selected. Figure 5 shows some indoor single-frame face detection results. Rows 1 shows the source images, row2 shows overlapped edge gradient strength, skin-tone area, detected ellipse, and fused face cropping frame before Kalman tracking, and row 3 shows the cropped face images after Kalman tracking. Column 1 to column 4 indicate that the skin-tone and ellipse detection cooperates well to detect faces on extreme situations like turned away face, highly cluttered background, and invasion of non-face skin-tone objects to the face blob. We currently tested the clips with the measurement noise variance set to 64 $(\text{pixel})^2$ and the random maneuver variance set to 512 $(\text{pixel})^2$. The standard deviation of the detected face alignment within the 64 by 64 face video after tracking is about 7 pixels.

Figure 5. Some results of the proposed multi-primitive face detection and tracking. See text for details.

For testing the face orientation estimations, we use a mobile video of 2300 frames for training and testing the CDHMM. The ground truth facing angles of the video are estimated manually frame by frame. The video is processed twice by the face detection and tracking to extract two face video sequences of the same length but different face alignments, due to the hardship of obtaining ground truth. The CDHMM is trained by one sequence as in Section 3, then tested by another sequence. We tried different CDHMM configurations as shown in Figure 6 to find a best combination of the numbers of states N, Gaussian mixtures M, effective dimension Dim, and initial covariance. The effective dimension is the first several PCA components of the full dimension of 135. They carry most of the orientation information [9]. The initial covariance is the value used to initialize the covariance matrices of the Gaussian likelihood functions. Note that to model rapid face turnings, a longer transition length (nonzero terms off the diagonal elements in the state transition matrix A) would be needed to allow farther jumping between the Markov chain states.

After the trial phase of the CDHMM, N=38, M=1, Dim=10, and σ=1/2 seems to exhibit the best performance as shown in Figure 8. Figure 7 shows the comparison of

the Kalman filter based schemes. Comparing the single-frame ML sequences of 5 quantization steps built from static images and 38 steps built from much more video frames by Baum-Welch (EM) training, the filtered versions of these two are definitely less noisy. Detailed comparison on the Kalman filtered 38-step ML sequence with the CDHMM output sequence indicate that there exists some delay and overshoot issue in the Kalman filter case, also the RMSE of the Kalman sequence (14°) is higher than the CDHMM one (12°). Thus the CDHMM scheme is preferable. The reason that the CDHMM approach performs better is that it is a *delayed decision* approach, i.e., the decision is not made before the final output. In the Kalman cases, ML decision is made before the filter for both 5-step and 38-step cases.

The current results support the effectiveness of the CDHMM scheme. In the future we still need to collect more videos on different people, gender, illuminations, and environments with a device that measures the ground truth face orientation synchronously with the face video. These videos are needed to train the CDHMM to increase the robustness so that the face orientation estimator can be used on broader range of environments, subjects, and setups. Also, pitch and roll angles of face orientation are to be included in the CDHMM. Lastly, the performance of the CDHMM-based face orientation estimation might be improved if other subspace methods such as LDA templates [13] are used, because these methods are proven to be more robust to uneven illumination conditions.

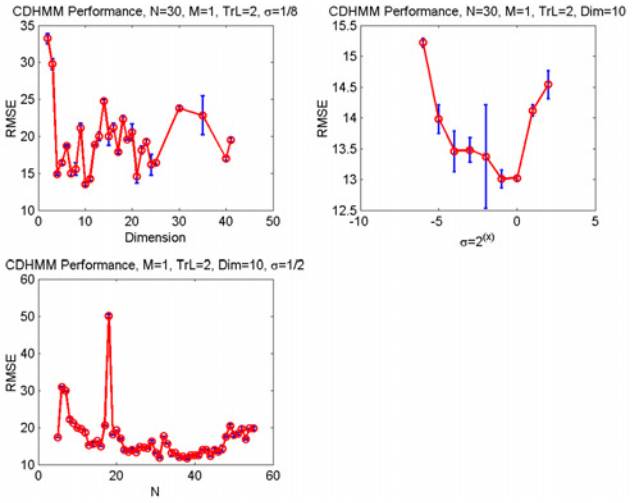

Figure 6. CDHMM performance in terms of the root MSE (RMSE) of the estimates to the ground truth. Top-left: Versus effective dimensions *Dim* with numbers of states *N*=30, Gaussian mixtures *M*=1, transition length *TrL*=2, initial covariance of the Gaussian likelihoods σ=1/8. The minimum RMSE occurs when *Dim*=10. Top-right: Versus σ with *N*=30, *M*=1, *TrL*=2, and *Dim*=10. Minimum RMSE occurs at σ=1/2. Lower-left: Versus *N* with *M*=1, *TrL*=2, *Dim*=10, and σ=1/2. Minimum RMSE occurs at *N*=38.

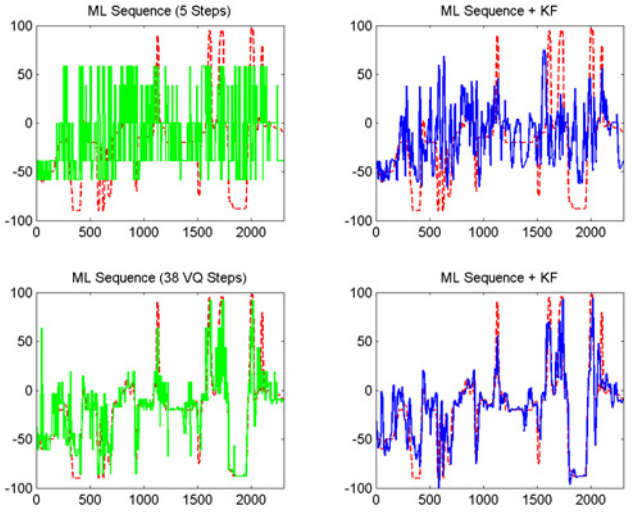

Figure 7. Face orientation estimation performance of the Kalman filter based scheme in Figure 3. The left column shows the ML sequence of 5 quantization steps as described in Section 3 and of the 38 observation likelihood functions borrowed from the CDHMM training. The right column shows the Kalman filtered ML sequences. The solid line is the estimated face orientation, and the dotted line is the ground truth value.

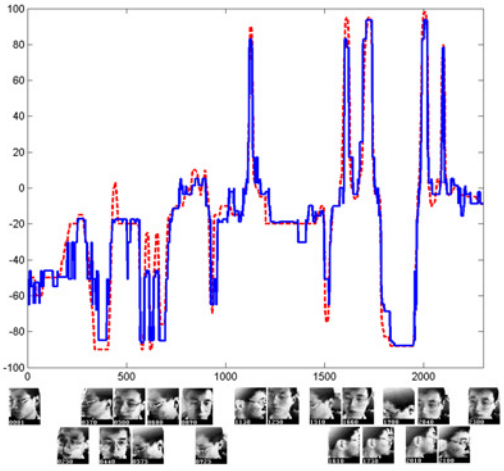

Figure 8. Face orientation estimation of the CDHMM scheme in Figure 4. The horizontal axis is the frame number. The solid line is the estimated face orientation, and the dotted line is the ground truth value. The CDHMM configurations are *N*=38, *M*=1, *Dim*=10, *TrL*=2, and σ=1/2.

5. Concluding Remarks

In this paper we have presented an integrated machine vision system for capturing the 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 Kalman filtering based tracker 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] E. Hjelmas and B. K. Low, "Face Detection: A Survey," *Comp. Vis. Img. Understd.*, vol. 83, pp. 236-274, 2001.

[2] Self reference.

[3] Self reference.

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

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

[6] J. Yang and A. Waibel, "A Real-Time Face Tracker," *Proceedings of WACV'96,* pp. 142-147, 1996.

[7] R. McLaughlin, "Randomized Hough Transform: Better Ellipse Detection," *Proc. IEEE TENCON DSP Appl.*, pp. 409-414, 1996.

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

[9] J. Ng and 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] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE Comp. Soc. Conf. on Comp. Vis. and Patt. Recog.*, pp. 586-591, Jun. 1991.

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

[12] 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.

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

Figure 9. Sample images from the test and training video sequences for face detection, tracking, and face orientation estimation. Left to right column: the source omnidirectional video, the unwarped panorama, human videos. Upper to lower row: indoor, outdoor #1, outdoor #2, mobile #1, mobile #2 environments.