A HYBRID REAL-TIME FACE TRACKING SYSTEM

Ce Wang and Michael S. Brandstein
Division of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
wangc,msb@hrl.harvard.edu

ABSTRACT

A hybrid real-time face tracker based on both sound and visual cues is presented. Initial talker locations are estimated acoustically from microphone array data while precise localization and tracking are derived from image information. A computationally efficient algorithm for face detection via motion analysis is employed to track individual faces at rates up to 30 frames per second. The system is robust to nonlinear source motions, complex backgrounds, varying lighting conditions, and a variety of source-camera depths. While the direct focus of this work is automated video conferencing, the face tracking capability has utility to many multimedia and virtual reality applications.

1. INTRODUCTION

The ability to do real-time human face detection and tracking is an important requisite of many developing technologies, including interactive human-computer communications, automated surveillance systems, and virtual reality interfaces. Of special interest is the video conferencing scenario where it is desired to have a computer-controlled camera isolate the image of a specific talker within a frame, adjusting for orientation and range as well as compensating for any source motion. Additionally, such a system would have to be capable of detecting and tracking many sources over a wide range of lighting conditions and room environments.

2. PREVIOUS WORK

A number of algorithms for face detection and tracking have been proposed in literature, such as [1, 2, 3]. While successful, these methods tend to be of a computational complexity that precludes a real-time application or require some form of interactive initialization which limits their practical utility. Several, less general, real-time face trackers based primarily on skin color information have been proposed [4, 5, 6]. Appropriate color statistics are easily extracted from an image and are frequently sufficient to distinguish a face from the background while being invariant to rotation. However, face detection based on color regions alone is susceptible to many factors, such as variations in lighting conditions, skin color, and the background image. For example, despite the use of the color normalization methods our experiments have shown that it is not unusual for commonly used white and yellow wall colors to be indistinguishable from the facial color distributions when lighting conditions were varied from those under which the original color statistics were produced. To compensate for these effects, adaptive methods for estimating the face color distribution have been employed [5, 6]. While effective to some degree, these procedures impose a heavy overhead on the face detection algorithm, particularly at the initialization stage. This situation may not be overly detrimental when the target and background are relatively stable. However, for the video conferencing scenario described above where the talkers and background environment change frequently, the computational requirements of these more complex schemes are prohibitive.

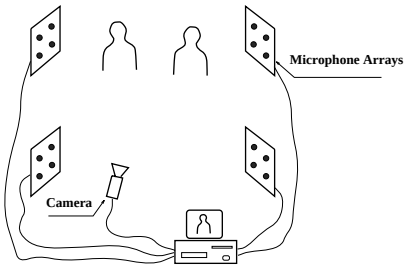

Figure 1. System framework

As an alternative to relying on color information as the primary detection criterion, we consider source motion as the feature of interest. This information may be easily extracted from inter-frame image differences and is relatively uninfluenced by lighting conditions. If properly utilized, it has the potential to reduce the initialization overhead while maintaining system performance. The hybrid face tracking system summarized in the following sections is designed to rely primarily on the motion information to detect and identify edges of the human body. Acoustic information is employed to orient the camera to sources outside of the image frame and to identify the active talkers.

3. TRACKING SYSTEM FRAMEWORK

The structure of the hybrid tracking system is shown in Figure 1. The unit consists of a visual tracking component and an acoustic talker localization component. The visual part is a mobile communication video camera (Cannon VC-C1) with three degrees of freedom: pan, tilt, and zoom. The acoustic portion consists of a set of four 4-element microphone arrays and is capable of localizing the positions of active talkers in the acoustic environment through a pro-

cess of time-delay estimation followed by a triangulation procedure. Furthermore, the microphone arrays may be electronically steered to provide spatially-selective speech acquisition. The computation required for this acoustic-based localization procedure is significantly less than that of the corresponding visually-oriented methods. The location estimate derived from the microphone array data is used as a coarse position measurement, providing an initial camera pointing vector and determining the region of active talkers. The visual portion of the system is then activated to refine the location estimate, frame the talker, and track any subsequent motion. During this period the acoustic portion is kept active in an effort to detect any voice activity outside of the camera viewing area. Details of the acoustic-based localization technique may be found in [7]. The following sections summarize the visually-based methods for the face detection and source tracking. Details may be found in [8].

4. FACE DETECTION

Humans in their natural state are rarely motionless. While the degree of movement can vary significantly, our experiments show that even under relatively sedentary conditions source motion is a sufficient indicator for distinguishing possible human bodies from the background environment. This observation is used as the basis of a computationally simple face detection algorithm.

4.1. Motion Detection and Noise Filtering

Motion detection is achieved through the process of inter-frame differencing. It is assumed that the camera is motionless during the analysis period. However, in what follows it would be possible to compensate for known camera movements. A motion plot, S, is generated by assigning image samples that are consistent across frames a value of 0. Otherwise the sample is set to 1. Subsequent to further processing, a neighborhood checking algorithm is employed to remove isolated noise samples. This denoising procedure is conducted as follows:

> *For each point X, s.t. $\mathcal{S}(X) = 1$, look at its neighborhood $\mathcal{N}(X)$ for all the points $\{Y_i\}$ such that $Y_i \in \mathcal{N}(X)$ and $\mathcal{S}(Y_i) = 1$. If $\|\{Y_i\}\| < T_{threshold}$ (where $\|\{Y\}\|$ is the cardinality of set $\{Y\}$) then discard this point, i.e. set $\mathcal{S}(X) = 0$. Otherwise, keep it.*

4.2. Adaptive Edge Detection

Face detection decisions are based upon contour shapes generated from an edge detection procedure. For simplicity it is assumed that all other moving objects are smaller than the target person in the camera image and maintain a minimal distance from the target body shape. This assumption will be relaxed after the initial detection stage.

An adaptive edge tracing algorithm is used to detect the human body contour. The following procedure describes the procedure for detection of the left side of the contour. Refer to Figure 2. The analysis for the right side is similar.

1. In S, compare the leftmost and rightmost points across horizontal scan lines. A region of consecutive lines with consistent leftmost and rightmost pairs is used to detect a starting position. The horizontal line at the boundary of the consistency region is denoted as the starting line, y.
2. Save the leftmost point of line y as $EdgeLeft[y]$.
3. Continue scanning downward across horizontal lines searching for the point on $y+1$ nearest to $EdgeLeft[y]$. Save this point as $EdgeLeft[y+1]$. Identify the nearest left neighbor to $EdgeLeft[y+1]$. If this neighbor exists and its separation from $EdgeLeft[y+1]$ is within a threshold T_l then update $EdgeLeft[y+1]$ to be the neighboring point. Continue this process until no other neighboring points within the threshhold T_l are found. If the current scan line is empty or has only one point, go on to the next line, and then interpolate the above empty lines.
4. Track the values $\Delta x_l(y) = EdgeLeft[y+1] - EdgeLeft[y]$. A sudden deep valley followed by a sharp peak in $\Delta x_l[y]$ is indicative of an excessive T_l value. In such a case, decrease T_l, go back to the valley, and trace the edge again. Similarly, a sharp upward sloping in $\Delta x_l[y]$ is due to an overly small T_l value. Go back to that position, increase T_l, and trace again. This process is illustrated in Figure 2.
5. After completing the lower portion of the contour, return to the starting line and apply the same procedure to the region above the starting line.

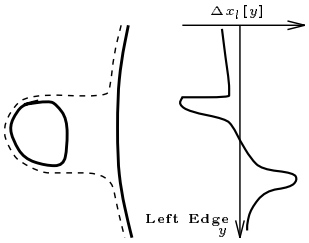

Figure 2. Adaptive edge detection (left side contour).

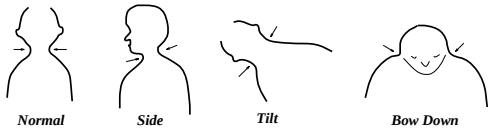

Figure 3. Necklines for various positions.

The above adaptive edge detection method, in addition to detecting the desired body contour, is capable of removing noise and any objects reasonably separated from the object of interest.

4.3. Face Detection

The final step in the face detection procedure is motivated by the observation that the neckline is a prominent feature of the body contour. As Figure 3 illustrates, the neckline is present across a variety of talker positions, orientations, and gestures, and may be roughly identified from the points on the body contour exhibiting the greatest concavity. The

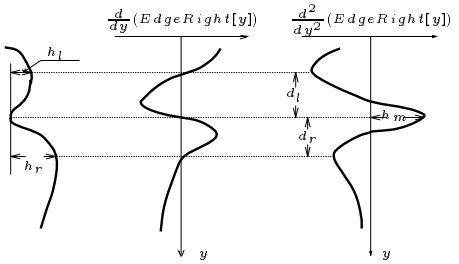

Figure 4. Neckline detection.

relative ease of estimating the neckline provides a straightforward method for identifying the separation between head and torso, thereby localizing the facial region.

The relation between the neckline and the nearby shoulder and head is shown in Figure 4. The points of peak concavity and convexity are found from the extrema of the second order derivative of the edge contour.

Because the second order derivative is sensitive to noise, the contour curve is first smoothed with a low pass filter. To improve the estimate reliability, the distances and heights of the extrema are incorporated into the following criterion for estimation of the neckline:

$$y_{neck\_point} = \arg\max\{(h_l h_r)^{w1}(d_l d_r)^{w2} h_m^{w3}\}$$

The three weights $w1, w2, w3$ determine the influence of the geometric parameters.

5. SOURCE TRACKING

Once the face is initially detected, a tracking algorithm is activated to predict the position and velocity of the target in the subsequent frames. The predicted position estimate is used to reduce the tracking search region, thereby decreasing processing requirements and providing a further resistance to noise.

5.1. Mathematical Model of Source Motion

A second order dynamic model is used to predict the source motion. The process noise is assumed to be symmetrically distributed. Model accuracy is evaluated through a simple statistical analysis.

Denote the measured position, velocity, and acceleration of a sample from the current frame t_0 to be s_0, $\dot{s}_0$, and $\ddot{s}_0$, respectively. These values may all be calculated directly from image samples from the current and past frames. Assuming knowledge of the past $n-1$ frames of image data (t_1, $\cdots$, t_{n-1}) the current state values of these motion features ($\tilde{s}_0$, $\dot{\tilde{s}}_0$, and $\ddot{\tilde{s}}_0$) are calculated from ensemble statistics as follows:

$$\begin{aligned}
\ddot{\tilde{s}}_0 &= \sum_{i=0}^{n-1} k_i \ddot{s}_i \\
\dot{\tilde{s}}_0 &= \sum_{i=0}^{n-1} k_i \dot{s}_i + \ddot{\tilde{s}}_0 \sum_{i=0}^{n-1} k_i \Delta t_i \\
\tilde{s}_0 &= -\frac{1}{2}\ddot{\tilde{s}}_0 \sum_{i=1}^{n-1} k_i \Delta t_i^2 + \dot{\tilde{s}}_0 \sum_{i=1}^{n-1} k_i \Delta t_i + \sum_{i=0}^{n-1} k_i s_i
\end{aligned}$$

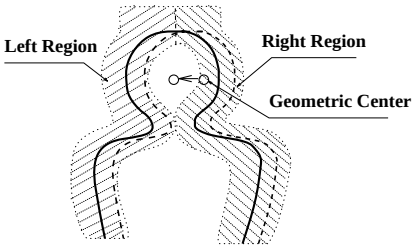

Figure 5. Illustration of edge searching regions.

where $\sum_{i=0}^{n-1} k_i = 1$, $k_0 \geq k_1 \geq \cdots \geq k_{n-1}$, and $\Delta t_i = t_0 - t_i$.

The sample position in the following frame is predicted from:

$$s'(t) = \frac{1}{2}\ddot{\tilde{s}}_0 (t - t_0)^2 + \dot{\tilde{s}}_0 (t - t_0) + \tilde{s}_0 \tag{1}$$

The confidence of the prediction is evaluated from the mean square error between the measured sample positions and the state value positions for the current and past $n-1$ frames:

$$\mathcal{E}_{\tilde{s}} = \left(\sum_{i=0}^{n-1} k_i (s_i - \tilde{s}_i)^2 \right)^{\frac{1}{2}} \tag{2}$$

5.2. Adaptive Face Tracking

Typically, head motion in space is very nonlinear, involving sudden accelerations, rotations, and changes in orientation. The corresponding image plane projection tends to be extremely complex. However, if the frame rate is sufficient, the shape differences between frames are small. This enables the use of inter-frame coherence and allows for the simplification of head image movement to a 2-D displacement according to the mathematical model described above.

The head edges are predicted to be in two regions as shown in Figure 5. The medial axis of each search region is estimated from the position prediction equation (1). The width of the region, W, is determined adaptively from the edge contour prediction accuracy statistic (2) by:

$$W = A\mathcal{E}_{\tilde{s}} + T$$

where T is a constant value designed to maintain a minimum width in the search regions. A is an adaptive factor which is updated according to the success or failure of the edge detector in the predicted regions.

6. CAMERA CONTROL

The Canon VC-C1 camera has three degrees of freedom: pan, tilt and zoom. Zoom control is evaluated from knowledge of the facial contour size and may be adjusted to provide consistent framing of the subject. The pan and tilt parameters are derived from the source position estimate and a camera motion model. The motion of the camera is approximated by a constant rotation rate plus a time delay. Details of this procedure may be found in [8]. In situations where the camera motion is too slow to track the target the system defaults to the acoustic localization procedure in an effort to relocate the source.

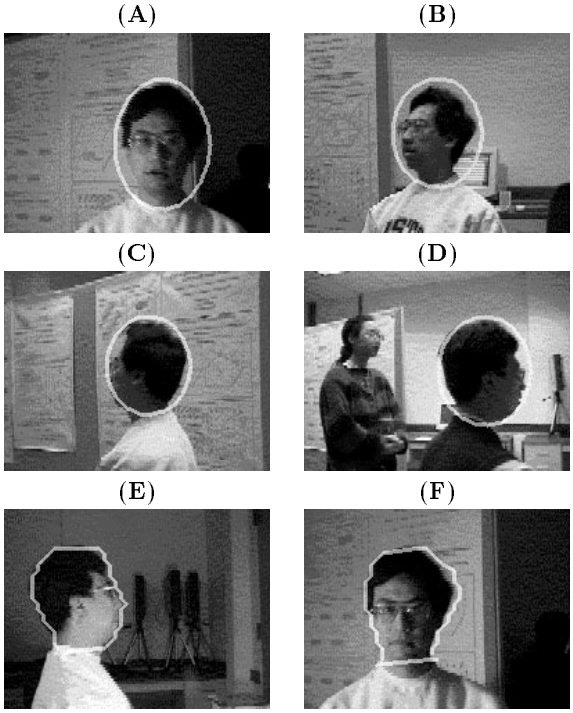

Figure 6. Detected faces.

7. PERFORMANCE EVALUATION

Computation for the visual component of the system is implemented on a Pentium Pro 200MHz PC running Windows 95. The image sizes are 144 by 192 pixels. The plots in Figure 6 display images derived from the tracking process. Plots A), B), and C) illustrate the results for various head orientations. In each case the detected head region is outlined by an oval boundary. As the results suggest the detection algorithm is relatively insensitive to facial orientation. In the case of plot C), the head is oriented away from the camera. While facial color is no longer a viable detection feature, the neckline is clearly evident and the system is capable of finding the head region based upon the edge contours derived from motion information alone. Plot D) illustrates the ability of the system to perform face detection with moving sources in the background. In this case the motion of the person in background is treated as noise relative to the movement of the foreground target, and the algorithm detects and tracks the more prominent source. This shows the noise resistance capability of the algorithm. Plots E) and F) show faces with their detected contours, illustrating the ability of the algorithm to identify the entire head contour rather than just the skin portion of the face.

This system has been tested by many people in an unconstrained laboratory environment under different lighting conditions. As a means of comparison, the advantages and similarities of the proposed system relative to a representative system are highlighted below. The CMU real-time face tracker [4] relies primarily on color statistics and operates on a HP-9000 workstation. The major differences and similarities can be listed as follow:

(1) The system presented here is capable of maintaining a nearly constant rate of 25-30 frames per second (fps) over a wide range of source-camera separations (0.5m to 10m). The CMU system reports a rate of 15 fps at a distance of 0.5m and only achieves a 30 fps rate at ranges greater than 2m.

(2) By dynamically adjusting the camera zoom, the proposed system is capable of normalizing the facial image size. The CMU system does not incorporate this feature.

(3) While the proposed system is capable of estimating complete head contours, the CMU system is limited to detecting only the skin region of the face.

(4) Both systems are capable of tracking a face when the head is still (except for the initial stage). In the proposed system this is achieved through a motion/rest decision detailed in [8].

8. CONCLUSION

This paper presented a real-time face tracker based on both visual and acoustic information, applied specifically to a video conferencing application. The major innovation of this work is the use on motion information to achieve face detection and tracking. The system requires only a moderate computational load to detect and track faces in a general environment and exhibits a robustness to various lighting conditions by virtue of its exclusion of color information in the analysis process. While the current system is limited to single source tracking, future work will focus on extending the algorithms to the multi-source scenario.

REFERENCES

[1] H. Rowley, S. Baluja and T. Kanade. Neural Network-Based Face Detection. *IEEE CVPR'96.*

[2] K. Sung and T. Poggio. Example-Based Learning for View-Based Human Face Detection. *AI Memo 1521, CBCL Paper 112, MIT*, December 1994.

[3] M. C. Burl, T. K. Leung and P. Perona. Face Localization via Shape Statistics. *Intl. Workshop on Automatic Face and Gesture Recognition*, Zurich, Switz., June 1995.

[4] J. Yang and Alex Waibel. A Real-Time Face Tracker. *Proc. of 3rd IEEE WACV,* Sarasota, Florida, Dec. 2-4, 1996, pp142-147.

[5] N. Oliver, A. P. Pentland and F. Berard. LAFTER: Lips and Face Real Time Tracker. *IEEE CVPR'97*, pp. 123-129.

[6] T. Jebara and A. Pentland. Parametrized structure from motion fro 3D adaptive feedback tracking of faces. *IEEE CVPR'97*, pp. 144-150.

[7] M. Brandstein and H. Silverman, A practical methodology for speech source localization with microphone arrays. *Computer, Speech, and Language*, 11(2):91–126, April 1997.

[8] Ce Wang and M. S. Brandstein. A Hybrid Real Time Face Tracking System. *Technical Report, Harvard University, DEAS*, September, 1997.