

ROBUST AUTOMATIC VIDEO-CONFERENCING WITH MULTIPLE CAMERAS AND MICROPHONES

Ce Wang, Scott Griebel, and Michael Brandstein

Division of Engineering and Applied Sciences
Harvard University, Cambridge, MA 02138
wangc@hrl.harvard.edu, griebel@fas.harvard.edu, msb@hrl.harvard.edu

ABSTRACT

An automatic video-conferencing system is proposed which employs acoustic source localization, video face tracking and pose estimation, and multi-channel speech enhancement. The video portion of the system tracks talkers by utilizing source motion, contour geometry, color data, and simple facial features. Decisions involving which camera to use are based on an estimate of the head's gazing angle. This head pose estimation is achieved using a very general head model which employs hairline features and a learned network classification procedure. Finally, a wavelet microphone array technique is used to create an enhanced speech waveform to accompany the recorded video signal. The system presented in this paper is robust to both visual clutter (e.g. ovals in the scene of interest which are not faces) and audible noise (e.g. reverberations and background noise).

1. INTRODUCTION

A popular area of multi-media research is automatic video conferencing, where a set of computer-controlled cameras capture the images of one or more individuals, adjusting for orientation, range, and source motion. In [1, 2, 3] we proposed a hybrid face tracking system which made use of an acoustic-based localization algorithm for initial camera steering followed by a motion-based technique to develop body contours and detect facial regions. The result was a computationally efficient multi-source tracker. This paper extends these ideas by adding head pose estimation and multi-channel speech enhancement techniques. The head pose estimation algorithm is used to decide which camera is best suited to capture a scene. This method is successful for estimating facial orientations, and does not require very high quality images. A second extension to previous results is the enhancement of the multi-microphone speech to reduce the effects of reverberations and background noise often present in teleconferencing environments. Currently, most speech acquisition systems require the user to be physically close to a device to achieve reasonable sound quality. However, a more desirable video-conferencing system would not require a physical link between participants and individual microphones. In this paper, we incorporate a model-based multi-channel speech enhancement technique first discussed in [4] and later refined in [5]. This enhanced speech signal accompanies the video signal to create a more enjoyable system for the end user.

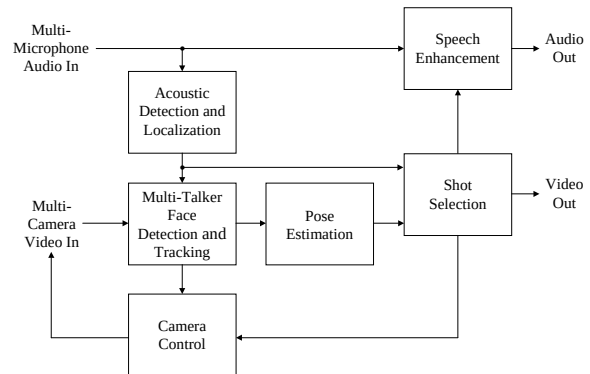


Figure 1: Video-Conferencing System Setup

2. VIDEO-CONFERENCING SYSTEM SETUP

The components of the hybrid tracking system are shown in Figure 1. The system consists of video (face detection and tracking, pose estimation, shot selection) and audio (talker localization, speech enhancement) components. The video system uses a set of cameras with three degrees of freedom: pan, tilt, and zoom. The audio system consists of a set of four 4-element microphone arrays and is capable of detecting voice activity and localizing the positions of talkers in the acoustic environment through time-delay estimation and an ensuing triangulation procedure. 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, estimate the poses of the targets, select the camera with the best view, and track any subsequent motion. The tracking information is fed to the speech enhancement procedure which then extracts the acoustic signal associated with the desired talker(s), attenuating unwanted sources, noise, and reverberation.

3. VISUAL FACE TRACKING AND POSE ESTIMATION

The methods used for face detection and tracking were first presented in [1, 2]. Multi-talker face detection is accomplished via a hierarchical procedure employing four features derived from the video stream: (1) source motion, (2) contour geometry, (3) color statistics, and (4) facial analysis. Source motion and contour geometry identify potential face regions while color and facial analysis are used as validation criteria. Because these latter criteria

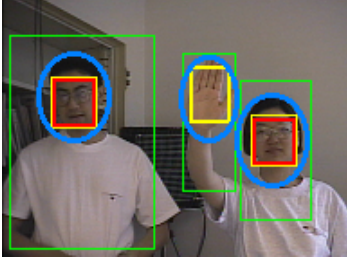


Figure 2: Face Tracking Example



Figure 3: Hairlines of Different Individuals

are used only for confirmation purposes rather than as the primary means of face detection, both the color test and facial analysis can be quite general and robust to varying conditions. At each stage of the algorithm, the appropriate features are estimated from the image data with only a limited amount of computation.

Figure 2 displays a sample image from the face detection process. The large boxes delineate the segmented regions derived from the motion information. The ovals are the result of the geometrical analysis which generates a two-sided contour boundary for each region and then estimates a neckline location. For this example each of the three candidates (two faces and a hand) were found to possess a headlike geometry. The color confirmation test is then applied within each of the ovals. The detected skin color regions are indicated by a light box. To this point, the hand is still misclassified as a face candidate due to its similar shape and color. After searching for eye regions, the hand is discarded from the set of possible face regions, as indicated by the darker boxes within the lighter boxes.

Given the tracking data provided by the face detection procedure, we now address the problem of automatically estimating an individual's head orientation. This information is useful for determining shot relevance and deciding which camera best captures the scene of interest. A number of methods for facial orientation or head pose estimation have been proposed in the literature [6, 7, 8]. The general approach involves estimating the positions of specific facial features in the image and fitting this data to a head model. Many of these methods still require manually selecting feature points and assume near-frontal views and high-quality images. For our applications, such conditions are not usually satisfied. Poor lighting conditions and hidden or unclear facial features are common problems.

In situations where most of the fine facial details, (e.g. eyes, lips, mouth) are difficult to detect reliably, we focus on a more overt property - the border between an individual's hair and skin. We are able to detect the hair-skin border over a wide range of viewing angles and light intensities. The curved shape of the hairline is sufficient for satisfactory estimation of the head orientation. Figure 3 illustrates the hairline contour derived from three people with different orientations.

Estimation of the hairline contour is performed in conjunction with the head tracker described previously. Within the head regions, hairlines are generated by clustering brightness, color and

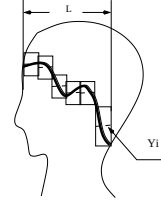


Figure 4: Hairlines as a feature

connectivity into hair and skin regions. Having the pre-segmented head region available significantly simplifies and improves this process. Once estimated, the hairline is parameterized into a simple feature vector. Referring to Figure 4, denote the extent of the hairline along the horizontal axis as L . Along L , the hairline is evenly segmented into 6 pieces. The average vertical position of the contour segment for each piece, Y_i , is normalized to form the 6-point feature vector $\vec{y}$, whose elements are given by $y_i = [Y_i - (Y_3 + Y_4)/2] / L$, ($i = 1, \dots, 6$). The purpose of this normalization is to make the feature vector invariant to the head size and relative frame position.

The problem of orientation estimation becomes one of finding the mapping function $f : \vec{y} \rightarrow \theta$, $\theta \in [-\pi, +\pi]$, where θ is the facial orientation. Given a training data set consisting of N vector-orientation pairs $\{\vec{y}_i, \theta_i\}$, the mapping function is specified from local linear projections associated with specific training data subsets. Specifically, the training samples are first partitioned into three classes C_1 , C_2 and C_3 , based upon their coarse orientation angles. These classes roughly correspond to right, left, and back facing head orientations. For each class and its corresponding subset of training vectors, S_j , the local orientation estimate is found from $\hat{\theta}_j = \vec{\alpha}_j^T \vec{y} + \beta_j$. The terms $\vec{\alpha}_j$ and β_j are obtained from minimizing:

$$\sum_{i \in S_j} p(\vec{y}_i, \theta_i) (\theta_i - (\vec{\alpha}_j^T \vec{y} + \beta_j))^2. \quad (1)$$

The probability density $p(\vec{y}_i, \theta_i)$ is included in the above expression to suppress the influence of outlier training data. This density function may be estimated by a Parzen window [9] approach:

$$p(\vec{y}, \theta) = \frac{1}{N} \sum_{i=1}^N \Phi(\vec{y}, \vec{y}_i, \Sigma) \Phi(\theta, \theta_i, \sigma), \quad (2)$$

where Φ is the Gaussian function. The variance matrices, Σ and σ , are predefined. Note that $p(\vec{y}, \theta)$ is estimated on the whole training set, not the individual subset. The least squares criterion in (1) may be recast as a matrix equation and solved via the pseudo-inverse.

In practice, an observed feature vector $\vec{y}$ is first classified into one of the three subsets using a Bayesian Classifier, i.e. finding the class which maximizes the *a posteriori* probability given by: $p(C_j | \vec{y}) \propto p(\vec{y} | C_j) p(C_j)$, $j = 1, 2, 3$. The *a priori* term is given by $p(C_j) = |S_j|/N$ and the class-conditional density function is found from:

$$p(\vec{y} | C_j) = \frac{1}{|S_j|} \sum_{i \in S_j} \Phi(\vec{y}, \vec{y}_i, \Sigma). \quad (3)$$

Without loss of generality, assume C_1 is the class which maximizes $p(C_j | \vec{y})$ and C_k , $k \in \{2, 3\}$ is the class which is second most likely. If the likelihood associated with C_1 is significantly greater than that of C_k , then the pose orientation is estimated directly from the subset 1 projection. Specifically, given a predefined

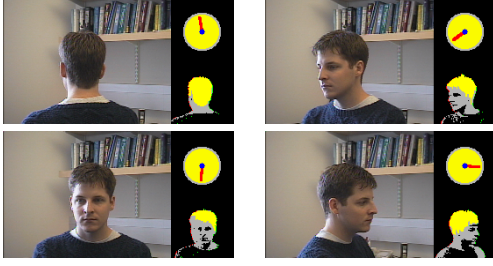


Figure 5: Examples of Face Orientation Estimation

threshold ϵ , if $|p(C_1|\vec{y}) - p(C_k|\vec{y})| > \epsilon$, then the estimated orientation is $\theta(\vec{y}) = \hat{\theta}_1 = \vec{\alpha}_1^T \vec{y} + \beta_1$. Otherwise the orientation angle is found using a weighted sum of the individual subset orientation angles according to:

$$\begin{aligned} \theta(\vec{y}) = & \left(\vec{\alpha}_1^T \vec{y} + \beta_1 \right) \left(\frac{p(C_1|\vec{y})}{p(C_1|\vec{y}) + p(C_k|\vec{y})} \right) \\ & + \left(\vec{\alpha}_k^T \vec{y} + \beta_k \right) \left(\frac{p(C_k|\vec{y})}{p(C_1|\vec{y}) + p(C_k|\vec{y})} \right). \end{aligned}$$

Note that the number of rough classes into which the continuous range of orientations are segmented is subject to the balance between the efficiency of the local prediction and the class separability. Increasing the number of classes improves prediction accuracy of the linear projection, but at the cost of decreased subset classification performance.

Figure 5 illustrates the pose estimation results derived from this procedure for a video stream. In each picture, a single image frame is shown along with a clock indicating the estimated facial orientation and a plot illustrating the segmented head region along with the detected hair and skin regions. Six o'clock and twelve o'clock correspond to forward and backward-facing subjects respectively. The pose estimation results are very accurate across the complete range of possible orientation angles despite the presence of object shadows and a complex image background. This illustrates the appropriateness of the extracted image features and the effectiveness of the proposed pose learning and estimation procedures. Our experiments indicate that the hairline contour can be reliably estimated from facial images encompassing a range of lighting conditions and source-camera distances using only relative intensity statistics, color disparity, and region connectedness. In situations where the skin or hair regions are partially misidentified, the pose estimator still provides reasonable results.

4. MULTI-CHANNEL SPEECH ENHANCEMENT

In our previous work, the audio signal accompanying the video recording was obtained through a simple delay-and-sum beamforming of the multi-channel speech data [3]. We now outline a multi-microphone procedure to produce an accompanying speech signal that is robust to reverberations and background noise. The reverberant speech signal, $x_i[n]$, observed at microphone i ($i = 1, 2, \dots, I$) can be modeled as:

$$x_i[n] = h_i[n] * s[n] + u_i[n], \quad (4)$$

where $s[n]$ is the clean speech utterance, $u_i[n]$ is noise, and $h_i[n]$ is the room impulse response between the speech source and microphone i . Previous work in [4, 5] introduced a multi-channel approach which used a nonlinear event-based processing method

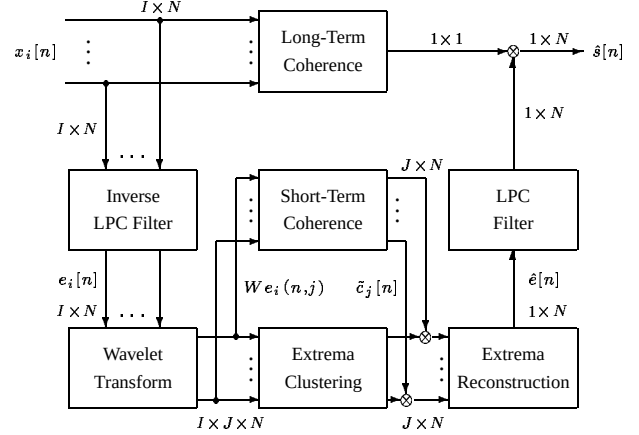


Figure 6: Outline of the Proposed Algorithm

to discriminate impulses in the received signals due to channel effects from those present in the desired speech.

The speech production system can be modeled by an all-pole filter excited by a quasi-periodic impulse train for voiced speech or by random noise for unvoiced speech. It is assumed that the noise and extra excitation impulses due to reverberations are relatively uncorrelated across the individual channels, while the excitation impulses due to the original speech are less affected by the environment. The approach is to identify the clean speech excitation signal from a set of corrupted excitations. The enhanced speech is then reconstructed by using the enhanced excitation signal as input to an estimate of the all-pole filter representing the vocal system. The overall procedure outlined in Figure 6 was first presented in [5]. The received signals, $x_i[n]$, are time-aligned either through *a priori* knowledge of the source and microphone geometry or by using a time-delay estimation procedure.

A joint LPC filter is derived from N -point windows of the I channels of speech. This LPC filter is found by using a multi-channel approach outlined in [5]. The resulting LPC coefficients can be used to generate a set of residual signals, $e_i[n]$, via inverse filtering. The goal of the algorithm is create a single enhanced residual signal, $\hat{e}[n]$. This residual will then be used as input to the LPC filter to generate an enhanced speech signal, $\hat{s}[n]$.

The first J scales of the wavelet transform (WWT) are calculated for each of the I residual signals. In this work, the WWT is calculated for the first 5 scales. The WWT of the residual signal from channel i , $e_i[n]$, is denoted by $W e_i(n, j)$ where j is the scale parameter and n is the temporal index within that scale. We use the non-decimated quadratic spline wavelet discussed in [10]. This set of wavelets was shown to be effective at detecting singularities in a signal. The set of extrema at each wavelet scale is an effective representation of the impulses in the excitation signals. By locating the extrema which are well clustered across all channels, we attempt to capture the underlying impulsive structure (for voiced speech) of the original non-reverberant LPC residual signal. Another advantage of working in the wavelet domain is the ability to approximately reconstruct a signal from the local extrema of its WWT coefficients at each scale [10]. The estimated 'clean' multi-scale extrema can be used to reconstruct the desired excitation signal.

Based upon the observation that residual impulses due to the original speech tend to be clustered across channels and scales while those due to reverberation effects are relatively uncorrelated in both amplitude and time, a clustering technique is now applied

to estimate the ‘clean’ extrema. The reverberant wavelet extrema at each scale tend to be clustered around the local wavelet extrema of the delay-and-sum beamformed speech. In general, for voiced speech, extrema due to glottal excitations tend to be more closely clustered than those due to additional impulses from the reverberant channel responses. This causes the less clustered peaks to have a relatively small amplitude. The clustering operation maps the multi-channel WVT extrema data to a single WVT extrema representation suitable for input to the WVT extrema reconstruction algorithm described in [10].

We now further de-emphasize the clustered peaks due to reverberations beyond what is achieved by the extrema clustering process. The following short-term coherence window is multiplied by the wavelet extrema at each scale j :

$$c_j[n] = \frac{2}{L-1} \frac{\sum_{l=1}^L \sum_{k=i+1}^L \sum_{l=-L}^L W e_i(n+l, j) W e_k(n+l, j)}{\sum_{l=1}^L \sum_{l=-L}^L W e_i(n+l, j)^2}, \quad (5)$$

where $2L + 1 \ll N$ is the length of a small subwindow within the analysis frame, on the order of 3 to 5 msec. With this window, any clustered extrema in areas of low coherence will be de-emphasized, while highly coherent extrema will be relatively unaffected. The coherence function as given above can assume values between -1 and 1 . Any coherence values below a threshold, T , are set to $T / \max_n c_j[n]$ while any values above T are set to $c_j[n] / \max_n c_j[n]$ to avoid excessive smoothing of the resulting speech signal.

After multiplying the clustered peaks for each wavelet scale by the appropriate scale-dependent coherence envelope, the enhanced LPC residual is reconstructed using the techniques discussed in [10]. For voiced frames, a quasi-pitch-synchronous smoothing approach is employed at the coarser wavelet scales to ensure that the peaks from consecutive voiced periods are slowly varying [5].

For each analysis frame, the reconstructed residual signal, $\hat{e}[n]$, is applied to the estimated LPC filter to generate the enhanced speech signal. The final step is to scale the reconstructed speech frames by a measure that is based on both the long-term coherence of the speech channels and the energy of the beamformed signal to obtain the enhanced speech frame, $\hat{s}[n]$. This second weighting serves two purposes (1) it adjusts the energy of each frame to provide a relatively smoothly varying output and (2) it allows for the de-emphasis of lengthy signal portions which are due primarily to reverberation effects (e.g. immediately after a sudden vowel stoppage).

Figure 7 shows the results of the overall enhancement method for a reverberation time of 400 msec and additive noise with an SNR of 10 dB relative to the original clean speech. The figure plots the original speech, the signal received at a single microphone, the product of delay-and-sum beamforming, and the result achieved using the proposed enhancement algorithm.

5. CONCLUSION

This work presented a robust automatic video-conferencing system combining our previous work involving face tracking, head pose estimation, and multi-channel speech enhancement. The face tracker is able to track different people in a dynamic environment. The head pose estimate can be used to select the most desirable camera to record a particular speaker. Finally, the speech enhancement algorithm provides an audio signal largely free of reverberation and additive noise to accompany the recorded video signal.

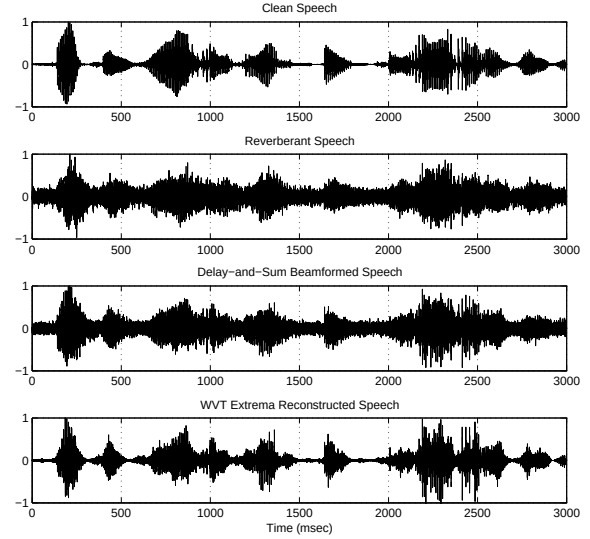


Figure 7: Results of Proposed Speech Enhancement Algorithm

6. REFERENCES

- [1] C. Wang and M. S. Brandstein. A hybrid real-time face tracking system. In *ICASSP'98*, volume 6, pages 3737–3741, Seattle, Washington, May 12-15 1998.
- [2] C. Wang and M. S. Brandstein. Multi-source face tracking with audio and visual data. In *Proceedings of IEEE 3rd Workshop on Multimedia Signal Processing*, pages 169–174, Copenhagen, Denmark, September 13-15 1999.
- [3] C. Wang and M. S. Brandstein. Real-time face tracking using acoustic and visual data. Technical Report 1, HIMMEL, Harvard University, Cambridge, MA, September 1997.
- [4] M. Brandstein. An event-based method for microphone array speech enhancement. In *ICASSP99*, pages 953–956, Phoenix, AZ, March 15-19 1999. IEEE.
- [5] S. Griebel and M. Brandstein. Wavelet transform extrema clustering for multi-channel speech dereverberation. In *IEEE Workshop on Acoustic Echo and Noise Control*, Pocono Manor, Pennsylvania, September 27-30 1999.
- [6] R. Lopez and T. S. Huang. Head pose computation for very low bit-rate video coding. In *6th International Conference on Computer Analysis of Images and Patterns*, pages 440–447. Springer-Verlag Berlin Heidelberg, 1995.
- [7] T. Maurer. Estimation of face position and pose with labeled graphs. In *BMVC*, 1996.
- [8] N. Kruger, M. Potzsch, and C. Malsburg. Determination of face position and pose with a learned representation based on labeled graphs. *Image and Vision Computing*, 15(8):665–673, August 1997.
- [9] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.
- [10] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(7):710–732, July 1992.