Multimodal Person Recognition using Unconstrained Audio and Video

Tanzeem Choudhury, Brian Clarkson, Tony Jebara, Alex Pentland
Perceptual Computing Group
MIT Media Laboratory
Cambridge, MA 02139
{*tanzeem,clarkson,jebara,sandy*}*@media.mit.edu*

Abstract

We propose a person identification technique that can recognize and verify people from unconstrained video and audio. We do not expect fully frontal face image or clean speech as our input. Our recognition algorithm can detect and compensate for pose variation and changes in the auditory background and also select the most reliable video frame and audio clip to use for recognition. We also use 3D depth information of a human head to detect the presence of an actual person as opposed to an image of that person. Our system achieves 100% recognition and verification rates on natural real-time input with 26 registered clients.

1 Introduction

Automatic identification of people has many applications in different areas. If the recognition can be performed in an unobtrusive manner it will be useful in secured access sites, for automatic banking, password-free computer login and also analysis of person dependent- behaviors and preferences.

Relatively high accuracy rates have been obtained in face recognition using computer vision techniques alone and by fusing with other modalities like speaker verification and different bio-metric measurements. But much less work has been done in person identification where there is little or no restriction on the person's movement or speech. Researchers have proposed different techniques that can handle varying pose by using template matching techniques or by modeling the pose variations as manifolds or subspaces in a high dimensional image space [5, 3].

The main goal of this paper is to recognize a person using unconstrained audio and video information. We derive a confidence scoring which allows us to identify the reliable video frames and audio clips that can be used for recognition. We also propose a robust method based on 3D depth information for rejecting imposters who try to fool the face recognition system by using photographs.

2 Face Recognition

In a realistic environment, a face image query will not have the same background, pose, or expression everytime. Thus we need a system that can detect a face reliably in any kind of background and recognize a person despite wide variations in pose and facial expression. The system also must be able to pick out reliable images from the video sequence for the recognition task. We propose a technique which uses real-time face tracking and depth information to detect and recognize the face under varying pose.

2.1 Face Detection and Tracking

The first step of the recognition process is to accurately and robustly detect the face. In order to do that we do the following:

1. Detect the face using skin color information.
2. Detect approximate feature location using symmetry transforms and image intensity gradient.
3. Compute the feature trajectories using correlation based tracking.
4. Process the trajectories to stably recover the 3D structure and 3D facial pose.
5. Use 3D head model to warp and normalize the face to a frontal position

We model the skin color (RGB values) as a mixture of Gaussians. To train our model we take samples from people with varying skin tone and under different lighting conditions. This model is then used to detect regions in the image that contain skin color blobs. The largest blob is then processed further to look for facial features e.g. eyes, nose and mouth. Our method does not require the face to be frontal for the detection stage. The loci of the features give an estimate of the pose. Using this pose estimate and a 3D head model we warp the detected face to a frontal view. This frontal face then undergoes histogram fitting to normalize it's illumination. For a detailed description please refer to [7].

2.2 Eigenspace Modeling

Once a face has been detected and its features identified, the image region containing the face is sent for recognition. The face finding stage gives us only an approximation of the feature locations. We refine these estimates by re-searching for eyes, and mouth within a small area around the previous estimate. This overcomes the time consuming stage of face and facial feature detection in the whole image and makes the recognition process suitable for real-time application. After the feature locations have been accurately detected the face is normalized such that the eyes and mouth are at fixed locations.

Our eigenspace is built using training images provided by the real-time face tracker. We use the thirty-five eigenvectors with the largest eigenvalues to project our images on to. Having a 3D model for pose normalization allows us to use a single eigenspace for a range of poses. This eliminates the requirement for storing and selecting from multiple eigenspaces. Thus our face detection algorithm does not constrain the user to maintain a still frontal pose.

To capture the facial variations for each person, we fit a Gaussian to their eigencoefficients. We define the probability of a match between a person and a test image to be the probability of the test image eigencoefficients given the person's model. In the unimodal case, the person that has the maximum probability for that test image is the claimed identity. You can see in Figure 1 the distribution of the eigencoefficients for two people and it also demonstrates the differences in the mean coefficients between two people.

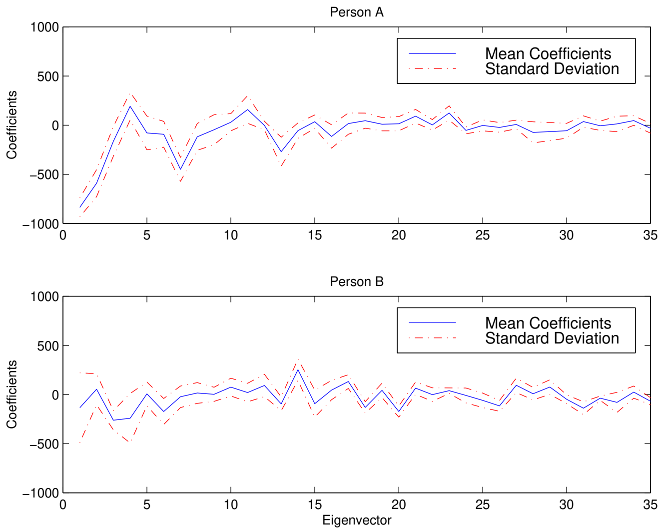

Figure 1: Distribution of the first 35 eigen coefficients for person A and B

2.3 Depth Estimate

If face recognition is used for security purpose, it is important that the system is not fooled by a still image of the person. The structure from motion estimate in the tracking stage yields depth estimates for each of the features. We can use this information to differentiate between an actual head and a still image of one. A picture held in front of the camera, even if it is in motion, gives a flat structure. Figure 2 shows the depth values extracted for a few test trials. The photograph yielded the same depth value over all of its feature points, while the depth values varied greatly for actual faces. We are also looking into reliably recovering the 3D structure of individuals to use for recognition purposes.

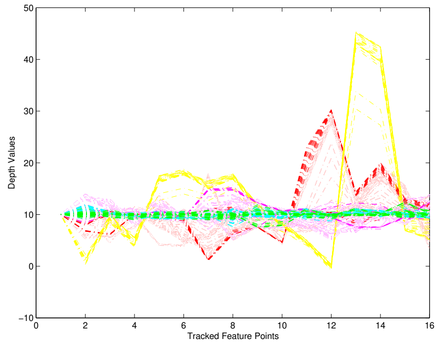

Figure 2: Depth values for tracked objects: the object with all of its features at the same depth is a photograph, the rest are faces.

3 Speaker Identification

Past work has shown that text-independent speaker identification (SI) relies on the characterization of the spectral distributions of speakers. However, convolutional and additive noise in the audio signal will cause a mismatch between the model and test distributions, resulting in poor recognition performance [8, 1]. Even if the audio channel is kept consistent so as to minimize convolutional noise, there will always be the problem of additive noise in natural scenarios.

Deconvolutional techniques such as RASTA [6] have had substantial success in matching the spectral response of different auditory channels. However, severe drops in performance are still evident with even small amounts of additive noise.

Work done by [1] has suggested that the presence of noise doesn't necessarily degrade recognition performance. They compared their system's error rates on a clean database (YOHO) and a more realistic database (SESP). When training and testing were done on the same database the error rates were comparable.

Building on this idea, our speaker identification system is based on a simple set of linear spectral features which are characterized with HMMs. This simple combination is well-suited for adapting the speaker models to various types of background noise.

3.1 Event Detection

The first state in the audio pipeline is the coarse segmentation of the incoming audio. The purpose of this segmentation is to identify segments of audio which are likely to contain speech. We chose this route because it makes the statistical modeling much easier and faster. Instead of integrating over all possible segmentations, we have built-in the segmentation as prior knowledge.

We used a simple and efficient event detector, constructed by thresholding total energy and incorporating constraints on event length and surrounding pauses. These constraints were encoded with a finite-state machine. The resulting segmentation yields a series of audio clips that can be analyzed for speaker identification.

This method's flaw is the possibility of arbitrarily long events. If for example there was a jack hammer nearby then the level of sound would always exceed the threshold. A simple solution is to adapt the threshold or equivalently scale the energy. The system keeps a running estimate of the energy statistics and continually normalizes the energy to zero mean and unit variance (similar to Brown's onset detector [2]). The effect is that after a period of silence the system is hypersensitive and after a period of loud sound the system grows desensitized.

3.2 Feature Extraction

After segmentation the (16 kHz sampled) audio is filtered with a weak high-pass filter (preemphasis) in order to remove the DC offset and boost the higher frequencies. We calculate Mel-scaled frequency coefficients (MFCs) for frames of audio that are spaced 16 ms apart and are 32 ms long. This frame size sets the lower limit on the frequency measurement to approximately 30 Hz. Mel-scaling increases the resolution for lower frequencies, where speech typically occurs.

MFC is a linear operation on the audio signal, so additive noise does not cause a nonlinear distortion in our features. This useful because it allows us to detect additive noise given a model of the noise in isolation.

3.3 Modeling

Our system uses HMMs to capture the spectral signature of each speaker. An HMM for each person is estimated from examples of their speech. The estimation was achieved by first using segmental k-means to initialize HMM parameters and then Expectation-Maximization (EM) to maximize (locally) the model likelihood. Since the examples of speech are text-independent there is no common temporal structure amongst the training examples. This situation requires the use of fully-connected (ergodic) HMMs.

In order to find the optimal model complexity for our task, we varied the number of states and number of Gaussians per state until the recognition rate was optimized. We tested HMMs with 1 to 10 states and 1 to 100 Gaussians. The best performance was achieved with a 1 state HMM with 30 Gaussians per state or, equivalently, a mixture of 30 Gaussians. This is not surprising given the lack of temporal structure in our text-independent training and testing examples. Arguably this makes the use of HMMs unnecessary. However, the use of HMMs is justified for our background noise adaptation.

3.4 Background Adaptation

Statistical models trained on clean speech (or speech in any specific environment) will perform badly on speech in a different environment. The changing environment causes distortions in the speech features which create a mismatch between the test speech and model distribution. Convolutional noise is caused primarily by differing microphone and sound card types, and microphone and sound source location. Additive noise is caused by the presence of other sound sources. We will assume that the microphone type and location is constant and concentrate on additive noise only.

The goal is to be able to adapt models of clean speech for use in noisy environments. However, the adaptation cannot require samples of the speech in the noisy environment because usually they are not available. So given only the clean speech models and recordings of the background noise, our adaptation technique can estimate the appropriate noisy speech models.

The model adaptation procedure (which is related to the parallel model combination algorithm of [4]) is based on estimating HMMs for noisy speech from HMMs separately trained on speech and noise. Since the background noise might have temporal structure, such as repetitive noises like motor noise, or randomly occurring changes like thunder in a rain storm, it is appropriate to use an HMM to model it. The feature extraction and HMM training was the same as above.

If the background noise and speech are assumed independent and the features are extracted using only linear operators then the distributions can be easily estimated. Let B be the background noise HMM with M states, S the clean speech HMM with N states, and S' the noisy speech HMM. The combination of the two HMMs, S and B, is the HMM S' with $M * N$ states in the state space constructed from the outer product of the S and B state spaces. The probability distributions for each state in S' are the convolution of the distributions in S with the distributions in B.

This adaptation was evaluated using the speech of 26 people(their collection is described below) and an auditory background scene of a street in a thunder storm. The noise scene contains frequent thunder and occassional passing cars against a constant background of rain. We created two sets of audio data: a *Speech Only* set with uncorrupted speech, and a *Speech + Noise* set which was constructed by adding the background recordings to the audio clips in the *Speech Only* set. They were mixed at a Signal-to-Noise Ratio (SNR) of 7dB. Each of these sets were further divided into training and test sets.

A single state HMM, S_i, was trained on the speech of each individual from the *Speech Only* set. A 3-state HMM, B, was trained on the background sounds. This HMM was used to adapt the S_i HMMs thereby creating a new set of HMMs, S'_i, which should match the speech in the *Speech + Noise* set. Although this is not an option for real-time adaptation, we also trained HMMs, call them C_i, on the *Speech + Noise* training set to evaluate the effectiveness of the adaptation.

Finally we test all HMMs on both the *Speech Only* and *Speech + Noise* test sets. Table 3 contains the recognition rates for two sets of 130 audio clips. As shown by the extremely poor performance of the S HMMs on the *Speech + Noise* test set, the background scene has clearly caused a mismatch between the speech models and the audio. The adaptation is able to regain 95% of the performance if we assume the C HMMs are exactly matched to the *Speech + Noise* set.

HMM Models	Speech Only	Speech + Noise
Speech Only (S)	71.5%	23.1%
Adapted (S')	N/A	65.4%
Corrupted (C)	N/A	69.2%

Figure 3: Recognition rates for the clean speech, corrupted speech and adapted speech models.

4 Classifier Fusion

The goal of classifier fusion is to complement one modality with the another. If a classifier is performing poorly then it is important not to let its suggestions skew the final decision. Therefore, careful considerations must be made to ensure the appropriate weighting of each classifier.

The derivation of this weighting relies on having a measurement of each classifier's reliability. Let $P(X_i = j)$ be the probability that classifier i assigns to person j. These probabilities are calculated from the model likelihoods, $L(X_i = j) = P(Data|Model_j)$:

$$P(X_i = j) = \frac{L(X_i = j)}{\sum_k L(X_i = k)}$$

While this normalization is necessary for comparing classifier scores, it also removes any measure of how well a test case is modeled by the classifier (i.e. $P(data|$ all models$)$).

4.1 Confidence Scores

We have tried numerous measures for estimating a classifier's confidence. For the face classifier, we tested confidences based on the following measures (x is a test image):

Distance from Face Space (DFFS)

$$DFFS(x) = ||x - \bar{x}||_{Eigenspace}$$

Aggregate Model Likelihood (AML)

$$AML(x) = \log\left(\sum_j P(x|Model_j)\right)$$

Maximum-Probability to Average-Probability Distance (MPAP)

$$MPAP(x) = \max_j\{P(X = j)\} - \frac{1}{N}\sum_j P(X = j)$$

The speech classifier was evaluated with only the AML and MPAP measures. Since the above measures can have arbitrary ranges and distributions we converted them to probabilities with the following transformation ($M(x)$ is one of the measures above):

Let $p(\omega)$ = pdf for the r.v., $\omega = M(x)$, then

$$\text{confidence}(\omega_0) = P(\omega < \omega_0) = \int_{\infty}^{\omega_0} p(\omega)\, d\omega$$

We estimate $p(\omega)$ from a set of images or audio clips using Parzen windows with Gaussian kernels. Figure 4 shows this mapping for the DFFS measure.

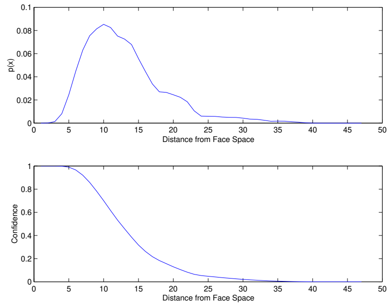

Figure 4: Mapping the DFFS to a probability: (top) the DFFS pdf for a set of images, (bottom) the derived cdf which is used for the mapping.

Table 5 shows how each confidence measure performs as a predictor for recognition. The percentages are based on the correlation between the confidence scores and the correctly or incorrectly recognized test cases. A score of 50% (chance) means that the confidence score is uncorrelated with recognition. The MPAP measure clearly outperforms the rest of the measures and hence it was adopted as the confidence measure for the system.

4.2 Bayes Net

In the fusion of classifiers, each knowledge source may be dependent on other sources. The full Bayesian approach assumes the least by assuming each knowledge source is dependent on all the other sources. This requires the estimation of many conditional distributions which in turn requires large amounts of training data. However, many of the dependencies are unnecessary and we will make our assumptions explicit with a Bayes Net.

The knowledge sources for each classifier, $i \in$ {(S)peech,(F)ace}, are:

1. $P(X|X_i)$ - classifier's probability for each person

Confidence Score	Speech	Face
DFFS	N/A	55.3%,90.0%
AML	50.2%,47.6%	N/A
MPAP	71.4%,50.3%	99.1%,53.4%

Figure 5: Comparison of Confidence Scores: Prediction rates of Correct Recognition (left) and Wrong Recognition (right).

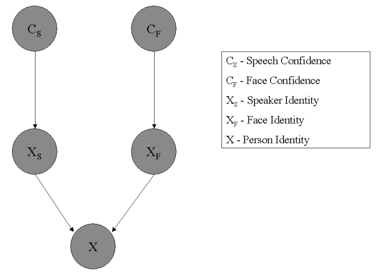

Figure 6: The Bayes net used for combining knowledge sources for each classifier to produce a final decision, X.

2. $P(X_i|C_i)$ - confidence in the classifier

where the r.v. C_i ={reliable, not reliable}, and the r.v. $X_i = \{j | j \in$ Client Database$\}$. Figure 6 displays the Bayes net we used to combine these knowledge sources.

The audio and video channels are assumed conditionally independent as depicted by the lack of direct links between C_S and C_F , and X_A and X_F . We are also assuming that the confidence scores are conditionally independent from X. This is equivalent to assuming that the distributions of confidence scores are the same for both classifiers.

$$P(X) = P(X|X_S)P(X_S|C_S)P(C_S) + P(X|X_F)P(X_F|C_F)P(C_F)$$

Finally, the prior on each confidence score, $P(C_i)$, is simply the recognition rate for each classifier. This prior should be estimated separately for each individual, but the lack of training data forced us to use the same prior for everyone.

5 Experiments

Both recognition and verification experiments were performed. We describe the data collection process and then discuss some of the results using various methods of evaluation.

5.1 Data Collection

We collected our data for training and testing using an Automated Teller Machine (ATM) scenario. The setup included a single camera and microphone placed at average head height. A speech synthesis system was used to communicate with the subjects rather than displaying text on a screen. The reasons for this are two-fold. First, the subjects won't be constrained to face the screen at all times. Second, it is more natural to answer with speech when the question is spoken as well. The subjects were instructed to behave as if they were at an actual ATM. No constraints were placed on their movement and speech.

The session begins when the subject enters the camera's field of view and the system detects their face. The system then greets the person and begins the banking transaction. A series of questions were asked and after each question the system waited for a speech event before proceeding to the next question. A typical session was as follows:

1. Wait for a face to enter the scene
2. System: "Welcome to Vizbank.Please state your name"
3. User: "Joe Schmoe."
4. System: "Would you like to make a deposit or a withdrawal?"
5. User: "Ummm, withdrawal."
6. System: "And the amount please ?"
7. User: "Fifty dollars."
8. System: "The transaction is complete. Thank you for banking with us"
9. Wait for the face to leave the scene
10. Go back to step 1

During the transaction process the system saves 40X80-pixel images centered around the face and audio at 16 KHz. We collected data from 26 people.

5.2 Evaluation Methods

We evaluated our system using both recognition and verification rates. Both procedures include a criteria for rejecting clients entirely based on the probability output of the Bayes net. Rejection means that the system did not get a suitable image clip or audio clip for recognizing or verifying the client. Usually an application would ask the client to repeat the session.

The recognition procedure is as follows:

1. The Client gives no information.
2. Recognized Identity = $\arg\max_j \{P(X = j)\}$.
3. Reject if $P(X =$ Recognized Identity$) <$ Rejection Threshold.

The verification procedure is as follows:

1. The Client gives a Claimed Identity.
2. Recognized Identity = $\arg\max_j \{P(X = j)\}$.
3. Reject if $P(X =$ Recognized Identity$) <$ Rejection Threshold.
4. Verify *iff* Recognized Identity = Claimed Identity *else* reject.

The results for each experiment are analyzed for hit rate and correct rejection rate over the entire range of rejection thresholds. The optimal operating threshold is theoretically where the sum of hit and correct rejection rates are maximized. This is assuming equal cost weights for hit rate and correction rejection rate. For each experiment we give the success rate at both zero threshold (i.e. no rejections) and the optimal operating threshold.

Modality	Per Image/Clip	Per Session
Audio	71.2 %	80.8 %
Video	83.5 %	88.4 %
Audio + Video	93.5 %	100 %

Figure 7: Recognition Rates (Zero Rejection Threshold): no rejections

Modality	Per Image/Clip
Audio	92.1% (28.8%)
Video	97.1% (17.7%)
Audio + Video	99.2% (55.3%)

Figure 8: Recognition Rates (Optimal Rejection Threshold): the rejection rates are in parentheses.

5.3 Results

Results for our recognition and verification processes were calculated based on audio information and video information alone and also by combining the outputs using the Bayes Net described above. We calculate rates both using all the images/clips and using only the "best" clip from each session. Where "best" is defined as the image/clip with the highest confidence score. For session-based applications the latter is more meaningful because it identifies the session accuracy rather than the accuracy per video frame and audio clip.

Table 7 gives an overview of the system's recognition performance when no thresholding is used. The recognition is perfect when done on a per session basis using only the most reliable image/clip pair. Table 8 shows what the optimal operating point is for per image/clip recognition. The high rejection rates are quite reasonable given that there were at least 7 image/clips per person.

The verification rates are in table 9. The verification is near perfect (99.5%) with only 0.3% false acceptance rate on a per image/clip basis. The session performance is perfect.

As is expected, when we prune away the less reliable images and audio clips, the performance increases appreciably. When we use only the most confident images and clips both the recognition and verification rates rise to 100% with no false acceptances.

6 Conclusions

We have implemented and evaluated a system that combines face recognition and speaker identification modules for high accuracy person recognition. Furthermore, both of these modules were designed to take a large variety of natural real-time input. The face recognition module achieves high recognition accuracies by combining face detection, head tracking, and eigenface recognition. The text-independent speaker identification module is robust to changes in background noise by incorporating adaptation in its event detection and modeling stages. We use a simple Bayes net to combine the outputs of our face and speech modules. However this method is made quite effective by deriving and including confidence scores that can predict each module's reliability. In fact, we have shown the system can select, based on the confidence scores, the most reliable images and audio clips from each session and in this case perform with perfect recognition and verification rates.

Modality	Per Image/Clip	Per Session
Audio	97.8 % (0.2%)	98.5 % (0%)
Video	99.1 % (0.2%)	99.6 % (0%)
Audio + Video	99.5 % (0.3%)	100 % (0%)

Figure 9: Verification Rates (Optimal Rejection Threshold): the false acceptance rates are parentheses.

Acknowledgments

The authors thank Sumit Basu for suggesting the use of depth map for recognition. We also thank Bernt Schiele, Jennifer Healey and Sumit Basu for many helpful discussions.

References

[1] F. Bimbot, H. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg, and J. Pierrot. Speaker verification in the telephone network: Research activities in the cave project. Technical report, PTT Telecom, ENST, IDIAP, KTH, KUN, and Ubilab, 1997.

[2] G. J. Brown. *Computational Auditory Scene Analysis: A representational approach.* PhD thesis, University of Sheffield, 1992.

[3] Brenden Frey Antonio Colmenarez and Thomas Huang. Mixture of local linear subspaces for face recognition. In *International Conference on Computer Vision and Pattern Recognition.*

[4] M.J.F. Gales and S.J. Young. Robust continuous speech recognition using parallel model combination. Technical report, Cambride University, 1994.

[5] Daniel Graham and Nigel Allinson. Face recognition from unfamiliar views: Subspace methods and pose dependency. In *Third International Conference on Automatic Face and Gesture Recognition.*

[6] Hynek Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn. Rasta-plp speech analysis. *ICSI Technical Report TR-91-069*, 1991.

[7] Tony Jebara and Alex Pentland. Parameterized structure from motion for 3d adaptive feedback tracking of faces. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[8] Jiyong Ma and Wen Gao. Text-independent speaker identification based on spectral weighting functions. *AVBPA Proceedings*, 1997.

[9] E.S. Bigun, J.Bigun, B. Duc, S. Fischer Expert Conciliation for Multi Modal Person Authentication Systems by Bayesian Statistics In *First International Conference on Audio- and Video-based Biometric Person Authentication.*