FaceFacts : Study of Facial Features for Understanding Expression

by

Tanzeem Khalid Choudhury

B.S., Electrical Engineering
University of Rochester, 1997

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of Master of Science at the
Massachusetts Institute of Technology

September 1999

Signature of Author .....

Program in Media Arts and Sciences

August 6, 1999

Certified by .....

Alex Pentland

Toshiba Professor of Media Arts and Sciences

Program in Media Arts and Sciences

Accepted by .....

Stephen A. Benton

Chair, Departmental Committee on Graduate Students

Program in Media Arts and Sciences

FaceFacts : Study of Facial Features for Understanding Expression

by

Tanzeem Khalid Choudhury

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning on August 6, 1999, in partial fulfillment of the requirements for the degree of Master of Science.

Abstract:

This thesis provides a framework for automatic detection, recognition and interpretation of facial expressions for the purpose of understanding the emotional or cognitive states that generate certain expressions. It provides the motivation and lays the initial foundation for the study and analysis of facial expressions in natural conversations — starting from data recording to feature extraction and modeling. Many of the existing techniques for feature tracking and modeling rely on noise free *perfect* data, which is impossible to obtain from real world events — the reliability and robustness of different existing techniques in real world situations are studied. All the analysis is done for person specific models. To allow the use of person specific models, a multi-modal person recognition system is developed for robust recognition in noisy environments. The study shows that it is very difficult to process and model events from spontaneous and natural interactions. The results show that some expressions are more easily identifiable, such as blinks, nods and head shakes, whereas expressions like a talking mouth and smiles are harder to identify. Data from conversations was recorded under different conditions, ranging from fully natural and unconstrained to having subjects' heads fixed in place. Observations made from comparing natural conversation data with constrained conversation data show that useful expression information can be lost due to imposing constraints on a person's movement. Thus, if automatic expression analysis is to be a useful input modality in different applications, it is necessary to study expressions in a natural and unconstrained environments.

Thesis Advisor: Alex Pentland
Toshiba Professor of Media Arts and Sciences

FaceFacts

Thesis Committee

Thesis Reader .....

Rosalind Picard

Associate Professor of Media Arts and Sciences

Massachusetts Institute of Technology

Thesis Reader .....

Nancy Etcoff

Assistant Professor

Harvard Medical School

Acknowledgments

First, I would like thank my advisor Sandy for his support and guidance throughout the last two years at the Media Lab. His advise and insightful discussions always steered me towards the right direction.

I would also like to thank my readers, Rosalind Picard and Nancy Etcoff for all the time and effort they put into reading my thesis and giving me helpful comments and suggestions.

I would like to thank Linda Peterson and Jennifer Ruotolo for their help with the administrative details of submitting a thesis.

Next I would like to thank Sumit Basu for many helpful discussions, and for his friendship. Thanks to Tony Jebara, Sumit Basu and Jacob Strom for letting me use the head-trackers they developed and answering all my questions. Thanks to Brian Clarkson for his HMM codes and to Bernt Scheile for his local receptive field histogram code.

Thanks you, all my friends, for being there for me and making my life richer and more exciting. Special thanks to Normin and Jenn, whom I could always count on to share my moments of sorrows and joy.

Thanks to Rajeev for your support, enthusiasm and love, for all our stimulating conversations, and for giving me hope when I almost lost it. I also want to thank apu (my sister) for giving me strength and confidence when I felt the weakest and most vulnerable. Finally, I would like to thank abbu and ammu (my dad and mom) for giving me the freedom to explore different things in life and for having faith in me — to you I am eternally grateful.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

After we have given the men all the suggestions we can that have to do with expressing ideas through the body, then we can come down to the value of the facial expression — the use of the eyes, eyebrows, the mouth — their relation to one another — how the eyes and the mouth work together (sometimes) for expression — how they work independently for expression at other times.

- Walt Disney

1.1 Introduction

The main focus of this thesis is to process and extract facial features from natural conversation sequences for the purpose of understanding emotional or cognitive states that generate certain expressions.

Realizing that there is always a tradeoff between depth and breadth, the thesis aims in building personal expression models as opposed to global models, i.e. the person's identity is known to the system. With current advances with person recognition technology this is not a difficult task, especially when the number of people to be recognized is not excessively large. This thesis also describes a method of person identification from unconstrained audio and video. In order to use personalized models it is necessary not only to identify a person, but to identify a person in a natural

environment where the person is allowed to move freely and talk in the presence of background noise.

1.2 Motivation

Understanding facial expressions of people gives insight about human emotions and behavior. Facial expression is one of the most important modalities used for non-verbal communication. People are very good at understanding and interpreting expressions of people of various backgrounds, ages and nationalities. In order to understand the subtleties of expression and the range of their expressiveness it is necessary to study and analyze naturally expressive data as opposed to pre-staged or *fake* expressions and transfer this power of interpretation to computers and machines.

Expression understanding is of great interest to researchers in different areas e.g. cognitive science, psychology, social psychology, computer vision, computer animation, affective computing, medicine. The ability to track, analyze and interpret natural expression will be useful in many ways (Picard, 1997). The two major areas of applications are:

1. For detailed and extensive studies of facial expression in a rigorous manner, which would be of interest to psychologists, doctors, cognitive scientists etc.

2. For machines towards understanding human behaviors through being able to detect and understand salient expressions in human-computer-interaction(HCI) or human-machine-interaction (HMI).

1.2.1 Meaning and Interpretation of Facial Expressions

One of the main interests in studying facial expression is to gain insight into people's emotional and cognitive states and to understand some personality traits. It is also of great importance in medical studies. For example, a link between certain negative expression and Type A myocardial eschemia is being studied (Ekman and J.,). Abnormal and characteristic facial expressions have been detected in chronic

schizophrenic patients (Jucke and Polzer, 1993). In psychology facial expressions are of great interest in studying how people react in different situations. What are the set of emotions people have and how do they express them ? Facial expressions provide information on whether a person is empathetic or hostile, abusive or non-abusive — this is useful for people in social psychology. Cognitive science researchers are interested in finding the link between facial expression and internal states. Computers empowered with the ability to analyze and understand expressions can assist these area of research.

1.2.2 Facial Expression and Human-Machine-Interaction

Development of systems that can parse natural interaction can revolutionize the scope of man-machine-interaction. There is a lot of interest in analyzing human expressions and simulating those expressions on responsive toys, e.g., Sony Robot, affective toys etc., to generate lifelike animation in computer graphics. Another aspect of expression analysis is to make machines understand the expressive spectrum and regulate or influence how machines interact with humans based on the expressions of the user.

1.3 Thesis Overview

This thesis focuses on the following elements:

Person Recognition: Recognizing people in natural environments with noisy audio and video as inputs to the recognition system. Start with detecting the presence of a person and then recognizing his/her identity. Robust combination of the audio/video modality for maximizing recognition accuracy.

Data Collection: Setting up a data collection environment which is conducive for natural interactions. The data collection process involves the recording of audio and video of each conversation sessions.

Head Tracking: Testing different head-tracking algorithms on recorded sequences in order to determine the robustness of these algorithms with real-life data and their applicability in real-time expression processing.

Feature Extraction: Selecting facial features that provide the most information. Selecting and extracting facial regions that are most expressive. Evaluating the effectiveness of different features in capturing the subtleties of expressions and their robustness in natural environments.

Feature Modeling: Temporal modeling of feature trajectories for identifying and detecting expressions that are important in interactions.

Observations: Discussion on observations made from the recorded data. Events that appear repeatedly and consistently in a conversation and give information about the conversation.

Results and Conclusion: Results obtained, directions for future research and conclusion.

Chapter 2

Background

2.1 Introduction

Facial expression is an important source of information in human-human communication and more recently in human-machine interactions. Facial expressions are outcomes of and are influenced by various human emotions and sensations.

2.2 Classical Theories in Facial Expression Understanding

Some of the very early studies of facial expression was done by Duchenne (Duchenne, 1990) and Charles Darwin (Darwin, 1872) in the early nineteenth century. In his book, *The Expression of the Emotions in Man and Animals*, Darwin proposes that some the chief expressions exhibited by humans are universal. For example, he argues that in suddenly looking at something or while looking around, everyone raises his/her eyebrows. Duchenne in *Mecanisme de la Physionomie Humaine* remarks the people often raise their eyebrows while trying to remember something. Later, Ekman in his famous study (Ekman and Friesen, 1975) provided experimental results that supported some of Duchenne and Darwin's claims. He demonstrated that, of the human expressions, six are universal across nationality and culture. These are:

happiness, sadness, anger, fear, surprise, and disgust. He later extended his list to include contempt. Also his findings suggest that a number of positive emotions — amusement, pride, relief, exhilaration — share the same facial expression, a particular type of smile (Ekman, 1992b).

Ekman and colleagues (Ekman, 1982a; Ekman, 1982b; Ekman, 1992a; Ekman, 1992b; Ekman and Friesen, 1975; Hager, 1985; Hager and Ekman, 1996) in their research explore many characteristics of facial expressions. Facial actions or signals greatly influence the information content of non-verbal communication. Non-verbal communication as defined in (Hager and Ekman, 1996) is the idea that bodily changes have an effect on another person during conversations. Hager and Ekman (Hager and Ekman, 1996) argue that the best way to measure facial signals is to measure the facial sign vehicles the signals are based on. Faces convey information through four general classes of signals or sign vehicles:

1. Static facial signals — facial constitution like bone structures, tissue mass etc.
2. Slow facial signals — changes in appearance over time.
3. Artificial signals — eyeglasses, cosmetics etc.
4. Rapid facial signals — phasic changes in neuromuscular activity. Rapid signals usually have a duration between 250ms and 5s.

Rapid signals have been the main focus of study for expression understanding. Rapid signals can convey five types of messages -

1. Emotions - happiness, sadness, anger, fear, disgust, surprise and contempt.
2. Emblems - Culture specific symbolic communicators such as winking
3. Manipulators - self-manipulative associated movements such as lip biting
4. Illustrators - actions accompanying and highlighting speech such as a raised brow.
5. Regulators - non-conversational mediators such as nods or smiles.

The temporal dynamics of facial signal vehicles are also important in understanding expressions. While a limited set of expressions can often be recognized from single snapshots, the ability to discriminate subtle expressions requires comparison over time. Hager and Ekman (Hager and Ekman, 1996) claim that these subtle differences not only signify which emotion is occuring but also provide information about emotion intensity, whether multiple emotions have blended together, or whether a person is trying to control expression of emotion.

To measure and quantify expressions Ekman introduced the Facial Action Coding System (FACS) which measures muscle actions in terms of action units. Usually more than one muscle action are combined into an action unit and each expression is a combination of different action units.

2.3 Computer Vision Approaches for Understanding Expression

The work done so far by computer vision researchers in automatic understanding of facial expressions has focused mainly on classification. In most cases, the goal is to identify the universal expressions defined by Ekman. Thus, the algorithms developed were tailored towards building models to recognize a set of six or seven expressions from static images or video sequences. Techniques based on sequences, as opposed to static images, were more successful in recognizing expressions in a more generalized environment.

Yacoob and Davis (Yacoob and Davis, 1994) looked at optic flow of high gradient points on the face for analysis and recognition of facial expressions. Bassili (Bassili, 1979) in his study concluded that principal facial motions provide powerful cues to subjects to recognize facial expressions Figure 2-1. Yacoob and Davis' recognition algorithm is based on this result — they use optic flow to come up with a dictionary for facial dynamics. They divide each expression into three temporal segments — beginning, epic, ending — and use a rule based approach for classifying different expressions.

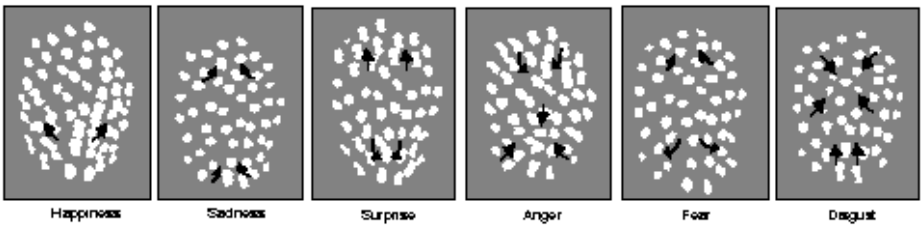


Figure 2-1: Cues for facial expression recognition proposed by Bassili

Essa and Pentland (Essa and Pentland, 1994; Essa et al., 1994) developed a computer vision system for facial expression representation (FACS+), which probabilistically characterizes facial motion and muscle activation. They use optic flow coupled with a geometric and physical face model. The systems outputs the spatio-temporal changes in the shape of a face and a parametric representation of muscle actions groups. The muscle action patterns are the basis of FACS+ and are used for analysis, recognition and interpretation of expressions.

Bartlett (Bartlett et al., 1996) attempts to capture the subtleties of facial expressions by detecting the muscular actions that comprise a facial expression. She proposes a method for automating facial action scoring based on holistic spatial analysis, feature measurement and optic flow, and presents experimental results for recognizing the upper facial actions. However, her analysis relies on manual time-warping of expressions to a fixed length. Many of her subjects were trained in performing FACS action units; therefore, her results might not be generalizable to non-trained people (Cohn et al., 1999).

Black and Yacoob (Black and Yacoob, 1997) model expressions using local parametric motion of facial features, such as the eyes, eyebrows, and mouth. Their system is able to track rigid and non-rigid facial motion using a set of parameterized optic flow models. They use the motion parameters to get a high-level semantic description of facial feature motions and to recognize facial expressions in presence of significant head motion.

Most of the work done so far is towards recognizing prototypic expressions. Cohn et al. (Cohn et al., 1999) are currently working on building a system capable of

distinguishing between the subtleties in expressions by discriminating FACS action units. The Automatic Face Analysis system consists of three modules - 1. feature point tracking that tracks the movement of preselected feature points, 2. dense flow tracking that calculates the flow of the entire face and 3. furrow detection that tracks changes in facial lines and furrows. The output of these modules are used to train HMMs that are able to discriminate between FACS action units. Cohn and Katz (Cohn and Katz, 1998) also combined vocal expressions using prosodic measurements with facial expression analysis to develop a system that can discriminate between subtle changes in facial expressions and recognize a speaker's communicative intent.

Some more work has been done in relating facial expressions to other responses i.e. speech. DeSilva, and Cheng and Huang (Desilva et al., 1997; Chen and Huang, 1998) use multi-modal inputs (audio and video) to classify expressions. They attempt to recover the modality that has dominance for a given expression. However, their findings are based on a small data set and the expressions are on demand, i.e. not taken from natural conversation.

Chapter 3

Person Recognition

3.1 Introduction

Some facial expressions have been identified as universal — nonetheless, facial expression of emotion is not necessarily a unique pan-cultural signal (Ekman, 1993). Even Ekman in *Facial Expression and Emotion* (Ekman, 1993) admits that there are large individual differences in a number of aspects of the timing of facial expressions. Also, there are variations in a given expression - 60 variations of the anger expression have been studied. Although there might be some core configurational properties common to all there are some differences. Thus, identity information of the person whose expressions are being studied may help resolve some conflicts between expressions and also allow a finer scale study of the expressional content.

In this chapter, I present the work we did for person recognition as a preprocessing step for expression analysis (Choudhury et al., 1999; Pentland and Choudhury, 1999). As expression is to be studied in a unconstrained environment, the recognition system also has to be robust to these changes. A bi-modal recognition system was developed using face and voice information. Voice is not only a good modality to use for recognition but also can be used for vocal expression analysis. Its been suggested that each expression that has a facial expression also has a vocal expression (Tomkins, 1963).

3.2 Person Recognition using Unconstrained Audio and Video

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 (Graham and Allinson, 1996; Colmenarez and Huang, 1998).

Our main goal was 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. As an added bonus of our head-tracker, we develop a robust method based on 3D depth information for rejecting impostors who try to fool the face recognition system by using photographs.

3.3 Face Recognition

In a realistic environment, a face image query will not have the same background, pose, or expression every time. 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 expressions. The system must also 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.

3.3.1 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:

(a) Input Image (b) Frontally Warped Image

Figure 3-1: Warping of input image to a frontal position for recognition

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. We take samples from people with varying skin tone and under different lighting conditions to train our model. 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. This 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 its illumination. For a detailed description of the system please refer to (Jebara and Pentland, 1996).

3.3.2 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 eigen-space is built using the normalized training images (zero mean and unit variance) provided by the real-time face tracker. We use the thirty-five eigen-vectors with the largest eigen-values, that capture 95% of variance, to project our images on to. Having a 3D model for pose normalization allows us to use a single eigen-space for a range of poses. This eliminates the requirement for storing and selecting from multiple eigen-spaces. 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 35 dimensional Gaussian to their eigen-coefficients from approximately 600 images per person. We define the probability of a match between a person and a test image to be the probability of the test image eigen-oefficients given the person's model. In the unimodal case, the person that has the maximum probability for that test image is the claimed identity. Figure 3-2 show the distribution of the eigen-oefficients for two people and it also demonstrates the differences in the mean coefficients between two people.

3.3.3 Depth Estimate

If face recognition is used for security purposes, 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

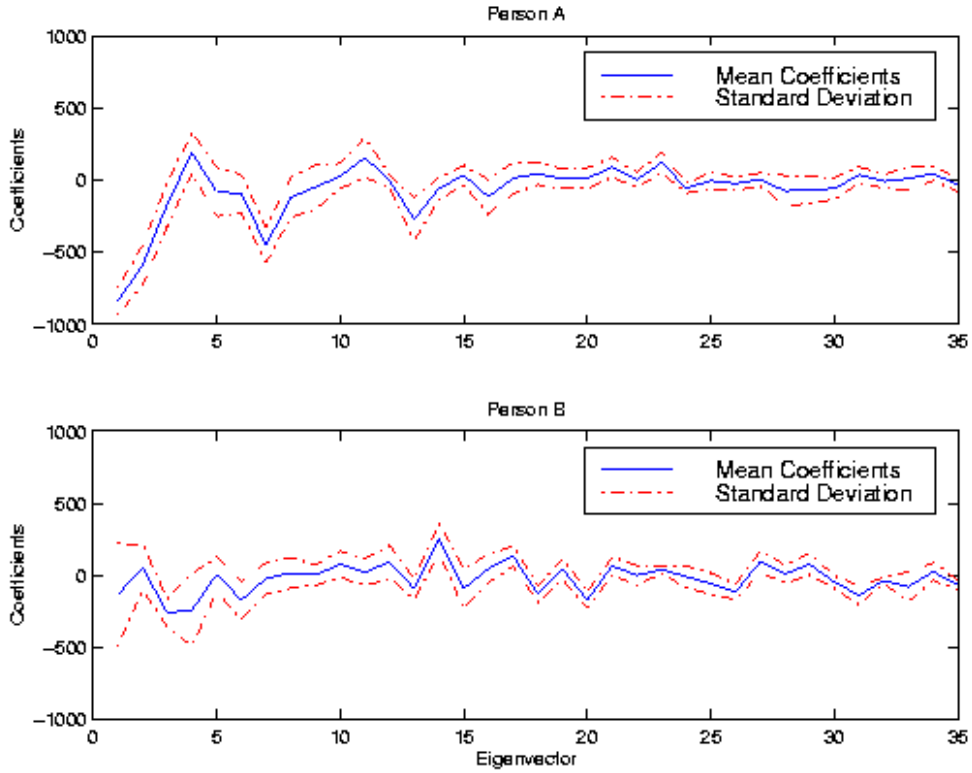


Figure 3-2: Distribution of the first 35 eigen-coefficients for person A and B

held in front of the camera, even if it is in motion, gives a flat structure. Figure 3-3 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.

3.4 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 (Ma and Gao, 1997; Bimbot et al., 1997). Even if the audio channel is kept consistent so as to minimize

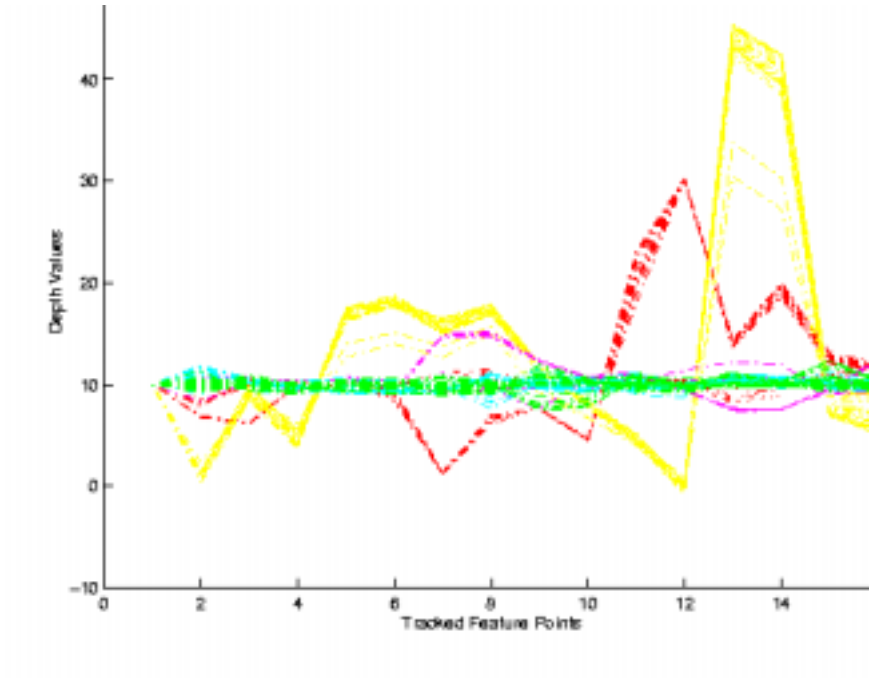


Figure 3-3: Depth values for tracked objects: the object with all of its features at the same depth is a photograph, the rest are faces.

convolutional noise, there will always be the problem of additive noise in natural scenarios.

Deconvolutional techniques such as RASTA (Hermansky et al., 1991) 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 (Bimbot et al., 1997) has suggested that the presence of noise does not 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.

3.4.1 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 (Brown, 1992)). The effect is that after a period of silence the system is hypersensitive and after a period of loud sound the system grows desensitized.

3.4.2 Feature Extraction

After segmentation, the audio (which is sampled at 16KHz) is filtered with a weak high-pass filter (pre-emphasis) 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 is useful because it allows us to detect additive noise given a model of the noise in isolation.

3.4.3 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 clustering to initialize HMM parameters and then Expectation-Maximization (EM) to maximize (locally) the model likelihood (Rabiner, 1989). 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 used the *leave one out* test on our training set for optimization purposes. 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.

3.4.4 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. In our experiments, we do not take into account the Lombard effect which can influence the speech of a person in presence of noise (Junqua et al., 1999; Zhou et al., 1999).

The model adaptation procedure uses parallel model combination (Gales and Young, 1994) and 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 (data collection is described below) and an auditory background scene of a street in a thunder storm. The noise scene contains frequent thunder and occasional 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*

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%

Table 3.1: Recognition rates for the clean speech, corrupted speech and adapted speech models.

+ Noise training set to evaluate the effectiveness of the adaptation.

Finally we tested all HMMs on both the *Speech Only* and *Speech + Noise* test sets — there was no overlap between the test and training sets. Table 3.1 contains the recognition rates for two sets of 130 audio clips. The first row shows the performance of clean speech model on test sets containing clean speech and speech with background noise, the second row shows the performance of the adapted speech model built using parallel model combination. The third row shows the performance of the model trained with the background noise — not adapted to the background noise. 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.

3.5 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)).

3.5.1 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, $\bar{x}$ is the mean image calculated from the training set and N is the total number of classes):

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

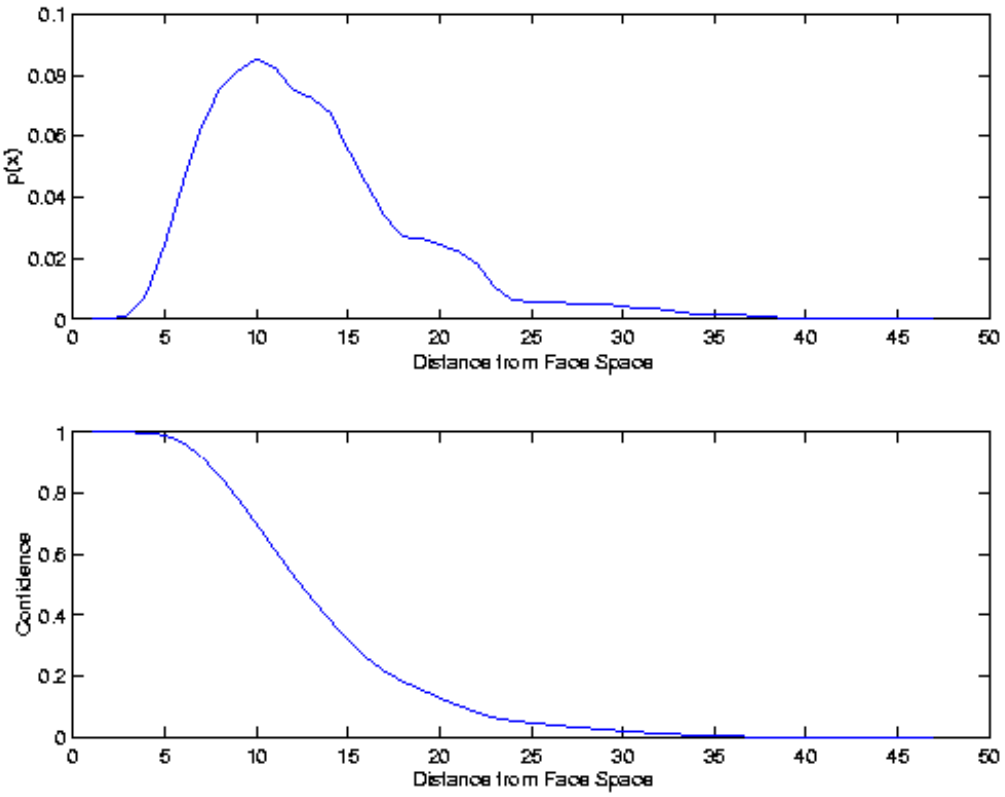


Figure 3-4: Mapping the DFFS to a probability: (top) the DFFS pdf for a set of images, (bottom) the derived cdf which is used for the mapping.

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

We estimate $p(\omega)$ from a set of images or audio clips using Parzen windows with Gaussian kernels. Figure 3-4 shows this mapping for the DFFS measure of faces of different people.

Table 3.2 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 confidence score that is high for recognized images is correlated with recognition. 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.

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%

Table 3.2: Comparison of confidence scores: Prediction rates of correct recognition (left) and wrong recognition (right).

3.5.2 Bayes Net

In the fusion of classifiers, each knowledge source may be dependent on other sources. The full Bayesian approach makes the minimal assumption that 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

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

where the r.v. $C_i = \{$reliability of classifier$\}$, and the r.v. $X_i = \{j | j \in$ Client Database$\}$. Figure 3-5 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_S and X_F . We are also assuming that the confidence scores are conditionally independent from X. We assume that the distributions of confidence scores are the same for both classifiers, i.e. $P(C_S) = P(C_F)$.

$$\begin{aligned} P(X) &= P(X|X_S)P(X_S|C_S)P(C_S) + \\ & \quad P(X|X_F)P(X_F|C_F)P(C_F) \end{aligned}$$

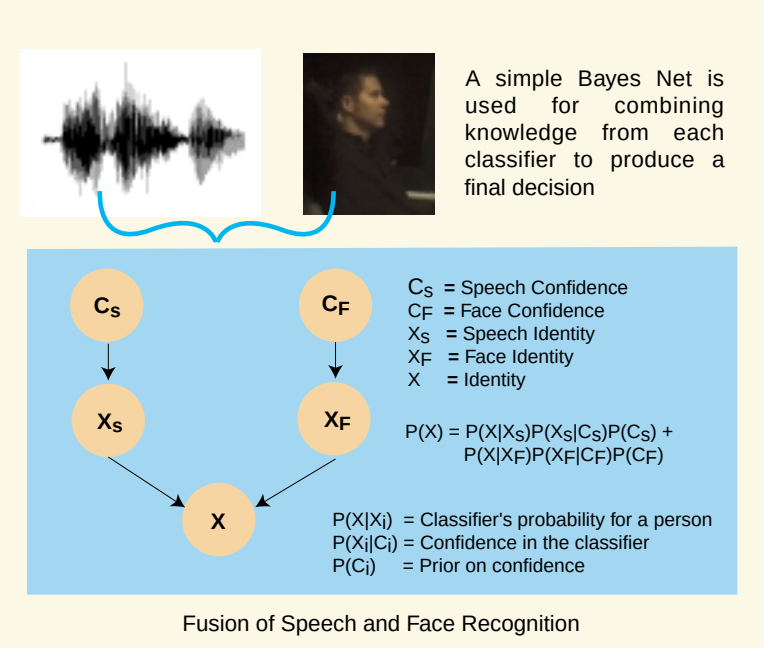


Figure 3-5: The Bayes net used for combining knowledge sources for each classifier to produce a final decision, X.

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.

3.6 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.

3.6.1 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 will not 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. We collected two sessions of data which was divided into training and testing sets of roughly equal size. Each transaction was about 20s long. The illumination condition were approximately the same for both session and there was no noise in the background at the time of recording.

3.6.2 Evaluation Methods

We evaluated our system using both recognition and verification rates. Both procedures include 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 $P(X = \text{Recognized Identity}) < \text{Rejection Threshold}$.

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 $P(X = \text{Recognized Identity}) < \text{Rejection Threshold}$.
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 correct rejection rate. For each experiment we give the success rate at both zero threshold (i.e. no rejections) and the optimal operating threshold.

3.6.3 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

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

Table 3.3: 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%)

Table 3.4: Recognition rates (optimal rejection threshold): the rejection rates are in parentheses.

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 3.3 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 3.4 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 3.5. 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.

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%)

Table 3.5: Verification rates (optimal rejection threshold): the false acceptance rates are in parentheses.

3.7 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 eigen-face 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.

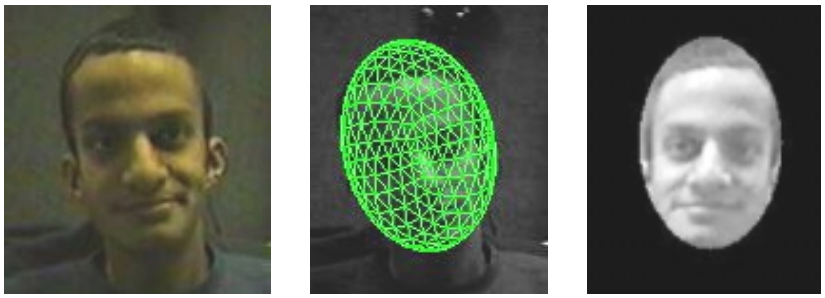
Chapter 4

Head Tracking

4.1 Introduction

People, to a certain extent, are able to understand different emotional expressions from static images. However, the dynamic evolution of images over time provides us with information about the context, flow and intensity of emotional expressions which is impossible to extract from single static images. Also, temporal relationships between different facial expressions and their duration are important information in the analysis and interpretation of facial expressions. Thus, it is essential for any system built for automatic expression analysis to track head movement in natural conditions over an extended period of time. Most robust head trackers today are 3D model based systems.

In our case, we needed to do analysis on unconstrained video with significant head movement. It is necessary to know changes in the head pose and in the facial features. We used three existing head-trackers in our group for our analysis purposes. In this section I'll discuss the pros and cons of each of these head-trackers and how feasible they would be to be incorporated in an automatic system. We ran the head-trackers on three different types of conversation sequences and noted their performance. The conversation types are — 1. Totally unconstrained sequences with lot of rapid head motion 2. Slightly restricted sequences with moderate head motion and 3. Severely restricted sequences with no head motion — for the rest of the chapter I'll call these



(a) Input image (b) Model Aligned (c) Frontal Warp

Figure 4-1: Output of the head-tracker based on regularized optic-flow

seq. 1, seq. 2, and seq.3 respectively.

4.2 Model-based head Tracking using Regularized Optic-flow

This system was developed by Basu, Essa, and Pentland (Basu et al., 1996) — they attempted to build a system that could track heads in different conditions including large head motion. Their work was a 3D extension of the tracker developed by Black and Yacoob (Black and Yacoob, 1995) using regularized optic-flow on a planar model.

The tracker uses a 3D ellipsoidal model for the head as opposed to a 2D planar model in order to track motion more accurately. It starts by computing the optic-flow for the entire image and then estimates the 3D motion of the rigid model that best accounts for the flow. These motion parameters are used to modify the location and rotation of the model and used for the next frame.

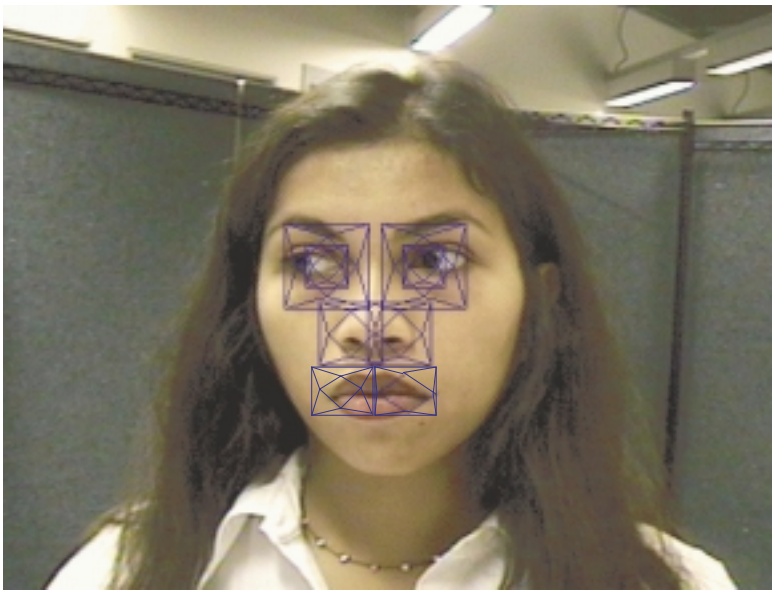
We used this tracker in all three sequences. It was able to track reliably for the first few hundred (300 - 400 frames on average) frames and then diverged significantly. Since motion on each frame depends on the accuracy of the previous frame, once an error is introduced it propagates and continues to grow. Even in seq. 3 the error was beyond acceptable limits after 500/600 frames. Thus it is unlikely the system could be used in extended sequences reliably. Also optic-flow is computationally expensive and the tracker requires a few second per frame to do the tracking. However, the

system is easy to re-initialize from the point of failure and thus useful tool for initial investigation and analysis.

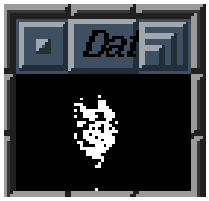
4.3 Parametrized Structure from Motion for 3D Tracking of Faces — Version 1.0

The second system used was a real-time system based on 2D feature tracking combined with an extended Kalman filter (EKF) to recover 3D structure, pose, motion and camera focal length (Jebara and Pentland, 1996). The system is real-time and automatically detects faces, facial features and computes the feature trajectories. The face is detected using skin color information. A mixture of Gaussians model is used to obtain the distribution of skin color in RGB. After locating the face, the facial features — eyes, nose and mouth — are located using symmetry transform and the image intensity gradient. The face and facial feature detection stage is not real-time and runs at 0.5Hz. Once the features are located their time trajectories are calculated using a 2D correlation based tracker and an EKF which also recovers the 3D structure and pose. The tracking stage runs at 30Hz. Eigenface projection is used to assess the reliability of the tracker output. If the output is a face it can be reconstructed using the eigenfaces and the reconstruction error will be small, whereas if the error is large it will indicate error in tracking and the system can revert to the face detection stage again.

One of the greatest advantages of this system is that the tracking stage runs in real-time. This eliminates the need to store extremely large files for processing. Also, the tracking performance can be seen instantaneously and the situations under which the tracker fails can be evaluated. However, the systems also had trouble tracking sequences with significant head motion. In seq. 1 whenever rapid head motion occured it needed to be re-initialized. Frequent re-initialization of the tracker is not ideal if temporal relations of expressions are to be measured. The output of the tracker, the position normalized frontally warped image is small (40x80) to ensure real-time performance. Thus, information about changes in facial features is more difficult to



(a) Input image



(b) Skin Blob (c) Frontal Warp

Figure 4-2: Output of the head-tracker based on parametrized SFM

obtain.

4.4 Parametrized Structure from Motion for 3D Tracking of Faces — Version 2.0

The third tracker, developed by Strom and Pentland (Strom et al., 1999) is built up from the previous tracker. It works almost in the same way with only a change in the feature points that are used for tracking. Instead of using a set of predefined features, namely the eyes, nose and mouth, the system selects features reliable for tracking based on the input image Hessians.

This system is still under development and was not used extensively on the

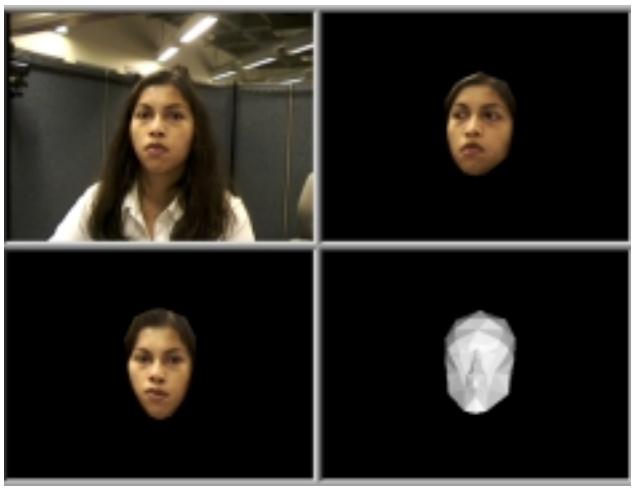


Figure 4-3: Output of the head-tracker based on SFM - clockwise from top left (a) input image (b) tracked head (c) candide head model (d) frontal warp

recorded coversation sequences. Initial testing showed that it performed well on sequence 2 and 3, but had problems with sequence 1 when rapid head motion occured. However, this system is expected to be more robust against head movements as it will not ultimately rely on a fixed set of feature points all the time but instead adapt to points that are more reliable to track — e.g., if a head rotates the system might rely on points from the ear for tracking rather than points from the occuled eye.

4.5 Conclusion

Eventually, the expression analysis system will require a real-time system that can robustly track the head in unconstrained environments. Tracker 1 is not suitable for such tasks. Both Tracker 2 and 3 have the potential to be applied to such tasks. Tracker 3 takes into account the problems faced by tracker 2 — i.e. fixed feature point for tracking in an extended sequence, and is mostly likely to be incorporated in the final system.

Chapter 5

Data Collection

5.1 Introduction

The design of the data collection process is very important in getting the span of emotional expressions on record. The first requirement was not to constrain the subject in any way. The cameras and microphones should be placed in a way that was very unobtrusive, yet able to capture the changes in facial and vocal expressions. To fulfill this requirement, we designed and built a conversation space which is described in the following section.

5.2 Equipment Setup for Conversation Space

The resources used to build the conversation space are :

1. Three Sony EVI-D30 pan-tilt cameras.
2. Two AKG 1000 cardioid microphones
3. Quad splitter
4. Hi8 video recorder
5. TV monitor

(a) The conversation Space

(b) A typical conversation scene

Figure 5-1: Conversation Space (CSpace) Set-up

6. Table and a set of chairs

7. Cables, connectors, camera mounts etc.

Two of the cameras were directly behind and slightly above each subject to get a frontal view of the subjects participating in the conversation. To capture the entire scene, a wide angle lens was attached to the third camera and placed in the center corner . Each microphone was directed towards a subject. The outputs of the cameras were fed into a quad splitter and the microphones to a mixer. Both the quad splitter and the mixer were connected to the recorder which recorded the time synchronized output of the two devices. The TV screen was used to monitor the recording process and to detect any problems.

The data collection was done in three stages. Initial data recordings were done in a totally unconstrained setting where the subjects were allowed to move around freely. However, these recordings were impossible for the trackers to track for an extended amount of time — usually whenever a rapid or abrupt head motion occured the trackers failed and needed to be re-intialized. At the next stage the subjects were asked to refrain from moving their heads too much or too rapidly. Finally, a sequence was obtained by locking the head at a fixed position and thereby allowing no head movement at all.

5.3 Questionnaire

Ideal recorded sequences should contain individuals demonstrating their entire span of expression. Though unlikely to acquire this in a short span of time the structure of the conversation should be such that it elicits a variety of emotional expressions and conversational cues. Thus, the theme of the conversation was guided but not the specific content of the conversation. For example, two friends were asked to talk about something they find interesting for the next ten minutes. In another setting one person asked a fixed set of questions to another to induce certain type of moods, e.g. happy, sad, etc. (Brewer et al., 1980; Hsee et al., 1991; Martin, 1990). Finally, the subjects were requested to make specific expressions, like happy, sad, etc., on demand.

Outline of the types of questions asked of subjects:

1. **Background/Control questions** — These were questions about the subject's name, place of work, and so on, to make them feel comfortable and provide data as to the nature of the subject's expressions and conversational structure in a neutral condition.

2. **Questions to elicit an affective response** — The intent of these questions was to have the user give expressive dialogue on a subject of their choice - example questions would be, "Describe a very pleasant experienc" or "Describe

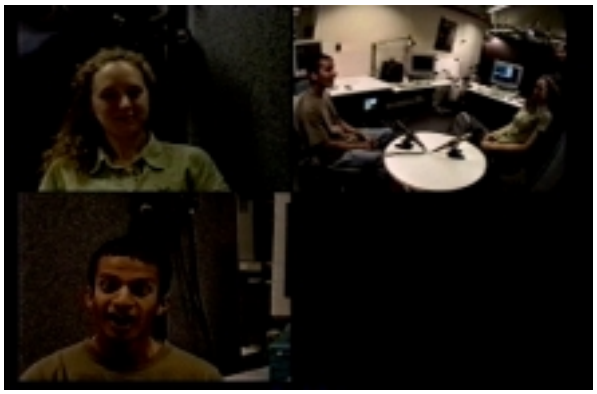


Figure 5-2: Input to the Hi8 recorder from a typical CSpace conversation

how you dealt with a stressful situation"

These questions were interspersed with neutral questions, such as those in section 1, in order to balance out the user's affective state before moving to a new topic.

3. **Labeling of subject's responses** — At the end of the interview, we ask the subjects to label their affective states during the course of the interview. They do this by reviewing a videotape of the interview and writing down a one-word description of their state during each minute.

Example questions for types 1 and 2:

1. What is your name ?
2. How are you doing today ?
3. We are working on building models of human expression and conversational structure in context of face-to-face conversations. I'll be asking you questions on different topics. If you do not feel comfortable answering any questions, let me know and we will switch topics.
4. Tell me about a pleasant incident in your life that made you very happy.
5. So how was your day today ?
6. How many classes did you have today ?
7. Tell me about a time when you felt very frustrated.
8. Are you going home for the holidays ?
9. Where is your home town ?
10. Tell about one of the most exciting events in your life.
11. In order to compare expressions by request with natural conversation will you please make a sad expression

12. Now a happy expression

13. An angry one

14. A frustrated expression

5.4 Consent Form

The thesis involves the participation of human subjects and recording of their conversations. In accordance to the regulations of the MIT Committee on the Use of human Experimental Subjects (COUHES), I obtained a written consent from the participants. The consent form provided to the subjects is included in Appendix A.

Chapter 6

Feature Extraction

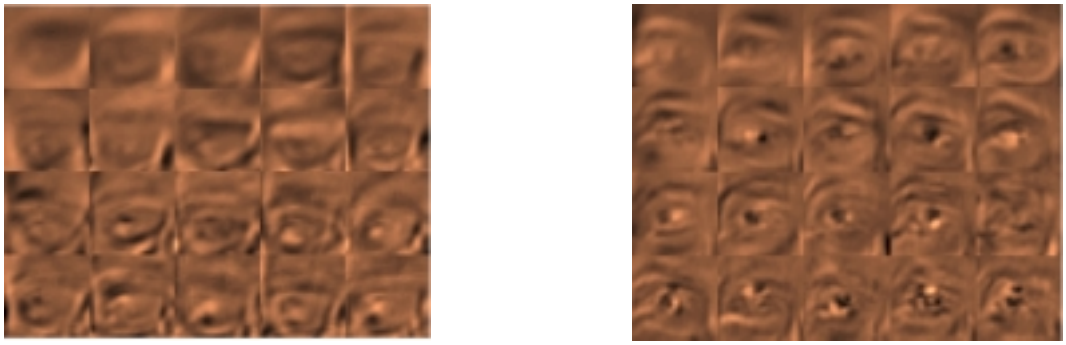
6.1 Introduction

The head-tracker gives the position and orientation of the head over time. We use the output to normalize and warp the face to a frontal position. Then the most expressive regions of the face are automatically cropped out for further processing. Evidence suggests that when people are engaged in a conversation the most frequently occuring movements are raising the eye-brow, lowering the eye-brow and some form of smiling (Hager and Ekman, 1996). Thus, I extracted features from the eyebrows, eye and mouth. I did analysis using three types of features extracted from the selected regions — principal component analysis (PCA), optic flow and local histogram.

In this chapter, I shall describe in detail the different features extracted and the advantages/disadvantages of using the different features.

6.2 Eigen Features

Experiments using eigen-faces for expression recognition (Pentland et al., 1993) have been done previously, and 85% accuracy was obtained for person dependent expression models using perfectly normalized images. Eigen-features will perform better than eigen-faces, because they capture changes in individual facial features rather than combined changes occuring in the face. As the initial intent is to focus on



(a) Eigen-features from normalized image (b) Eigen-features from after re-normalization

Figure 6-1: Sensitivity of eigen-features to changes in position

person-dependent expressions, we used eigen-features as our first approach. Once we calculated the eigen-templates for the facial features of interest, we computed the eigen-coefficients for these parts in a given sequence, which was used as the input feature for analysis.

The major disadvantage of using eigen-features calculated from raw images was its sensitivity to precise normalization. Although the head was normalized to a standard size and location, the features, i.e. eyes, brows and mouth normalization, were not precise in every frame. Thus, in a lot of cases the coefficients capture more of the changes in the feature locations rather than changes in the features themselves. For example, changes in the location of the center of the eye were captured as opposed to changes in eyebrow due to a brow raise. None of the head trackers were able to give

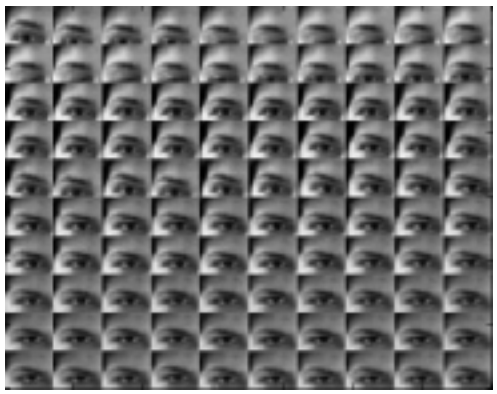


Figure 6-2: Typical normalization error in a sequence

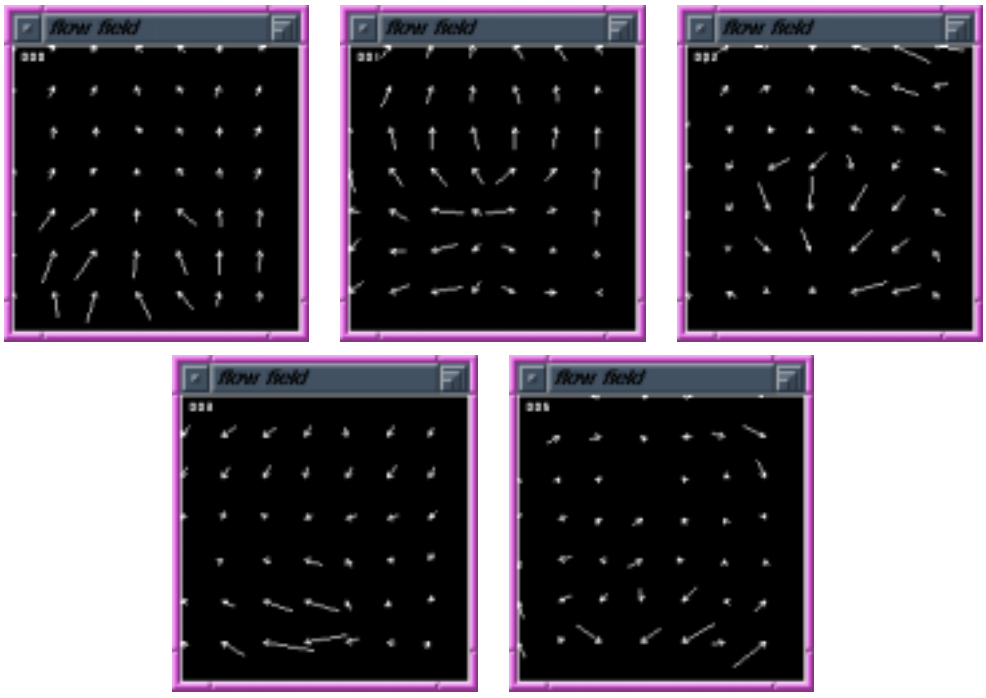


Figure 6-3: Eigen-vectors calculated from optic-flow of the mouth region

precisely normalized feature locations for the eigen-features to capture subtleties like changes in the eye-ball location or brow raise versus lower and mouth open versus smile etc. Figure 6-1(a) shows the eigen-features calculated from the output images of the head-tracker and Figure 6-1(b) are calculated after re-normalization of the output — it can be seen that Figure 6-1(b) captures more of the changes in the eye and eye-brow.

6.3 Optic Flow

Our goal is to capture changes in features of different regions of the face, thus we calculated the optic-flow of the selected regions. Optic-flow is sensitive to the motion of image regions and the direction in which different facial features move.

First, we computed the dense optic flow of the feature images cropped from the normalized face image. Then we calculated the eigen-vectors of the covariance matrix of the flow images. The top 10 eigen-vectors were used to represent the motion

observed in the image regions. These captured 85% of the variance. The eigen-coefficients computed for each image region were the new input features. Optic-flow was able to capture more of the variations in the features which were missed by the eigen-features obtained from the raw images. However, large tracking errors leading to improper normalization can cause optic-flow to provide misleading results.

6.4 Multidimensional Receptive Field Histograms

Lack of accuracy in normalization of face images leads to difficulty in recognition of changes in facial features. Thus, features that are less sensitive to small position and scale changes are likely to prove more reliable for our task. For this reason we uses an object recognition system proposed by Schiele and Crowley (Schiele and Crowley, 1996). They build a statistical model for an object using local object descriptors. This has proven to be a reliable representation for objects and considerably robust to scale and position changes.

Objects are represented as multidimensional histograms of vector responses of local operators. Gaussian derivatives are easily generalizable and robust, thus the multidimensional vectors calculated are the Gaussian derivatives and Laplace operators at two or three scales. In my experiment, I used only the first derivative and the Laplacian. Recognition is performed by matching of the histograms using χ^2 statistic, histogram intersection, or mahalanobis. The resolution of the histogram axis was either 16 or 32 pixels. For more detail on this method please refer to (Schiele and Crowley, 1996).

The probability of an object O_n given local measurement M_k is:

$$p(O_n|M_k) = \frac{p(M_k|O_n)p(O_n)}{p(M_k)}$$

where $p(O_n)$ is the prior probability of the object which is known and $p(M_k)$ as the prior probability of the filter output which is measured as $\sum_i p(M_k|O_i)p(O_i)$. If we have K independent measurements $M_1, M_2, ..., M_K$ then the probability of the object O_n is:

$$p(O_n|M_1, M_2, ..., M_K) = \frac{\prod_k p(M_k|O_n)p(O_n)}{\prod_k p(M_k)}$$

Since, the measurement locations can be chosen arbitrarily and only a certain number of local receptive field vectors need to be calculated, the method is fast and robust to partial occlusion. This method can also work without correspondence between the object database and the test image (Schiele and Crowley, 1996).

6.5 Conclusion

Although initially we assumed that eigen-features calculated from the raw normalized image would be a good enough feature to capture changes in the eye and mouth region, it proved not to be the case. These features were too sensitive to position and scale variations in the normalization process. However, if the head-tracker is robust and performs with minimum error this would be a good feature to use. Optic-flow does capture the motion in the image and consequently the changes in the features. But, this does not take into account the structure of the features like the previous method. Some knowledge or dependency on the structure might prove useful for fine scale analysis of facial expressions. The multidimensional histograms have been shown to be very successful in object recognition even in the presence of occlusion. It is relatively insensitive to scale and position changes thus a suitable feature to use for the proposed task. Initial testing done using this method on simple mouth and eye expressions were encouraging. However, a thorough analysis of this method and its effectiveness in distinguishing between very fine and subtle changes in the same facial feature have not been done in this thesis. This is a promising line of investigation for future research.

Chapter 7

Feature Modeling

7.1 Introduction

After the feature extraction stage comes the modeling stage — in this stage we would like to capture the relationships among feature movement in expressive gestures. For example, an eye-blink always consists of an open eye gradually closing and then opening again — there is an explicit temporal pattern in the eyelid motion. There are many such cases where there is a clear temporal pattern, e.g during a smile, raising of eyebrows etc. It is important to capture these expressive patterns to build models for a person's facial expressions. However, a smile or an eye-blink does not provide enough information about emotional expressions — it is the combination and temporal relationship between the short time expressions that can explain what a person is trying to convey through his/her facial expressions.

7.2 Temporal Clustering Using Hidden Markov Models

Following widespread use and success of hidden markov models (HMMs) in speech recognition technology, HMMs are being widely used in computer vision for gesture and expression recognition. Expressions have explicit temporal patterns and HMMs

are particularly useful for this task because they code the temporal sturcture of the observations by producing first order Markov Models. Each state in an HMM is associated with the observation probability distribution $b_j(x)$ and the probability of making a transition from state i to state j in one time step is denoted as A_{ij}. Thus, the structure of the models can also be constrained by constraining the state transition matrix A of an HMM. For a tutorial on Hidden Markov Models please refer to (Rabiner, 1989).

The duration of each state of an HMM is guided by its frame/state allocation. Varying the frame/state allocation number guides the clustering algorithm to model time-series at varying time scales. The frame/state T is provided explicitly at the intialization state such that the HMM initally has a duration of TS samples where S is the number of states for the model. Expectation Maximization (EM) is used for estimating the final parameters from the initial guess. The model duration is changed during the re-estimation stage, but the change is not drastic because of the local optimization property of EM. The length or time-resolution of a model can thereby be controlled somewhat by changing the frame/state allocation number (Clarkson and Pentland, 1998).

7.3 Modeling of Short-time and Long-time Expressions

To model emotional expressions, we need to model the long-time expressions as well as the short-time expressions that make up the longer and more expressive events. Thus we need to model our data at different time resolutions. The temporal clustering of events at different time-scales is based on the system developed by Clarkson and Pentland (Clarkson and Pentland, 1998). They propose a hierarchy of HMMs to capture relationships between expressive motion at different time granularities.

7.3.1 Short-time Events

In our first attempt at modeling short-time events, we only specified the time resolution of the HMM states and the number of HMMs we wanted to model — a totally unsupervised learning scheme. However, the output was not satisfactory — models tended to localize in time, i.e. each HMM captured events that were close together in time rather than events that were similar in terms of feature movement. For modeling the short-time events, we selected in advance the expression that we wanted to capture and trained our HMMs based on that. The algorithm used by Clarkson and Pentland was as follows:

1. We choose the number of HMMs (N), number of samples allocated to a state(T) and number of states per HMM (S).

2. Initialize N HMMS of length TS using linear state segmentation (Clarkson and Pentland, 1998).

3. Compile the N HMMS using a fully connected grammar and then re-segment cluster membership for each HMM. The HMMs are constrained to be left-right with no jump states and the output of each state is modeled by a single 10 dimensional Gaussian with diagonal covariance matrix.

4. Estimate HMM parameters using the Forward-Backward algorithm. Then iterate on the segments until the HMMs converge and then go back to step 3. Repeat steps 3 and 4 till models converge.

7.3.2 Long-time Events

Clustering events at different time-scales will not always capture the structure of expressions. For example, person X when annoyed always furrows his eyebrows and then tightens his lips, or person Y when agreeing with someone always smiles and then nods her head. Events like this that are a sequence of temporally structured events, need to be first identified individually and then as sequences with specific

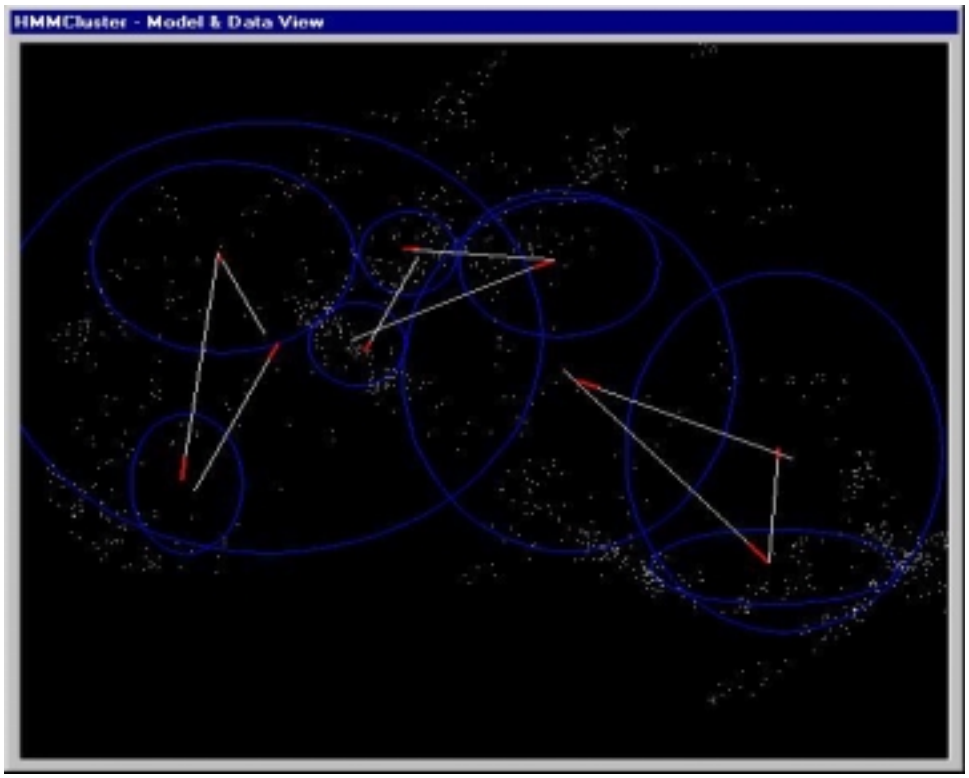


Figure 7-1: A Typical structure - 3 HMMs each with 3 states

temporal relationships. Changing the time-scale of HMMs cannot model such events — so a hierarchy of HMMs much like a grammar is useful to capture these structures (Clarkson and Pentland, 1998). Relationships of short-time scale HMMs can be encoded in long-time scale HMMs — the algorithm proposed by Clarkson and Pentland to encode the relationships is used:

1. Obtain a likelihood $L_\lambda(t) = P(O_t, ..., O_{t+\triangle t}|\lambda)$ for each HMM λ using the Forward algorithm and a sliding window of $\triangle t$.

2. Construct a new feature space based on the likelihoods,

$$F(t) = \begin{bmatrix} L_1(t) \\ \vdots \\ L_N(t) \end{bmatrix}$$

3. Use the same technique used to model short-time HMMs to obtain long-time HMMs using the new feature.

7.4 Conclusion

In this chapter we described the algorithms that will be suitable for modeling facial expression. In the thesis we explored variations in short-time expression modeling. All the modeling described uses a single Gaussian with a diagonal covariance matrix for each output state of the HMMs. This proved not to be quite sufficient for capturing the structure of some expressions. So, currently we are looking into modeling the feature space using HMMs whose output states are Gaussian mixture models.

Chapter 8

Results and Conclusion

8.1 Introduction

In this chapter, I present the results obtained from my experiments. I shall elaborate on the methods used and circumstances under which the results were obtained. I shall discuss some of the observations made from the recordings that have not yet been automatically detected but provide insight on how natural conversations capture crucial information that is not obtained from restricted or *expression on demand* type of experiments. I shall also discuss some of the strategies I plan to take in the future to continue onwards with this work.

8.2 Results

Unsupervised clustering of features did not give any useful and consistent expressions — mainly because of the insufficient of training data and also because the modeling technique tends to cluster events localized in time. Thus, if a certain expression occurs in significant time gaps, it does not get clustered into the same event.

To avoid this problem and yet validate the techniques used, I selected segments that occured enough times to build models for them. All the models were person specific, and I trained HMMs on selected segments only. Testing and training sequences were obtained from two different data recording sessions. For eyes, the most common

Expression	All Sequence Lengths	Sequence Length > 3 frames
Blink	81.82 %	92.73 %
Open	75.00 %	91.96 %

Table 8.1: Experiment 1 — Recognition rates for eye-expressions

Expression	All Sequence Lengths	Sequence Length > 3
Blink	67.71%	85.42 %
Open	89.17 %	96.67 %

Table 8.2: Experiment 2 — Recognition rates for eye-expressions

patterns were blinks and long segments where the eyes are open. So, tests were done using blink and open eye expressions — I used 30 blink sequences for training and 34 for testing and 30 open eye sequences for training and 37 for testing. The first experiment was using two three state HMMs for each open and blink expressions. All the HMMs had a frame/state allocation of 15. The second experiment had HMMs of varying lengths — for blinks I had 4 models, one three state HMM with frame/state = 10, one 5 state HMM with frame/state = 10, two 5 state HMMs with frame/state = 15 — for open eye expression I also had 4 models, two 5 state HMMs with frame/state = 10, one 4 state HMM with frame state = 15 and one 5 state HMM with frame/state = 15. Table 8.1 and Table 8.2 show the recognition rates for the two experiments.

Mouth expressions were much harder to recognize. I selected segments containing open-to-close, close-to-open, and smile sequences. However, each type of sequence had variations - in term of intensity, teeth showing vs. teeth covered and the models did not provide consistent result. It seems that each HMM models a large expression space that incorporates multiple expressions. I used 19 open-to-close sequences for training and 10 for testing and 21 close-to-open sequences for training and 9 for testing. Experiment 3 used 4 single state HMMs with frame/state = 25 — the first two HMMs were trained on open-to-close sequences and the remaining two were trained on close-to-open sequences. Instead of presenting recognition rates, I shall provide tables (Table 8.3 and Table 8.4) with the HMMs and the expressions they

HMM	Close	Open	Open-to-Closed	Closed-to-Open
0	2	2	22	24
1	3	12	4	6
2	2	2	3	1
3	-	1	-	-
4	-	1	-	-
5	2	2	1	1
6	1	3	1	-

Table 8.3: Experiment 3 — Model classification for mouth expressions

HMM	Close	Open	Open-to-Closed	Closed-to-Open
0	1	4	7	16
1	3	17	21	18
2	-	2	-	-
3	2	9	2	3

Table 8.4: Experiment 4 — Model classification for mouth expressions

classified. For example, HMM 0 classified 2 closed mouth, 2 open mouth, 22 open-to-close and 24 close-to-open expression — this implies that HMM 0 is too broad a model and is capturing most of the data. Ideally, we would want each HMM capturing only one expression. The HMMs also tend to break the sequences into subsequences and thus although trained on open-to-close and close-to-open sequences, some HMMs classify pure open and close subsequences.

Experiment 4 used 7 two state HMMs with frame/state = 25 — the first three were trained on open-to-close sequences and the remaining four were trained on close-to-open sequences. Again, we observe that HMM 1 is too broad and is not selectively classifying one type of expression.

I also did a preliminary experiment using local receptive field histograms on static images. The experiment used 200 labeled images of the eyes, 100 for training and the other 100 for testing. The labels were for open, semi-open and closed eyes and the recognition rate for the test set was 92% using histogram intersection measurement (Schiele and Crowley, 1996). I later incorporated a larger database of labeled images

Eye-Expression	Close	Open	Semi-Open
Intersection	93.49%	75.71%	8.70%
Mahalanobis	82.00%	69.24%	18.84%

Table 8.5: Recognition rates for eye expressions using receptive field histograms

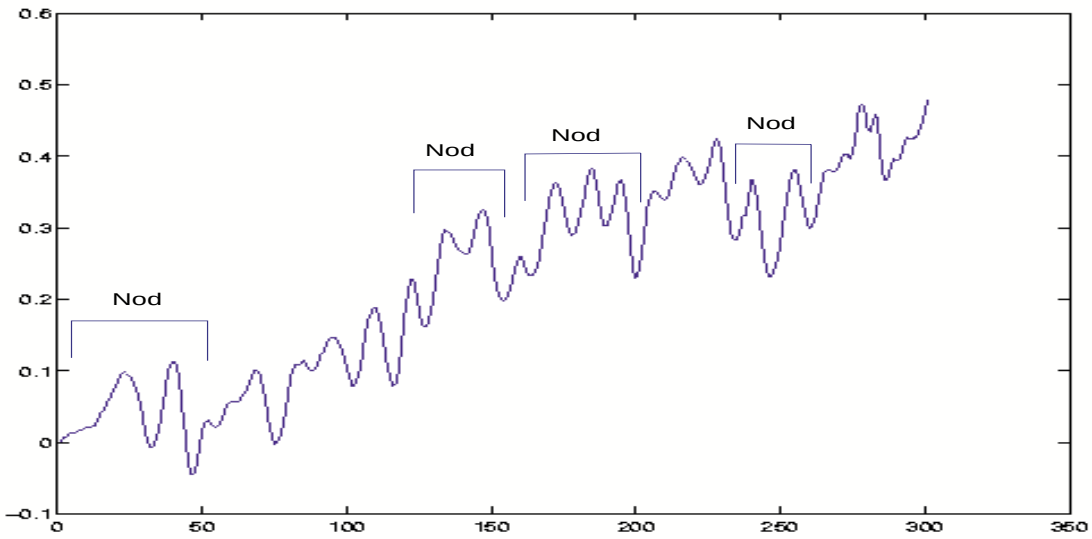
Mouth-Expression	Close	Open	Smile	Tight-lipped
Intersection	17.72%	72.85%	66.98%	49.55%
Mahalanobis	12.03%	76.16%	20.95%	10.81%

Table 8.6: Recognition rates for mouth expressions using receptive field histograms

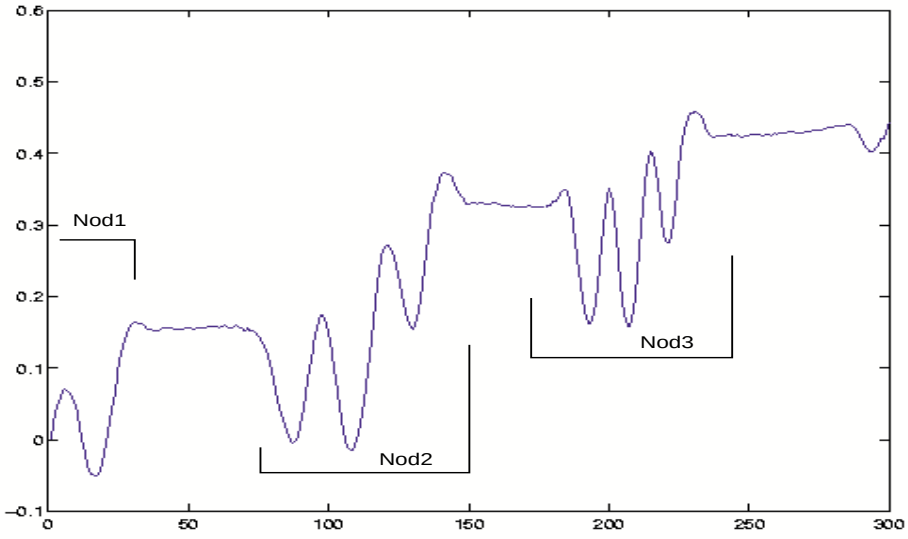
consisting of 1000 training and 1000 testing images — the recognition rate for this set was 66.9% using the mahalanobis distance and 99.69% within the top 7 matches, and 74.30% using intersection measurement with 99.00% within top 4 matches. For the mouth images again, the results were much worse — training and testing were done on a database of 1000 images each and images were divided into 4 classes — open, close, smile and tight-lipped. The recognition rate using the intersection measurement was 34.80% and 88.40% within the top 2 matches. Using mahalanobis distance the recognition rate was 29.50% and 65.30% using top 2 matches. The above recognition rates are the combined rate for all eye expressions and all mouth expressions — Table 8.5 and Table 8.6 show the recognition rate for each expression separately.

Apart from temporal modeling using HMMs, it is possible to get additional information from the trajectories of the rotational parameters, alpha, beta, and gamma (rotation about x,y,z axis) obtained from the 3D head tracker.

For example, a nodding or shaking of the head can be detected by measuring the periodicity of the gamma and beta parameters respectively. Figure 8-1 shows the trajectory of the gamma parameter during two nodding sequences. Figure 8-1(a) is a sequences with continuous head nodding and Figure 8-1(b) is a nodding sequence where each nod is followed by a pause — notice how easily this can be detected from the trajectory of the rotational parameter gamma. Similarly Figure 8-2 shows the trajectory of the beta parameter during two head shaking sequences —

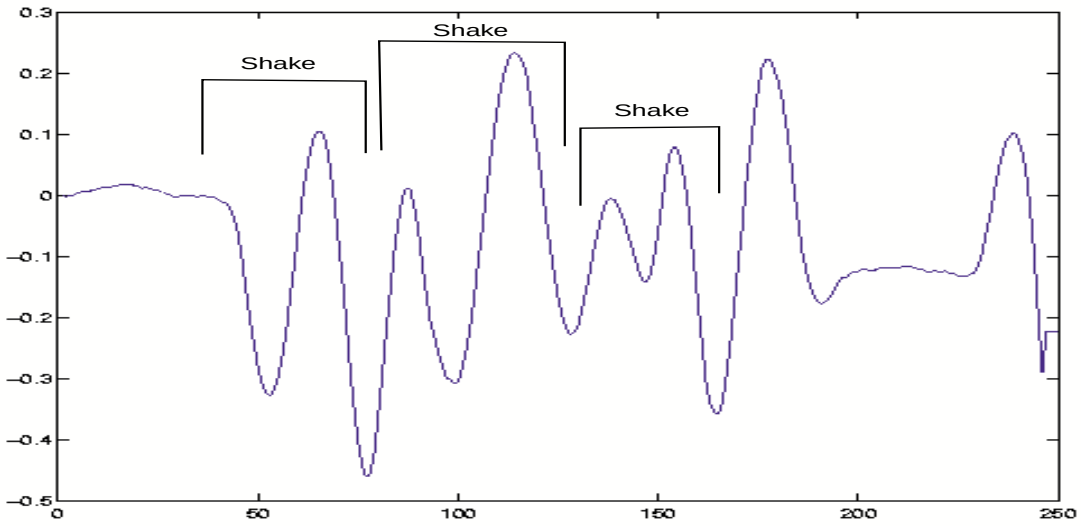


(a) Nod Sequence 1 — Continuous nodding

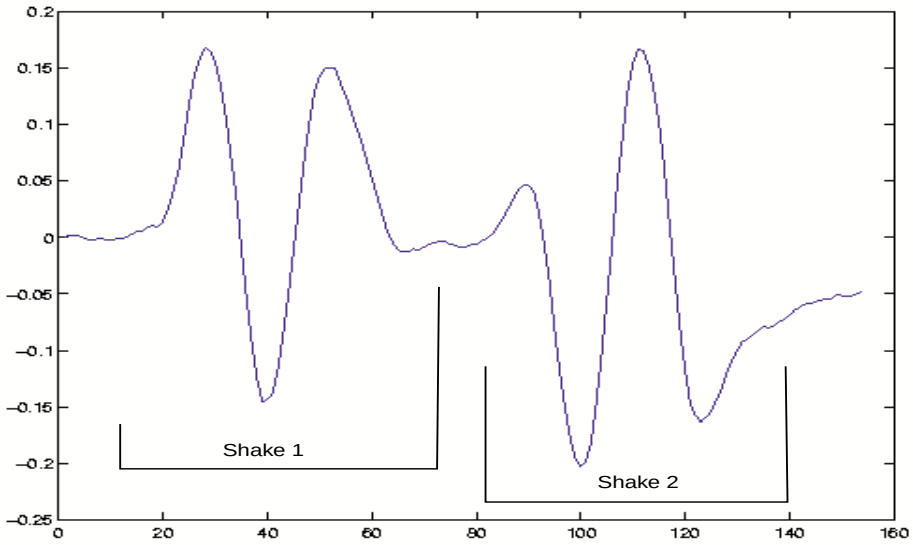


(b) Nod Sequence 2 — Nodding with intermediate pauses

Figure 8-1: Time trajectory of the gamma parameter for sequences containing nods



(a) Shake Sequence 1 — Continuous head shaking



(b) Shake Sequence 2 — Head shaking with intermediate pause

Figure 8-2: Time trajectory of the beta parameter for sequences containing head shakes

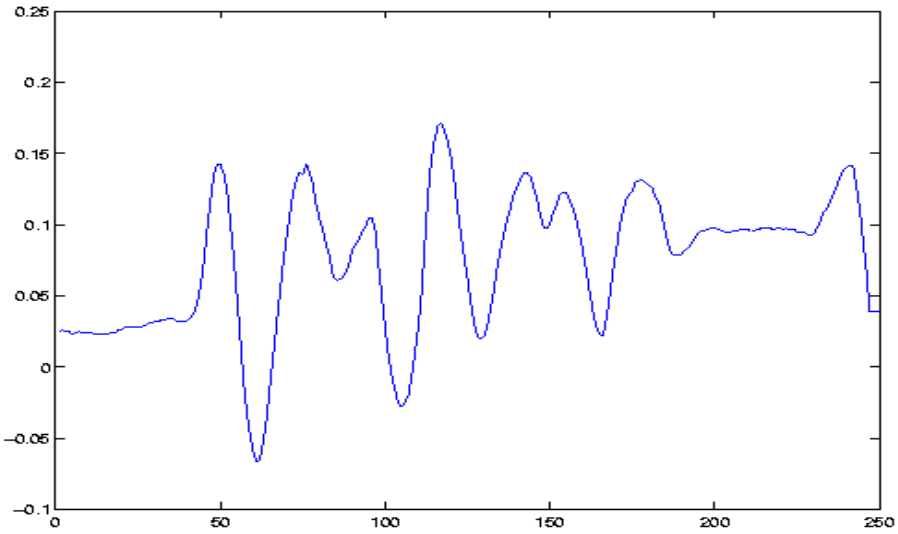


Figure 8-3: Time trajectory of the alpha parameter for Shake Sequence 1

where Figure 8-2(a) shows a continuous shaking sequence and Figure 8-2(b) shows a shaking sequence with pause in between. In the first shaking sequence there was also rotation about the x axis which can be seen from the alpha parameters Figure 8-3. However, work on an automatic detection of these expression is in progress and not yet complete.

8.3 Observations

At this stage, the higher level structure of emotional expressions has not been modeled. However, some interesting observations were made about the structure, which validates the importance of using natural conversation data and provides motivation for future work towards building higher level expression models. Some of these observations are probably person specific and are likely to give information about the particular person's interaction patterns. For example, recording showed that person X had a higher blink rate while listening to someone as opposed to talking, and also had significantly more head movement while talking. Another piece of interesting information obtained from the recordings with varying degree of constraint in head

movement was the intensity of expression, which went down when a person was asked to keep his/her head completely still — also the expressions seem less genuine or spontaneous when compared to recordings without any constraints on movement. Although testing across a wide group of people has to be done ensure consistency in these observations, this does provide motivation to work with natural interaction data.

8.4 Conclusion and Future Work

This thesis lays the initial foundations and provides the motivation for the study and analysis of facial expressions in natural conversations. I provide a framework for the study of facial expressions, starting from data recording to feature extraction and modeling, and explore the advantages and disadvantages of various recognition and modeling methods in natural, unconstrained environments. The results presented in this thesis are preliminary — however, I plan to continue and build on the work in the future. Amongst the possible approaches to explore are using 3D models for the facial features similar to the lip model built by Basu and Pentland (Basu and Pentland, 1998), building spatio-temporal models using HMMs, whose output are Gaussian mixture models with full-covariance matrices, analyzing of the receptive field histogram clusters to generate class labels as opposed to having predefined labels etc. This thesis is the first step towards a comprehensive study of facial expressions occuring during natural interactions in order to build an automatic and robust system, capable of detecting, recognizing, and understanding expressions in real-time.

Appendix A

Consent Form

I understand the participation in this experiment is completely voluntary and I can withdraw from the experiment at any time.

In this session I'll be asked questions that elicit affective response. I will be asked to talk about situations when I felt satisfaction, enjoyment or happiness, and also felt frustration, anger or disappointment. I understand that the purpose of this experiment is to gather data that will enable the experimenters to investigate ways to build models of human expressions and conversational structure in context of face-to-face conversations. I also understand that the entire session will be videotaped and be used for research purposes only. Each session will be between 15 - 30 minutes long. Sessions will be taped in E15 - 384, 20 Ames Street, Cambridge, MA 02139. Only one session per subject is necessary.

No harmful effects are expected. However, if I feel uncomfortable I can request to discontinue participation at any point and the data collected from me shall not be used.

If I have any questions about the experimental procedure or further questions about how the data will be used, I will feel free to ask the experimenters.

In the unlikely event of physical injury resulting from participation in this research, I understand that medical treatment will be available from the M.I.T. Medical Department, including first aid emergency treatment and follow-up care as needed, and that my insurance carrier may be billed for the cost of such treatment. However,

no compensation can be provided for medical care apart from the foregoing. I further understand that making such medical treatment available; or providing it, does not imply that such injury is the Investigator's fault. I also understand that by my participation in this study I am not waiving any of my legal rights.

I understand that I may also contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T. 253-6787, if I feel I have been treated unfairly as a subject." Consent forms used in cooperating institutions must assure that the rights of the subjects are protected at least to the same degree.

*Further information may be obtained by calling the Institute's Insurance and Legal Affairs Office at 253-2822.

I hereby give consent for the data (audio and video) collected from the experiments in which I have participated to be used in research papers, presentations, and demonstrations. I understand that my name will be withheld from all such displays. Furthermore, if I wish to keep any part of the experiment from being made public, the experimenters will fulfill such requests. I understand that after the study is complete, the audio/video data will be archived. If other researchers wish to use this data in the future, they will have to abide by the conventions set forth in this document.

Name of the subject:

Signature of the subject:

Date:

Bibliography

Bartlett, M., Viola, P., Senjowski, T., Golomb, B., Larsen, J., Hager, J., and Ekman, P. (1996). Classifying facial action. *Advances in Neural Information Processing*, 8.

Bassili, J. (1979). Emotion recognition: The role of facial movement and the relative importance of the upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–2059.

Basu, S., Essa, I., and Pentland, A. (1996). Motion regularization for model-based head-tracking. In *13th International conference on Pattern Recognition*, Vienna, Austria.

Basu, S. and Pentland, A. (1998). 3d lip shapes from video: A combined physical-statistical model. *Speech Communication*, 26:131–148.

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

Black, M. and Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric model of image motion. In *Proceedings of the International Conference on Computer Vision*, pages 374–381, Cambridge, MA. IEEE Computer Society.

Black, M. and Yacoob, Y. (1997). Recognizing facial expressions in image sequences using local parameterized models of image motion. *Int. Journal on Computer Vision*, 25(1):23–48.

Brewer, D., Doughtie, E., and Lubin, B. (1980). Induction of mood and mood shift. *Journal of Clinical Psychology*, 36:215–226.

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

Chen, L. and Huang, T. (1998). Multimodal human emotion/expression recognition. In *International Workshop on Automatic Face and Gesture Recognition*, pages 366–378, Nara, Japan.

Choudhury, T., Clarkson, B., Jebara, T., and Pentland, A. (1999). Multi-modal person recognition using unconstrained audio and video. In *Proceeding of the Second International Conference on Audio- and Video-based Biometric Person Authentication*, pages 176–181, Washington, D.C.

Clarkson, B. and Pentland, A. (1998). Unsupervised clustering of ambulatory audio and video. Technical Report 471, MIT Media Laboratory.

Cohn, J. and Katz, G. (1998). Bimodal expression of emotion by face and voice. In *The Sixth ACM International Multimedia Conference*, Bristol, England.

Cohn, J., Zlochower, A., Lien, J., and Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Journal of Psychophysiology*, 36(1):35–43.

Colmenarez, B. F. A. and Huang, T. (1998). Mixture of local linear subspaces for face recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 32–37.

Darwin, C. (1872). *The Expression of Emotion in Man and Animal*. New York: Philosophical Library.

Desilva, L., Miyasato, T., and Nakatsu, R. (1997). Facial emotion recognition using multimodal information. In *IEEE Int. Conf. on Information, Communication and Signal Processing*, pages 397–401, Singapore.

Duchenne, G.-B. (1990). *The Mechanism of Human Facial Expression.* Cambridge University Press. Translation of: Mecanisme de la Physionomie Humaine.

Ekman, P. (1982a). *Emotion in the Human Face.* Cambridge University Press.

Ekman, P. (1982b). *Handbook of Methods in Non-verbal Behavior Research*, chapter Methods of Measuring Facial Actions, pages 45–90. Cambridge University Press.

Ekman, P. (1992a). An argument for basic emotion. *Cognition and Emotion*, 6(169-200).

Ekman, P. (1992b). Facial expression of emotion: New findings, new questions. *Psychological Science*, 3:34–38.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, pages 384–392.

Ekman, P. and Friesen, W. (1975). *Unmasking the Face.* Prentice-Hall.

Ekman, P. and J., B. Transient myocardial ischemia — relationship of facial behavior and emotion to incidence of ischemia in cardiac patients. http://mambo.ucsc.edu/psl/joehager/research.html.

Essa, I., Darrell, T., and Pentalnd, A. (1994). Tracking facial motion. In *Workshop on Motion of Non-Rigid and Articulated Objects*, Austin, TX.

Essa, I. and Pentland, A. (1994). A vision system for observing and extracting facial action parameters. In *CVPR*, pages 76–83.

Gales, M. and Young, S. (1994). Robust continuous speech recognition using parallel model combination. Technical Report TR 172, Cambridge University.

Graham, D. and Allinson, N. (1996). Face recognition from unfamiliar views: Subspace methods and pose dependency. In *Third International Conference on Automatic Face and Gesture Recognition*, pages 348–353.

Hager, J. (1985). A comparison of unit for visually measuring facial actions. *Behavior Research Methods, Instruments, and Computers*, 17:450–468.

Hager, J. and Ekman, P. (1996). Essential behavioral science of the face and gesture that computer scientists need to know. In *International Workshop on Automatic Face and Gesture Recognition*.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. (1991). Rasta-plp speech analysis. *ICSI Technical Report TR-91-069*.

Hsee, C., Hatfield, E., and Chemtob, C. (1991). Assessment of the emotional states of others: Conscious judgments versus emotional contagion. *Journal of Social and Clinical Psychology*, 11:119–128.

Jebara, T. and Pentland, A. (1996). Parameterized structure from motion for 3d adaptive feedback tracking of faces. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Jucke, G. and Polzer, U. (1993). Fine analysis of abnormal facial expressions in chronic schizophrenic patients — a pilot study. *German Journal of Psychiatry*.

Junqua, J., Fincke, S., and Field, K. (1999). The lombard effect: A reflex to better communicate with other in noise. In *Proceedings of the Conference on Acoustics, Speech and Signal Processing,'99*, Phoenix, Arizona.

Ma, J. and Gao, W. (1997). Text-independent speaker identification based on spectral weighting functions. *AVBPA Proceedings*.

Martin, M. (1990). On the induction of mood. *Clinical Psychology Review*, 10:669–697.

Pentland, A. and Choudhury, T. (1999). Personalizing smart environments: Face recognition for human interaction. Submitted to IEEE Computers Special Issue on Biometrics.

Pentland, A., Starner, T., Etcoff, N., Masoiu, A., Oliyide, O., and Turk, M. (1993). Experiment with eigenfaces. In *IJCAI*, Chamberry, France.

Picard, R. (1997). *Affective Computing*. MIT Press.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–284.

Schiele, B. and Crowley, J. (1996). Probabilistic object recognition using mutlidimensional receptive field histograms. In *Proceedings of International Conference on Pattern Recognition*, Vienna, Austria.

Strom, J., Jebara, T., Basu, S., and Pentland, A. (1999). Real time tracking and modeling of faces: An ekf-based analysis by synthesis approach. In *International Conference on Computer Vision: Workshop on Modelling People*, Corfu, Greece.

Tomkins, S. (1963). *Affect, imagery, consciousness: The negative Affects*, volume 2. Springer-Verlag, New York.

Yacoob, Y. and Davis, L. (1994). Computing spatio-temporal representation of human faces. In *CVPR*, Seattle, WA.

Zhou, G., Hansen, J., and Kaiser, J. (1999). Methods for stress classification: Nonlinear TEO and linear speech based features. In *Proceedings of the Conference on Acoustics, Speech and Signal Processing,'99*, Phoenix, Arizona.