Bimodal Information Analysis for Emotion Recognition

Malika Meghjani

Master of Engineering

Department of Electrical and Computer Engineering

McGill University

Montreal, Quebec

October 2009

Revised: February 2010

A Thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Engineering

DEDICATION

This thesis is dedicated to my dear family: Maa, Paa, Di, Nizar, Sha and Amyn.

ACKNOWLEDGEMENTS

This thesis is an outcome of the motivation and support of various individuals who deserve a very special mention. I would like to acknowledge each one of them for their invaluable contributions to my work. First of all, I would like to convey my sincere gratitude to my supervisors Professor Gregory Dudek and Professor Frank P. Ferrie for giving me the opportunity to work under them. Their vivid and friendly interactions helped me understand and develop new ideas and concepts for my research work. I am thankful to Dr. Antonia Arnaert from McGill School of Nursing, for providing me the Tele-Health care session recordings and introducing me to the concepts related to the nursing intervention protocols.

I would like to express my gratitude to my colleagues at the Centre of Intelligent Machines particularly the members at the Artificial Perception Lab and Mobile Robotics Lab who in their own special ways helped me have a splendid experience. Special thanks to Mitchel Benovoy for providing useful resources to commence my research and to Mohak Shah for helping me understand the concepts related to Feature Reduction techniques and Support Vector Machines. I thank Meltem Demirkus for her ideas on solving the semi-supervised training problem. I would also like to acknowledge Prasun Lala and Karim Abou-Moustafa for their timely suggestions. Special thanks to Junaed Sattar for his encouragement and guidance at several occasions.

I am grateful to Rajeev Das for patiently reviewing my technical writings and providing me his invaluable suggestions. My gratitude to: Nicolas Plamondon for the

French translation of my thesis abstract. My special appreciation to Parya Momayyez Siahkal, for being a wonderful friend, support and guide all throughout.

Finally, I would like to thank my family members for their unconditional love, support and trust. It is their strong believe in my aspirations that kept me going all throughout. Special thanks to Di, for listening to my daily technical talks and trying hard not only to understand but also provide ideas to solve some of my problems.

ABSTRACT

We present an audio-visual information analysis system for automatic emotion recognition. We propose an approach for the analysis of video sequences which combines facial expressions observed visually with acoustic features to automatically recognize five universal emotion classes: Anger, Disgust, Happiness, Sadness and Surprise. The visual component of our system evaluates the facial expressions using a bank of 20 Gabor filters that spatially sample the images. The audio analysis is based on global statistics of voice pitch and intensity along with the temporal features like speech rate and Mel Frequency Cepstrum Coefficients. We combine the two modalities at feature and score level to compare the respective joint emotion recognition rates. The emotions are instantaneously classified using a Support Vector Machine and the temporal inference is drawn based on scores obtained as the output of the classifier. This approach is validated on a posed audio-visual database and a natural interactive database to test the robustness of our algorithm. The experiments performed on these databases provide encouraging results with the best combined recognition rate being 82%.

RĖSUMĖ

Nous présentons un système d'analyse des informations audiovisuelles pour la reconnaissance automatique d'émotion. Nous proposons une méthode pour l'analyse de séquences vidéo qui combine des observations visuelles et sonores permettant de reconnaître automatiquement cinq classes d'émotion universelle : la colère, le dégoût, le bonheur, la tristesse et la surprise. Le composant visuel de notre système évalue les expressions du visage à l'aide d'une banque de 20 filtres Gabor qui échantillonne les images dans l'espace. L'analyse audio est basée sur des données statistiques du ton et de l'intensité de la voix ainsi que sur des caractéristiques temporelles comme le rythme du discours et les coefficients de fréquence Mel Cepstrum. Nous combinons les deux modalités, fonctionnalité et pointage, pour comparer les taux de reconnaissance respectifs. Les émotions sont classifiées instantanément à l'aide d'une « Support Vector Machine » et l'inférence temporelle est déterminée en se basant sur le pointage obtenu à la sortie du classificateur. Cette approche est validée en utilisant une base de données audiovisuelles et une base de données interactives naturelles afin de vérifier la robustesse de notre algorithme. Les expériences effectuées sur ces bases de données fournissent des résultats encourageants avec un taux de reconnaissance combinée pouvant atteindre 82%.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABELS

Chapter 1

Introduction

We present a bimodal system for the study of voice patterns and facial expressions of human subjects to identify their expressed emotions. Our system is a combination of audio analysis and computer vision algorithms which works in an analogous manner to the human perceptual system. The combination of audio and visual information from the two channels provides complementary data which together enhance the emotion recognition rates.

Emotions are a universal means of communication which can be expressed non-verbally without any language constraints. They are recognized through facial expressions, voice tones, speech and physiological signals. The speech signals provide the context for the expressed emotions but cannot be used as a universal indicator for recognizing them as they are dependent on the language content. The physiological signals such as blood pressure, body temperature and force feedback are relatively more accurate and universal indicators of emotions but they require user-aware and intrusive methods for collecting the data. Hence, in this thesis, we primarily focus on two universal and non-intrusive modes of emotion recognition, namely, voice tones and facial expressions.

The motivation for designing and implementing an emotion recognition system is to automatically create video summaries of the audio-visual data in order to study the emotion trends of clinical subject in a Tele-Health care application. The Tele-Health care scenarios we are dealing with are the recordings of videoconferencing calls between a nurse and a patient. The goal of these Tele-Health care sessions is to help the patients suffering from Chronic Obstructive Pulmonary Disease (COPD) to better self-manage their health in order to prevent and survive any chronic attacks. During these sessions, the nurse is required to monitor the health and emotion states of the patient. The health status is monitored based on the current and previous physiological measurements recorded during each session. The emotion condition is examined based on the present state of the patient and any subjective notes made by the nurse from the previous sessions. This limited assessment of the emotion state of the patient necessitates the need for validating the role of various nursing procedures which help in providing the emotional care to the patient. The role of these nursing procedures is evaluated by mapping the interventions of the nurse to the corresponding emotion states of the patient as represented in Figure 1-1.

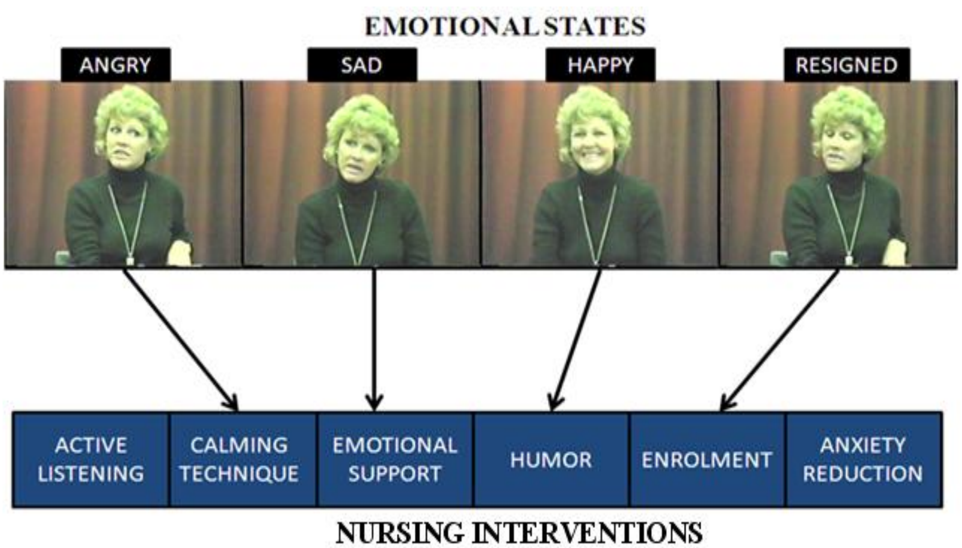


Figure 1-1: Emotion recognition in Tele-Health care application

These analyses are presently performed manually by a human annotator who labels the nursing interventions and subjectively evaluates their overall contribution by collecting the patient’s feedback. In such a case, an automatic annotating system would provide an efficient quantitative measure for developing and improving nursing intervention protocols which can be utilized for educational and teaching purposes.

1.1 Problem Statement

Our goal is to design an automatic emotion recognition system as a proof of concept for the Tele-Health care application, described in the previous section. For this purpose, we propose an audio-visual information analysis system for recognizing five universal emotions: ‘Anger’, ‘Disgust’, ‘Happiness’, ‘Sadness’ and ‘Surprise’. A major challenge for the design of this system is to deal with varying temporal nature and intensity of the emotions expressed by different individuals. This characteristic of the emotions makes it difficult to sequentially train the system.

Our audio-based emotion recognition system deals with this issue by measuring the global statistics of the acoustic features instead of the local dynamic features, to eliminate the effects of noisy measurements obtained during the temporal analysis of the auditory signal. Our visual-based emotion recognition system resolves the problem of temporal variations by selecting a single key frame instead of using all the frames in the sequence for training purposes. The selected frame represents the peak intensity of the facial expression. This single frame, image-based approach avoids the computations required for the temporal analysis of the visual data and yet, provides the necessary information for training the system.

A crucial aspect of the image-based training process, described above, is the selection of the key representative frame from an emotion sequence. We initially obtain these frames by manually selecting one frame from each input emotion sequence. We later implement a partially automated process for frame selection and compare it with the results obtained from the manually selected frames.

The temporal relation between the visual frames in a sequence is obtained during classification. For this purpose, we use the image-based training process and instantaneously classify each frame in the test sequence. The classification probabilities of all the frames are sequentially aggregated to compensate for the loss of the temporal relation at the feature level.

Another important factor, which is common to most recognition systems, is the availability of an ideal database for training the system. The characteristics of ideal training data for emotion recognition are that the subjects in the database are expected to explicitly express the required emotions while facing the camera under proper lighting and sound conditions. These conditions are, however, not practical for many real-time applications. Hence, we consider a combination of an ideal posed database [28] and an unconstrained natural database [20] for training and validating our approach. These two databases are standard and are available to the research community for comparison of different emotion recognition techniques.

1.2 Approach

Our bimodal emotion recognition system is made up of three major components: audio analysis, visual analysis and data fusion. An outline of our approach is presented in Figure 1-2.

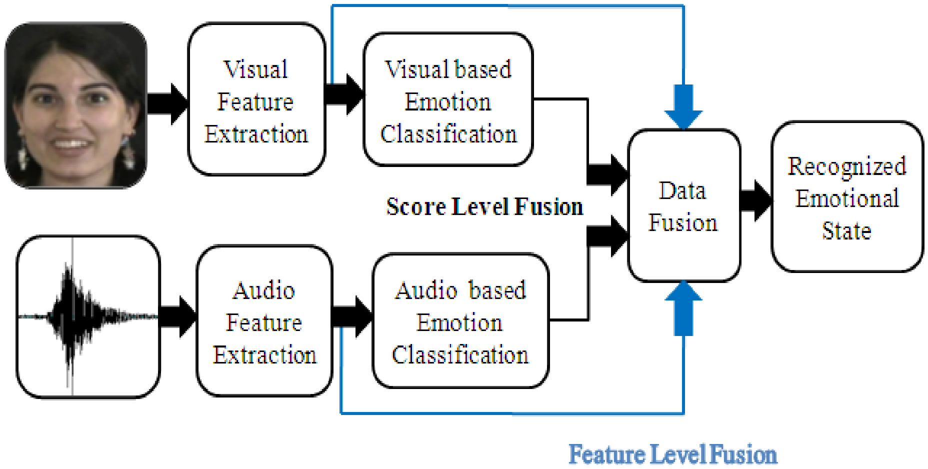


Figure 1-2: Bimodal emotion recognition system.

Our system initially separates the input data into audio and visual information for their individual analysis. The audio analysis includes the extraction of the global acoustic features related to pitch and intensity of the auditory signal, along with dynamic features such as speech rate and Mel Frequency Cepstral Coefficients (refer to Page 25). For the visual analysis, we extract appearance-based features using a bank of Gabor filters at selected frequencies and orientations.

The two modalities are combined using either a feature-level or score-level fusion technique. The feature-level fusion is obtained by concatenating audio and visual features to obtain a single feature vector for joint emotion classification. The score-level fusion is performed by initially classifying the individual modalities and obtaining their respective classification scores. These scores are later fused to obtain a combined audio-visual emotion recognition rate. The classification step in the two techniques is performed using a multi-class Support Vector Machine.

1.3 Applications

The scope of automatic emotion recognition systems is investigated in a wide range of domains such as tele-health care monitoring, tele-teaching assistance, gaming, automobile driver alertness monitoring, stress detection, lie detection and user personality type detection. Some of the multimodal emotion recognition systems developed by the research community are: an animated interface agent that mirrors a user's emotion [47], an automated learning companion [40] to detect a user's frustration for predicting when they need help, and a computer aided learning system [21] for developing user centric tutoring strategies. The target application of our system, as mentioned earlier, is the analysis of the Tele-Health care recordings for automatic annotation of the emotion states of the patient such that it can be used for evaluating the role of various nursing interventions.

1.4 Research Goals

The aim of our research is to achieve the following results:

I. To combine the best features of instantaneous and temporal visual-based emotion recognition systems in order to overcome computational complexity and maintain reasonable recognition rates.

II. To partially automate the training process of the visual system and avoid manual selection of samples for training the system.

III. To evaluate the performance of temporal aggregation methods used for the visual system.

IV. To assess the performance of global statistics of acoustic features for audio-based emotion recognition.

V. To validate the assertion made by past research in the field of bimodal emotion recognition that data fusion improves recognition rates significantly.

1.5 Outline

This thesis is structured in five chapters. It begins with a review of the relevant research work in Chapter 2. This chapter presents separate discussions on audio-based emotion recognition, visual-based emotion recognition and bimodal emotion recognition systems. An overview of application-specific bimodal emotion recognition systems is also included at the end of this chapter. Chapter 3 provides the details of the feature extraction methods for the audio and visual analysis systems respectively. In Chapter 4, we describe the feature selection and classification techniques. We compare different feature selection methods and choose the method which provides the best cross-validation accuracy. The implementation of the entire system along with the experimental results is illustrated in Chapter 5. This chapter provides the details of training, testing and fusion methods. The thesis is concluded in Chapter 6 with a detailed analysis of the experimental results and comparison of our results with our goals. We discuss the glitches in our present system along with the possible future improvements.

Chapter 2

Related Work

This chapter covers a range of work from the field of emotion recognition, specifically audio-visual based bimodal emotion recognition. The basic components of an audio-visual bimodal emotion recognition system include the audio information analyzer, facial expression recognizer and a data fusion scheme for combining the two modalities. This structure of bimodal emotion recognition system is widely used in the literature [41], but the methods adopted for the analysis at each step vary depending upon the required application. We discuss the research work from each of these primary domains and highlight the novel contributions in the respective fields. We mainly focus on algorithms which are computationally inexpensive and can be implemented in practical applications.

2.1 Audio-based Emotion Recognition

The research for audio-based emotion recognition mostly focuses on two measurements: linguistic and paralinguistic [1]. Linguistic measurement for emotion recognition conforms to the rules of the language whereas paralinguistic measurement is meta-data which are related to the way in which the words are spoken, i.e. in terms of variations in pitch and intensity of the audio signal that are independent of the identity of the words in the speech. The decision regarding the relative utility of these two categories of features for emotion recognition is inconclusive in the literature [41]. Hence, in order to obtain an optimal feature set, researchers [2] have combined acoustic features with language information using a Neural Network architecture. In this section, we only focus on paralinguistic based emotion recognition methods since they can be generalized to any language database.

There are four aspects of paralinguistic features: tone shape (e.g., rising and falling), pitch variations, continuous acoustic measurements (e.g., duration, intensity and spectral properties) and voice quality as discussed in [1]. A comprehensive relation between the statistical properties of the paralinguistic features and the respective emotion classes they represent is obtained from [3] and presented in Table 2-1 as a reference.

Table 2-1: Summary of human vocal effects described relative to neutral speech. (Derived from [3])

	Anger	**Happiness**	**Sadness**	**Fear**	**Disgust**
Speech Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much slower
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wide	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy	Blaring	Resonant	Irregular	Grumbled

In order to identify the nature of the paralinguistic features to be used for multi-class emotion recognition, Shuller et al. [4] investigated the performance of two feature sets: global statistical features and instantaneous continuous features. The global statistical features are computed for the entire utterance by deriving statistical measures from the pitch and intensity variations whereas the instantaneous features are the local continuous measurements of these variations which represent the dynamics in the speech signal. These two sets of features namely, global and instantaneous features, are classified using a Gaussian Mixture Model (GMM) and a continuous Hidden Markov

Model (cHMM) respectively. The findings of these experiments suggested that the global statistical features improved the recognition rates by at least 10% when compared to the instantaneous features. In another work [2], they evaluated the performance of the global statistical features using seven different classifiers and reported that Support Vector Machines (SVM) performed best at classification and the K-Means algorithm performed worst. They also observed that the derived pitch and energy features were major contributors for accurate recognition rates with an individual contribution of 69.8% and 36.58% respectively.

Similar results were reported by other researchers which confirm the importance of pitch, energy and intensity features for audio-based emotion recognition. Yongjin et al. [5] explored a list of audio features for emotion recognition including, pitch, intensity, Mel-frequency Cepstral Co-efficient and formant frequencies and found that the pitch and intensity features contributed to 65.71% of recognition rate against an overall recognition rate of 66.43%.

A significant, but rarely addressed, aspect of multi-speaker emotion recognition is the normalization process for speaker independent recognition. One of the relevant works in this area has been done by Sethu et al. in [6], where they proposed a normalization method by using a speaker specific feature wrapping technique. Their method involves extraction of a feature vector comprising pitch, signal energy along with its slope measurements, and signal zero crossing points over short sequences of consecutive voiced frames, which are wrapped based on the speaker’s neutral speech signal and classified using an HMM. An important result of their research was that the feature normalization process improved the recognition rate relatively by 20% when compared to the recognition rates obtained without normalization. The overall recognition rate for five classes was however quite low (45% after normalization using the signal wrapping technique).

2.2 Facial Expression Recognition

Facial expression recognition systems can broadly be classified into two categories based on the methods used for its analysis, these are: target-oriented and gesture-oriented emotion recognition systems [1]. The target-oriented systems select a single frame consisting of the peak facial expression from a given sequence. The gesture-oriented system tracks specific facial feature points over all the frames in the sequence. The facial features which are used for identifying an expression are generally referred to as Action Units (AU) in the literature [7]. The activation of a specific combination of the AU indicates the presence of a facial expression, e.g. the AU like eye brow raised and jaw dropped together indicates surprised expression.

The two above mentioned systems have their corresponding limitations. The gesture-oriented systems require accurate tracking of all the facial action units and the interpretation of their occurrence with respect to the individual emotion class. On the other hand, the target-oriented systems require the selection of key frames which can summarize the expressed emotion. The following section presents some significant contributions based on these two approaches.

The fundamental step for any recognition system is the feature extraction process. The extracted features are required to best represent the underlying physical phenomena of interest [29]. The feature extraction methods used for target-oriented emotion recognition systems include analysis of image filter responses, subspace analysis [30], shape [31] or appearance model fitting analysis [32], [33]. On the other hand, the methods widely used for gesture-oriented systems are based on optical flow analysis which is measured using either the image gradient or inter-frame correlation information. A detailed discussion on these methods can be reviewed from the survey in [1].

An important contribution among the few real time facial expression recognition systems based on the target-oriented approach, was reported by Bartlett et al. [8]. They applied a bank of Gabor filters at 5 spatial frequencies and 8 orientations to extract the features for facial expression recognition. A novel combination was implemented using the Adaboost algorithm for feature selection and Support Vector Machines (SVM) for classification. The person-independent recognition rate based on the above combination (AdaSVM) was reported to be 90% for a posed database [25]. The emotions in this database were deliberately expressed with smooth temporal transitions beginning with neutral and ending with peak intensity facial expressions. In their implementation, the first and last frames of each sequence were selected for training their system to recognize six universal emotion classes along with neutral samples. One of the important outcomes of their work highlighted the fact that the Gabor filter responses were not sensitive to the facial feature alignment prior to the feature extraction stage since they preserve the spatial relation between features regardless of their in-plane alignment.

The training samples used for facial expression recognition system are either obtained from an image database of subjects expressing the emotions with peak intensity [42], [43] or they are selected from the visual sequence with the frame representing the peak emotion as mentioned above. The second method of selecting training samples is easier if the emotions are always present at the same frame such as in [25], where the emotions are always present in last frame of the sequence. This task however, becomes complicated when the presence of emotions is not guaranteed in the estimated frame. In order to address the problem of finding the useful information frames from a sequence of frames, approaches such as supervised and unsupervised learning techniques are implemented.

One of the supervised learning techniques was proposed by Huang et al. [9] which involved training the system using manually segmented sequences with continuous labels. Their system comprised a two-level HMM to automatically segment and recognize the facial expressions. The first level was made up of six HMMs, one for each of the six universal emotion classes. The state sequence of these individual HMMs is decoded using the Viterbi algorithm, and is used as the observation vector for the next level HMM which consists of seven states, representing the six emotion classes and a neutral state. The decoded state of the higher level HMM gives the recognized emotion. The advantage of this implementation is that it can automatically segment and classify a continuous test sequence containing different emotions one after the other.

Although the modeling of the temporal dynamics of the facial features in the above case is efficiently performed using HMM, it is poor at discriminating features for classification on a frame to frame basis. In order to overcome this drawback of HMM for classification, while taking advantage of its temporal modeling characteristic, Pantic et al. [10] implemented a technique which combined SVM and HMM. The output of the SVM classifier is obtained in form of the distance between the test pattern and the separating plane which is then used to obtain the posterior probability distribution over different emotion classes based on Platt's method [11]. The posterior probabilities are converted into likelihood measures which are used as observation vectors for the HMM. The primary goal of their research [10] was to segment a video sequence into AUs in terms of their temporal phases, e.g. neutral, onset, offset and peak.

The two methods discussed above, use a posed label database for training the system for automatic segmentation of the test sequence into respective emotion classes. In a practical scenario, the training sequences are not posed and it is tedious to obtain expert labels for each frame. Such issues were addressed by Torre et al. [33] who proposed an unsupervised learning technique for application on a spontaneous database. Their method uses shape and appearance features to cluster similar facial action units in any visual sequence. These clusters are grouped into coherent temporal facial gestures to identify the displayed AUs over any given period of time. Their system could only group the facial action units but it did not interpret these AUs in terms of the respective emotional classes.

2.3 Bimodal Emotion Recognition

The literature in the field of bimodal emotion recognition [12] suggests that recognition rates can be improved by at least 10% by combining the information from audio and visual modalities as compared to the best recognition results from an individual mode. This suggests that the bimodal information is complementary in nature which can be exploited for improvement of the system's performance. In order to combine the audio and visual data, there are three widely used fusion techniques: data level fusion, feature level fusion and decision level fusion. A comprehensive survey of the fusion techniques is given by Corradini et al. [13]. The focus of this section is on different methods of data fusion in the context of bimodal emotion recognition, which takes advantage of the complementary nature of the audio-visual information.

The pioneering work in integrating audio-visual information for automatic emotion recognition was proposed by De Silva et al. [14]. They studied the human subjects' ability to identify six universal emotion classes in order to derive a weighting function for audio and visual modalities respectively. The important conclusions of their findings were that emotions like 'Anger', 'Happiness' and 'Surprise' were better recognized in the visual domain whereas 'Sadness' and 'Fear' were more easily detected in the audio domain.

This idea was later implemented by Chen et al. [15] who demonstrated the complementary nature of the audio-visual features which can be used to resolve ambiguities between the confusing emotion classes and hence improve the recognition ability of the system. A simple rule based approach was applied by considering one feature at a time to obtain coarse to fine discrimination between different emotion classes. The visual features used for analysis had to, however, be manually recorded based on user dependent rules.

A feature level fusion technique for the audio-visual analysis of the emotion data was described in [5]. It proposes a combination of a key representative visual frame with a list of statistical audio-based features for the bimodal emotion recognition. The criterion for selecting the key frames from the audio-visual sequences was based on the heuristic that peak emotions are displayed at the maximum audio intensities. This idea of frame selection was based on the general observation of human subjects which suggested that facial features are most explicit at the highest voice amplitudes.

Dragos et al. [16] addressed the problem of analyzing continuous audio-visual data on the same time scale by re-sampling individual modalities to obtain a uniform sampling rate for the combined analysis. They also proposed a data fusion technique where they relied only on the visual data in the silent phase of the video sequence and the fused audio and visual data during non-silent segments. The visual modality during non-silent segments focused only on the upper half of the facial region to eliminate the effects caused by changes in the shape of the mouth due to speech.

In a similar work, Song et al. [17] proposed an approach for multimodal emotion recognition which focused on temporal analysis of three sets of features namely, 'audio features', 'visual only features' (upper half of facial region) and 'visual speech features' (lower half of facial region), using a triple HMM for each of the information modes. This model was proposed in order to deal with state de-synchronization of the audio-visual features and also to maintain the original correlation of these features over time. A comparative list of the bimodal emotion recognition systems is presented in Table 2-2.

It can be observed from Table 2-2, that most of the successful bimodal emotion recognition systems are person dependent and have their own database, so it is difficult to draw definitive conclusions. It can however, be noticed that the recognition rates in the case of the standard database used by Dargos et al. have high recognition rates for the individual modalities (audio: 85.15% and visual: 86.3%), but they did not report the results on any combined audio-visual databases. Whereas, in case of Marco et al. the recognition rates obtained on the standard bimodal database 'eNTERFACE' [28], was quite low (feature level: 42.24% and decision level: 45.23%).

Table 2-2: Recognition rates for bimodal emotion recognition systems

Researcher	Database Size	Classification Method	Fusion Technique	Emotions	Evaluation Method	% Correct Recognition Rates
Dragos et al. [16]	Berlin Audio Database	GentleBoost	Audio	Anger, Boredom, Disgust, Fear, Happy, Sad	2-Folds Cross Validation	85.15
	Cohn Kanade Visual Database	SVM	Visual	6 Universal Emotions		86.3
	eNTERFACE	Dynamic Bayesian Network	Feature	6 Universal Emotions		Unknown
Marco et al. [18]	eNTERFACE	Neural Networks	Audio	6 Universal Emotions	Person Independent	41.63
		SVM	Visual			41.43
		Neural Networks	Feature/ Decision			42.24/45.23
Yongjin et al. [5]	Unknown, 8 Subjects	Fisher's Linear Discriminate Analysis	Audio	6 Universal Emotions	Person Independent	66.43
			Visual			49.29
			Feature (One V/s All Classifier)			82.14
Busso et al. [19]	Unknown, 1 actress	SVM	Audio	Angry, Sad, Happy, Neutral	Person Dependent	70.9
			Visual			85.1
			Feature/ Decision			89.1/89
Zeng et al. [22]	Unknown, 1 male/ 1 female	HMM	Audio	Positive, Negative	Person Dependent	65.15/75.09
		Locality Preserving Projection HMM	Visual			84.85/87.50
		Multi-Stream HMM	Decision			89.39/90.36
Song et al. [17]	Unknown	HMM	Audio	6 Universal Emotions + Neutral	Person Dependent	68.42
		HMM	Visual			82.52
		THMM	Feature			90.35

2.4 Application Specific Emotion Recognition System

The ultimate challenge for an automatic emotion recognition system is to identify the spontaneous emotions from natural conversations and to implement a system for a real time application. Some of the major concerns for such a system are: availability of a standard database for training, categorization of spontaneous emotion classes, consistency in the methods for ground truth labeling and detection of subtle and short span emotions. In this section we discuss a range of application domains where these issues are practically resolved.

The foremost task in standardizing the spontaneous recognition systems is to decide the categories in which the emotions can be meaningfully classified, since the six universal emotion classes are too explicit to occur in natural conversations. Therefore, researchers have borrowed ideas from psychology to determine a meaningful set of labels. For example, Cowie et al. [20] used the concept of a four quadrant emotion space called 'activation-evaluation' space, Figure 2-1. They used this space to continuously map the perceived emotional states of the subjects in terms of the intensity (vertical axis: activation) and nature (horizontal axis: evaluation) of the expressed emotions.

Figure 2-1: 'Activation-Evaluation' Emotion Space [20]

Based on this model, Picard et al. [21] designed an interactive learning tool called "Learning Companion" which assessed the affective states of the subjects and adjusted the system's difficulty level, in order to help them complete any learning task optimally. In another contribution, Zeng et al. [22] used simpler labels for their system to classify natural interview conversations as either positive or negative expressions. This method of evaluating the affective states in the emotion space is relatively practical but requires consistent labeling of the audio-visual sequences by the experts.

On the other hand, it is possible to recognize the prototypical emotions of the subjects for spontaneous real time applications, provided there are multi-modal information channels which can help in affirming the results obtained by the individual modalities. For example, Lisetti et al. [23] proposed a multimodal interface for a Tele-Health care application. Their system comprised five input modalities: physiological signals, voice pitch and intensity information, visual signals for facial expression recognition, haptic signals and speech signals for natural language processing. A combination of these modalities helped them obtain a comprehensive trend of the patient's emotional state over time. The analysis of patient's emotional trend facilitated the process of early detection, diagnosis and treatment for "manage by exception" patients.

2.5 Databases

The most common concern for any recognition system is to have an ideal database for training purposes as it plays an important role in deciding the system's performance. Table 2-3 summarizes a list of databases used by the emotion recognition community which range from posed to natural databases. The two italicized databases ([28], [20]) in this table are used for evaluating our work.

Table 2-3: List of emotion databases

Database	**Modalities**	**Database Size**	**Nature**
Japanese Facial Expression Database [24]	Images	213	Posed
Cohn Kanade [25]	Visual	625	Posed
MMI [26]	Visual	197	Posed
Berlin Database [27]	Audio	800	Posed
e'NTERFACE 2005 [28]	*Audio-Visual*	*1166*	*Posed*
Belfast Naturalistic Database [20]	*Audio-Visual*	*24*	*Natural Interactions*

CHAPTER 3

Feature Extraction

This chapter discusses the theory and implementation of the feature extraction process used for our audio and visual analysis systems individually. We describe in detail the required pre-processing and parameter selection methods to be applied at various stages of the feature extraction process. We conclude the chapter with the discussion on the post-processing steps that are necessary to complete the feature extraction process in order to obtain a subset of most discriminative feature set for emotion classification.

3.1 Audio Analysis

The audio-based emotion recognition is efficiently analyzed using the prosody information of the speech signal. The prosody information is a combination of data related to pitch, intensity and duration of the speech signal. The audio analysis component of our system extracts the pitch and intensity (amplitude) contours which represent the variation and the rhythm in the voice. We do not consider any semantic information for emotion recognition in order to avoid the dependency of the system's performance on the language content. We evaluate other speech related temporal features such as speech rate, spectral analysis and MFCC which highlight the dynamic variations in speech. We derive a list of global statistical features from the pitch and intensity contours as listed in Table 3-1. In addition to these features, we obtain five statistical features (minimum, maximum, mean, median and standard deviation) derived from the MFCC which results in a total of 85 audio features for each audio-visual sequence.

The feature extraction process initially involves the pre-processing of the audio signal by removing any leading or trailing silent edges of the signal where there is no information present. The pitch and intensity contours of the audio signal are obtained using PRAAT [34], the speech analysis software. Once the global statistical features are extracted, they are normalized individually for each subject. The normalization of the features removes the speaker variability effect before performing the classification. An outline of the audio analysis process is represented in Figure 3-1.

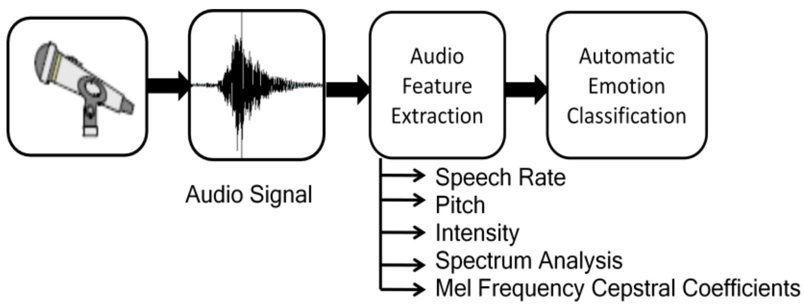


Figure 3-1: Audio analysis

3.1.1 Pitch Contour

Pitch is the fundamental frequency or the lowest frequency of the sound signal. It is one of the useful features for prosody analysis of the speech signal and is considered as an important cue for recognizing the speaker's emotions [35]. We estimate the pitch of the sound signal using the autocorrelation method as implemented in [34]. The following sub-section describes this method in detail.

The autocorrelation function $r(\tau)$ of an input time signal $x(t)$ is given by:

$$r(\tau) = \int x(t)x(t+\tau)\,dt\,, \tag{3.1}$$

where, the signal $x(t)$ is said to be periodic with period T_0 if the autocorrelation function $r(\tau)$, is maximum for every integer n, i.e. when $\tau = nT_0$. The fundamental frequency of this periodic signal is given by $F_0 = 1/T_0$.

The autocorrelation method for pitch estimation is a short-term analysis technique using which the input speech signal is analyzed in short temporal windows for local pitch estimation. The input speech signal is divided into short segments of length T which is equal to at least three minimum periods $(3T_0)$. The minimum period is determined based on the expected minimum frequency of the input signal. The segmented input signal centred around t_{mid} is subtracted by the local average value μ_x of the signal and multiplied with a window function $w(t)$. We selected Hanning window [35], for the analysis since it constitutes short length periods which can capture the characteristics of rapidly changing sound signals. The windowed function thus obtained, is given by the equation:

$$\alpha(t) = \left(x\left(t_{mid} - \frac{T}{2} + t\right) - \mu_x\right) w(t), \tag{3.2}$$

where, Hanning window $w(t)$ is,

$$w(t) = \frac{1}{2} - \frac{1}{2}\cos\left(\frac{2\pi t}{T}\right). \tag{3.3}$$

Therefore, the windowed autocorrelation function is given by,

$$r_a(\tau) = \int_0^{T-\tau} \alpha(t)\alpha(t+\tau)dt . \tag{3.4}$$

The autocorrelation function of the original input signal is obtained by dividing the autocorrelation function of the windowed signal by the autocorrelation function of the window itself. This process removes the effects of the window function,

$$r_x(\tau) \approx \frac{r_a(\tau)}{r_w(\tau)}. \tag{3.5}$$

The maximum value for the autocorrelation function gives the local period estimation from which the local fundamental frequency is obtained. The local fundamental frequencies are interpolated to obtain the required pitch contour of the entire input speech signal.

3.1.2 Intensity (Amplitude) Contour

The measurement of the audio intensity or amplitude is straightforward and is obtained by initial processing of the raw input signal. The raw speech signal is subtracted by the mean air pressure from the input signal to nullify the effect of the constant air pressure (DC offset) that might be added during different stages of the recording process.

The amplitude signal thus obtained is smoothened in order to remove the intensity variations caused by the local changes in the fundamental period. The smoothing is performed by measuring the amplitude at each point as a weighted average of the amplitude over the neighbouring points in time. A Gaussian function is selected for weighting the input signal to obtain a smooth amplitude contour. The extracted amplitude and pitch contours are shown in Figure 3-2.

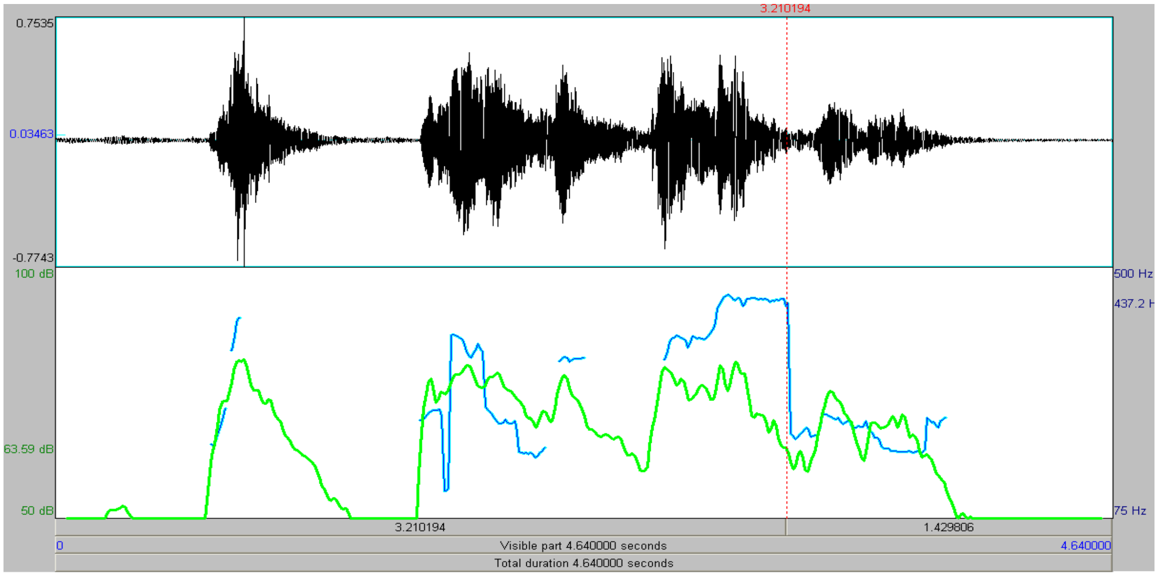


Figure 3-2: The input speech signal and corresponding pitch (blue) and amplitude (green) contours.

3.1.3 Mel-Frequency Cepstral Coefficients

MFCCs are widely used for speech and speaker recognition applications as they represent the phonetic features of the speech signal. We extract these features in order to evaluate the utility of the speech related features for emotion recognition. MFCC is a representation of the real cepstrum (spectrum of a spectrum) of the speech signal on a Mel-frequency scale. The Mel-frequency scale is a non-linear scale which approximately models the sensitivity of the human ears and gives a better discrimination of different speech segments.

The steps involved in the extraction of MFCC from the input speech signal can be summarized as:

I. The Discrete Fourier Transform of the input signal $x[n]$ of length N is obtained in order to perform the spectral analysis of the signal,

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-\frac{j2\pi k}{N}} \quad 0 \leq k < N \,. \tag{3.6}$$

II. A bank of M triangular overlapping filters H_m with increasing bandwidth is defined to measure the average energy of the above spectrum around a centre frequency k on the Mel-scale.

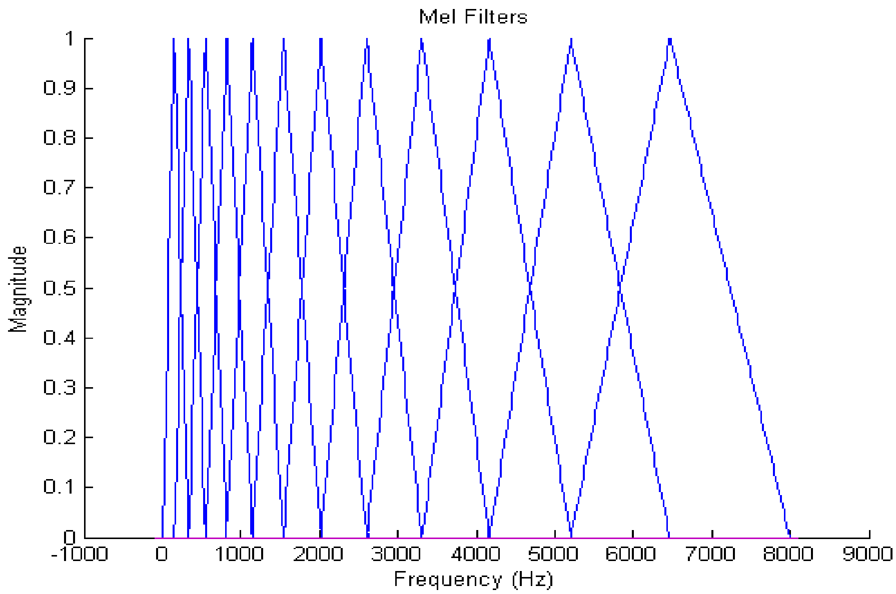


Figure 3-3: Bank of triangular filters used for Mel-Cepstrum analysis, given by:

$$H_m[k] = \left\{ \begin{array}{cc} 0 & k < f[m-1], \\ \frac{k - f[m-1]}{f[m] - f[m-1]} & f[m-1] \leq k \leq f[m], \\ \frac{f[m+1] - k}{f[m+1] - f[m]} & f[m] \leq k \leq f[m+1], \\ 0 & k > f[m+1]. \end{array} \right\} \tag{3.7}$$

III. The logarithm of the filtered response (Mel-log energy) $S[m]$ is obtained and the discrete cosine transformation is applied to the signal as given by the following equations:

$$S[m] = \ln[\sum_{k=0}^{N-1} |X[k]^2| H_m[k]] \quad 0 < m \leq M \,, \tag{3.8}$$

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos\left(\frac{\pi n\left(m-\frac{1}{2}\right)}{M}\right) \quad 0 \leq n < M. \tag{3.9}$$

IV. The discrete samples of the amplitude of the resulting signal are the MFCCs.

We consider the first 13 MFC coefficients, as it has widely been accepted and experimented in the literature [35] that these coefficients have the most salient features for speech and emotion recognition [5] tasks. Once the coefficients are obtained, we measure the five statistical values from these coefficients in order to deal with the varying length of the resulting cepstrum obtained from the varying length of the input speech signals. Thus, the statistical features obtained from the MFCC provide a uniform dimensional feature vector for any length of the input signal.

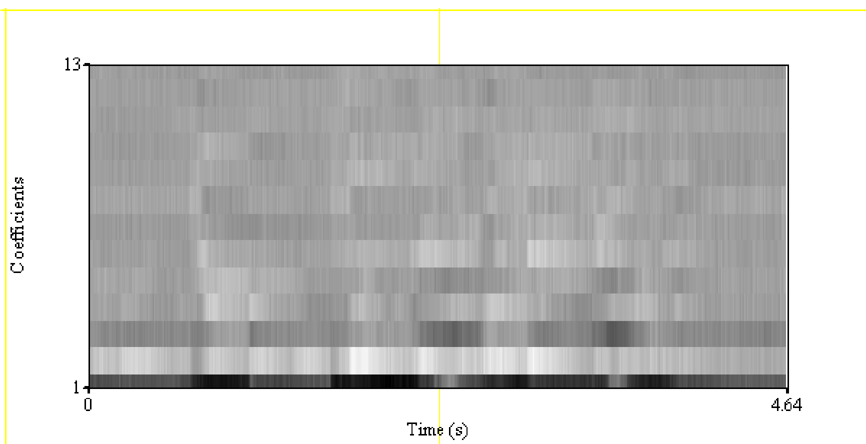


Figure 3-4: 13 MFCC of the speech signal represented in Figure 3-2.

3.1.4 Global Statistical Features

We determine a list of global statistical features from pitch and intensity contours mentioned in Table 3-1. A majority of these features are the basic statistics (mean, median and standard deviation) obtained directly from the contours whereas the other derived features are discussed as follows:

(F-1/F-7) Pitch relative maximum (minimum) position = $\frac{pitch\ maximum(minimum)\ time}{total\ duration\ of\ speech}$

(F-3) Pitch mean absolute slope = $\frac{pitch\ difference}{rise(fall)\ pitch\ duration}$

(F-16) Speech rate = $\frac{number\ of\ voice\ sements}{total\ length\ of\ voice\ segments}$

Table 3-1: List of acoustic features

Feature number	Feature description
F-1	Pitch relative maximum position
F-2	Pitch standard deviation
F-3	Pitch mean absolute slope
F-4	Pitch mean
F-5	Pitch maximum
F-6	Pitch range
F-7	Pitch relative minimum position
F-8	Pitch minimum
F-9	Voiced mean duration
F-10	Spectral energy below 250 Hz.
F-11	Intensity standard deviation
F-12	Intensity mean fall time
F-13	Intensity mean
F-14	Intensity mean rise time
F-15	Unvoiced mean duration
F-16	Speech rate
F-17	Signal mean
F-18	Intensity maximum
F-19	Spectral energy below 650 Hz.
F-20	Intensity relative maximum position

(F-10/F-19) Spectral energy below 250 Hz (650 Hz): The spectral properties of the speech signal represents its phonetic content which is measured using a spectrogram. A spectrogram is a spectral-temporal representation of the signal with time on the horizontal axis and frequency on the vertical axis. The darker regions of the spectrogram indicate the presence of high energy densities and the lighter regions indicate vice-versa. Since the spectral properties depend to a great extent on the spoken content and not the emotions, we measure values only in a particular range (< 250 Hz for males, < 500 Hz for females).

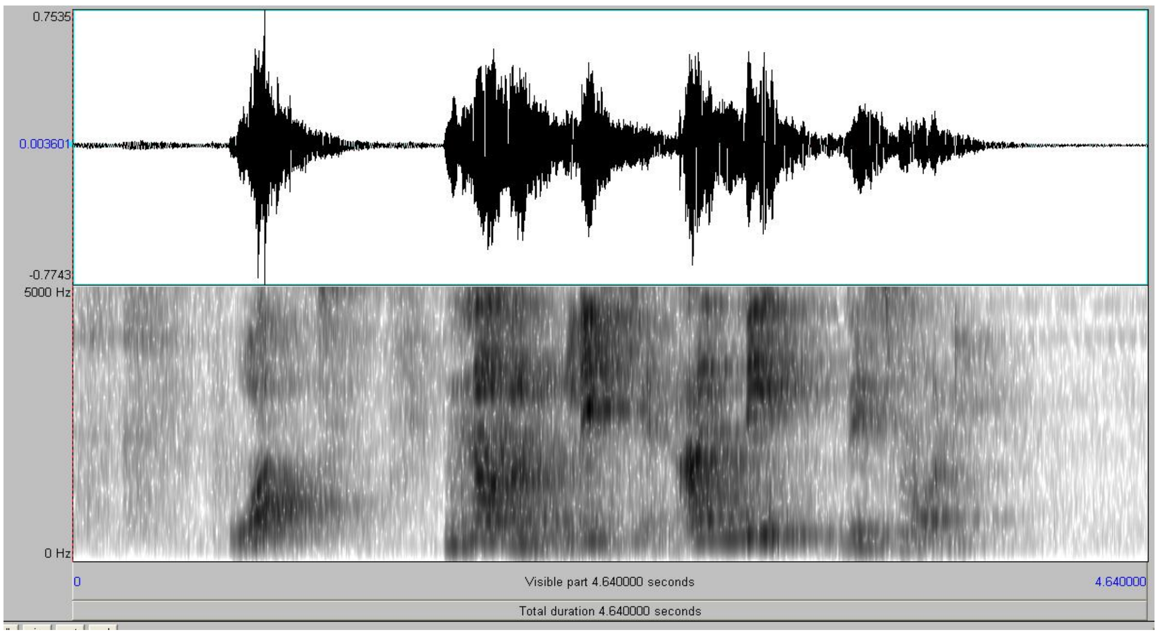


Figure 3-5: Speech signal and its spectrogram.

3.2 Visual Analysis

Our visual analysis system evaluates the facial expressions of the subjects in the scene as a cue for recognizing the expressed emotion. We detect frontal faces of the subjects in each visual frame, which is processed by a bank of Gabor filters at different spatial frequencies and orientations for feature extraction. The advantage of using Gabor filters is that it provides a multi-resolution analysis of the input image. The cost of this feature extraction process is the increased complexity and the high dimensionality of the feature vector obtained at each instance of time.

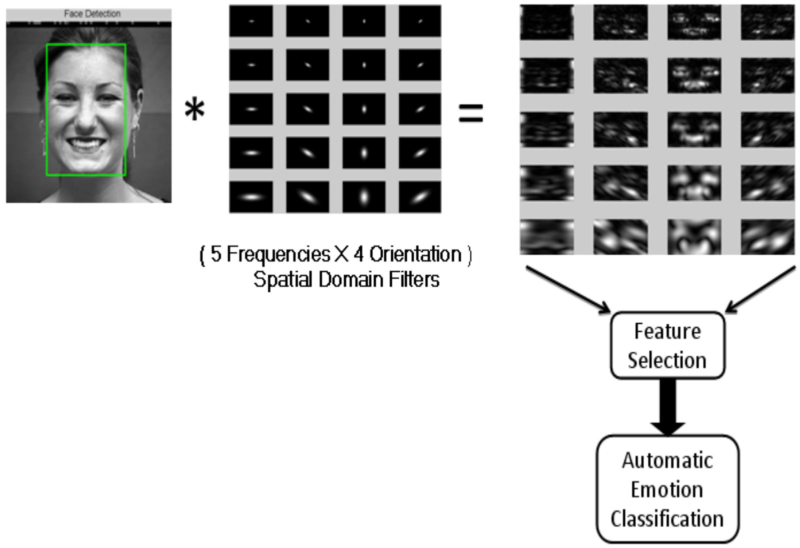


Figure 3-6: Visual analysis using bank of 20 Gabor filters (5 spatial frequencies and 4 orientations). Face image © Jeffrey Cohn.

Figure 3-6 represents the feature extraction process where Gabor filters in the spatial domain are convolved with the input image to obtain the Gabor responses. The computational complexity of the feature extraction process is reduced by performing filtering in the frequency domain instead of the spatial domain. The dimensionality of the feature vector is reduced by selecting most discriminative features using a feature selection process. The selected features are classified using a multi-class Support Vector Machine for visual based emotion recognition. We discuss each step of visual processing in the following subsections along with the relevant theory to provide a comprehensive view of our system.

3.2.1 Face Detection

The first step in recognizing the facial expressions is to detect the facial regions. We train our system to recognize frontal face facial expressions, i.e. we consider the subjects to be mostly facing the camera as in the Tele-Health care application where the patient and nurse converse in a video conferencing type of a setup.

This consideration of frontal face subjects gives us the leverage of not tracking the face over the entire visual sequence and hence, avoiding tracking drifts and undue computation complexity. Therefore, we only detect faces frame by frame. We use Viola-Jones face detector [36] with implementation derived from the code available in the OpenCV library [37]. In this section, we briefly describe the implementation of the face detector.

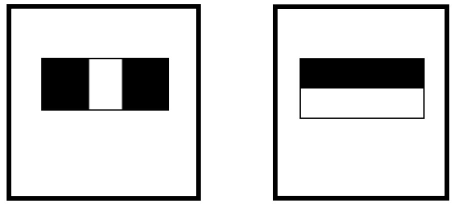
The Viola-Jones face detector [36] is widely used in many real-time applications due to its efficient and fast implementation. This detector has four major components:

I. Haar-like rectangular features for detecting facial regions.

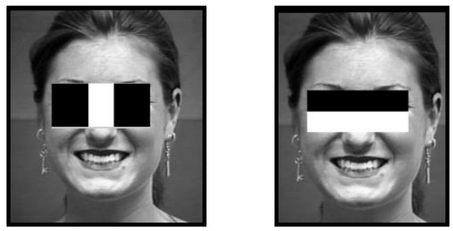
II. Integral images for fast computation and searching for facial regions over different scales and locations.

III. Adaboost algorithm for obtaining weights from several weak Haar-like features which are combined to describe the entire facial region.

IV. Cascade classifier for efficiently classifying face and non-face regions.



(a)



(b)

Figure 3-7: (a) First two ranked Haar-like feature used for Viola-Jones face detector (b) Haar-like features overlapped on a face image. Face image © Jeffrey Cohn.

The Viola-Jones face detector uses features similar to Haar wavelets. These features are simple rectangular combinations of black and white patterns as shown in the Figure 3-7(a). The rectangular patterns help in detecting different facial regions, for example, the two black regions in Figure 3-7(b) are used to detect the eye location which is comparatively darker than the nose region in the image domain. A facial region is detected by subtracting the average pixel value under the dark regions of the patterns from the average pixel value belonging to the bright regions. If the value exceeds a threshold, the expected facial region is said to be present at the evaluated image region.

In order to search for the presence of these rectangular patterns at each image location and different scales, Viola-Jones' technique uses an integral image to reduce the number of computations. The integral image is formed by integral values at each image location which are obtained by calculating the sum of all the pixels to the left and above it. The integral value at the location (x, y) in the Figure 3-8(a) is represented by sum of all the pixels in the black shaded region. The average value of the rectangular region is obtained by dividing the integral value at the bottom right corner of the rectangle by the area of the rectangle. The integral image approach helps in fast calculation of the average integral value at each pixel location for the rectangles which have a corner at the top left of the image and the diagonal corner at any location(x, y).

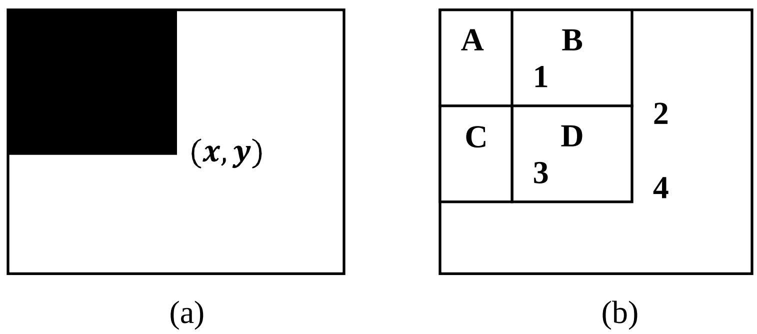


Figure 3-8: Integral image from Viola-Jones face detector. (a) Pixel location (x, y), contains the sum of pixels in the black shaded region. (b) The sum of pixels in D is $(x_4, y_4) - (x_2, y_2) - (x_3, y_3) + (x_1, y_1)$.

The sum of all the pixels (integral value) for the rectangles which do not have a corner at top left of the image is calculated by the method represented in Figure 3.9(b). The sum of all the pixels in the rectangle 'D' can be obtained by

$$D = (A+B+C+D) - (A+B) - (A+C) + A. \tag{3.10}$$

In general, with an integral image, we can obtain the sum of pixels for any rectangle by using only three operations:

$$(x_4, y_4) - (x_2, y_2) - (x_3, y_3) + (x_1, y_1). \tag{3.11}$$

The next step is to train the detector to select the Haar-like features specifically for face detection and to obtain a threshold which indicates their presence. This is achieved by using Adaboost algorithm, which combines several weak Haar-like features that can only marginally detect the presence of facial regions. The weak features are weighted by Adaboost algorithm which together form a strong feature set for the face detection. These features are combined in a chain structure as represented in Figure 3-9.

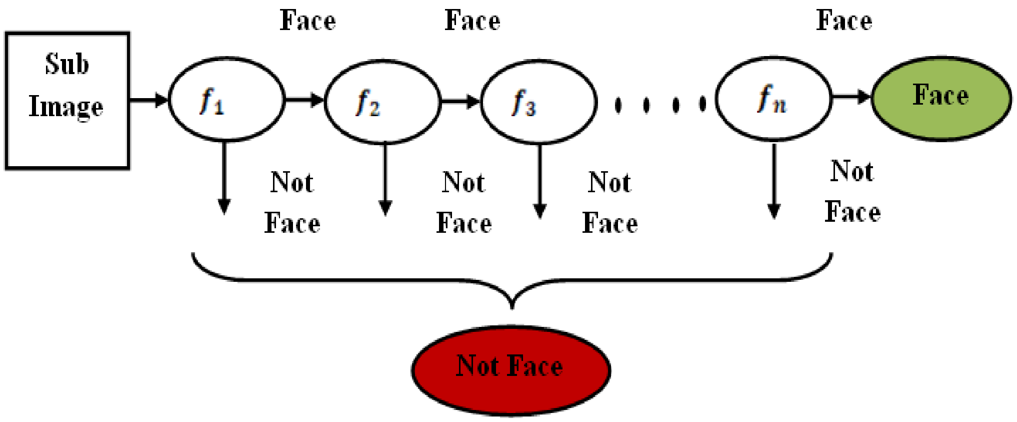


Figure 3-9: Cascade classifier: The sub-region in the image which contains the presence of all the features is classified as "Face", and others are classified as "Not Face".

For each sub-region, the presence of the weighted Haar-like features is tested beginning with the highest weighted feature. If the higher weighted features are absent in the sub-region, and then the algorithm skips the evaluation of later features in the cascade to save on computations. On the other hand, if the sub-region in the image clears the presence of all the features in the cascade, then the face is said to be detected in that region.

The facial region detected by the original detector consists of extra regions around the face which act as noise in the feature extraction process. We reduce these facial regions obtained from the original detector in order to closely bind the detector to the face as represented by Figure 3-10. The removal of these unwanted regions makes the features robust and consequently provides better recognition results.

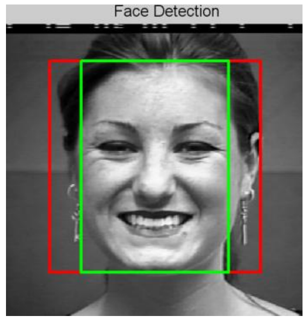


Figure 3-10: Face detection with tight bounds (green). Face image © Jeffrey Cohn.

3.2.2 Gabor Filter

Gabor filters are widely used for feature extraction in many computer vision applications like texture segmentation, edge detection, document analysis, iris recognition and face detection [38]. These filters have the property of optimally localizing the features in both spatial and frequency domains. A bank of Gabor filters, tuned at different spatial frequencies and orientations, provide a multi-resolution analysis of the input image. A combination of multi-resolution responses of the input provides better discrimination for classification.

A Gabor filter by definition, is a complex sinusoidal plane wave modulated by a Gaussian envelope at any desired frequency and orientation. A two-dimensional Gabor filter in the spatial domain at the location (x, y) is given in Equation (3.12) and represented by Figure 3-11.

$$g(x, y; f_0, \theta) = \frac{f_0^2}{\pi\gamma\eta}\ e^{-\left(\frac{f_0 x'}{\gamma^2} + \frac{f_0 y'}{\eta^2}\right)}, \tag{3.12}$$

$$\text{where, } x' = \quad x\, cos\theta + y\, sin\theta,$$

$$y' = -x\, sin\theta + y\, cos\theta.$$

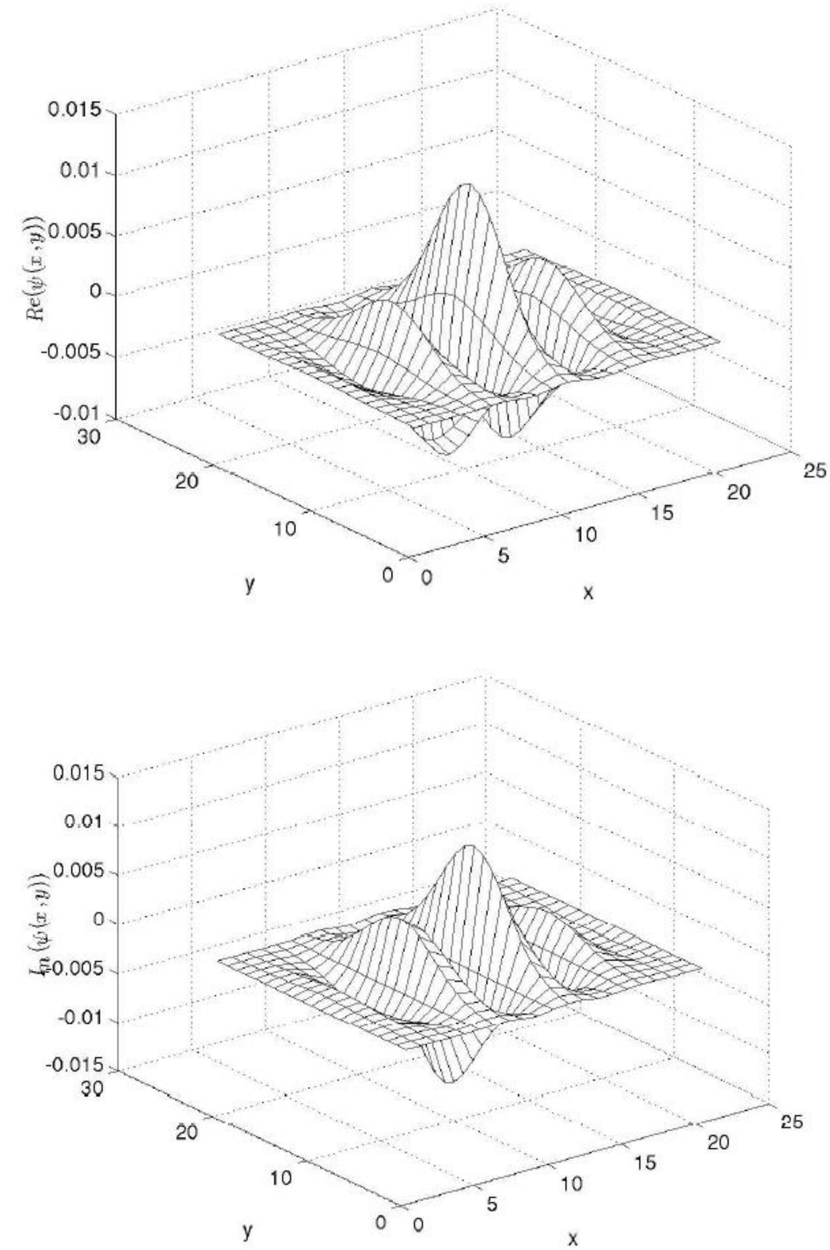


Figure 3-11: 2-d Gabor filter in spatial domain with real (top) and imaginary (bottom) components, $f_0 = 0.2, \gamma = \eta = 1$.

The center frequency of the filter is represented by f_0, the orientation of the Gaussian function's major axis and the plane wave is denoted by θ whereas, the major and minor axes of the Gaussian function are given by γ and η respectively. The Gabor filter response for an input image $i(x, y)$ is obtained by convolving the filter with the image as follows:

$$r(x, y; f, \theta) = g(x, y; f, \theta) * i(x, y). \tag{3.13}$$

The multi-resolution analysis of the image is obtained by using a bank of Gabor filters at different spatial frequencies and orientations which ensures that the extracted features are almost invariant to changes in scale and orientation. This is achieved based on the invariance property of the filter which suggests that the transformation of the object can be represented by transformation of the filters i.e.

$$r'(x, y; f, \theta) = r\left(ax, ay; \frac{f}{a}, \theta - \phi\right), \tag{3.14}$$

where, $r'(x, y; f, \theta)$ is the response of a transformed image and $r\left(ax, ay; \frac{f}{a}, \theta - \phi\right)$ is the original response which is transformed according to the input transformation. The orientation angles for the bank of filters are discretely selected from a set of uniformly spaced orientations $\theta = \frac{l2\pi}{n}$, $l = \{0, \dots, n-1\}$, where n is the number of orientations. We analyze the inputs for four equally spaced orientations in the interval $[0, \left(\frac{3\pi}{4}\right)]$ for extracting the facial features. Since, the responses of the filter for the orientations $[\pi, 2\pi)$ are the complex conjugate of the responses for $[0, \left(\frac{3\pi}{4}\right)]$. This reduces the computations by half.

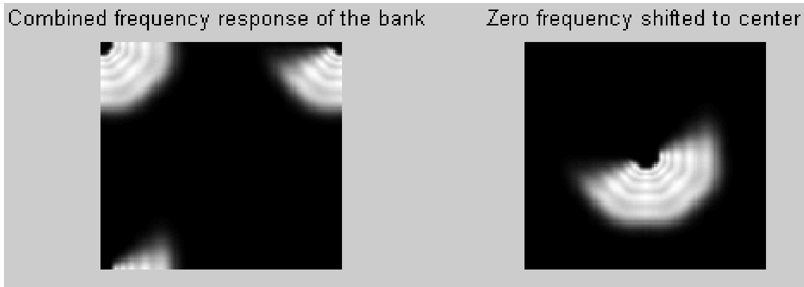


Figure 3-12: Frequency domain bank of filters at 5 spatial frequencies and 4 orientations.

The spatial frequencies of the filter are selected based on $f_l = \frac{f_{max}}{\mathrm{k}}$ where, $l = \{0, \ldots, m-1\}$ and m is the total number of orientations. The common choices for filter spacing k are half octave ($k = \sqrt{2}$) and full octave ($k = 2$). We chose the half octave filter spacing since the filter bank thus obtained captures all the important frequencies of the input facial region as mentioned in [38]. Based on these criteria for orientation and frequency spacing for the filter bank, we obtain the filter response for the input image. A list of other parameters to be selected for the implementation of the Gabor filter bank is given in Table 3-2. These parameters are selected based on the guidelines provided in [39].

Table 3-2: Parameters for Gabor filter bank (Following [39]).

Parameters	**Description**	**Selected values**
p_1	Crossing point (overlap) between adjacent frequencies	0.5
p_2	Crossing point (overlap) between adjacent orientations	0.5
k	Scaling factor for filter frequency	$\sqrt{2}$
m	Number of filters in different frequencies	5
n	Number of filters in different orientations	4
γ	Filter sharpness along major axis	1.54
η	Filter sharpness along minor axis	0.67
f_{max}	Tuning frequency of the highest frequency filter	0.25
$f_{\min}$	Tuning frequency of the lowest frequency filter	0.0625

A major drawback of the bank of Gabor filters is their cost and complexity of the computations, which makes them difficult to be applied in real time. The convolution of the input image (NxN pixels) with a Gabor filter (MxM pixels) has $O(N^2M^2)$ complexity. If we use the convolution theorem and perform the computations in the frequency domain the complexity can however be reduced to $O(N^2 \ln N)$. The filtering process in the frequency domain is performed by taking the Fourier transform of the image using the Fast Fourier Transform (FFT), and multiplying it with the frequency domain Gabor filter to obtain the frequency domain response. The responses are converted back to the spatial domain using an inverse FFT (IFFT). The entire filtering process is represented in Figure 3-13. A two-dimensional Gabor filter in the frequency domain is given by the Equation (3.15).

$$\psi(u,v;f_0,\theta) = e^{-\pi^2\left(\frac{(u'-f_0)}{\alpha^2}+\frac{v'}{\beta^2}\right)}, \tag{3.15}$$

where, $u' = \quad u\,cos\theta + v\,sin\theta$,

$v' = -u\,sin\theta + v\,cos\theta$.

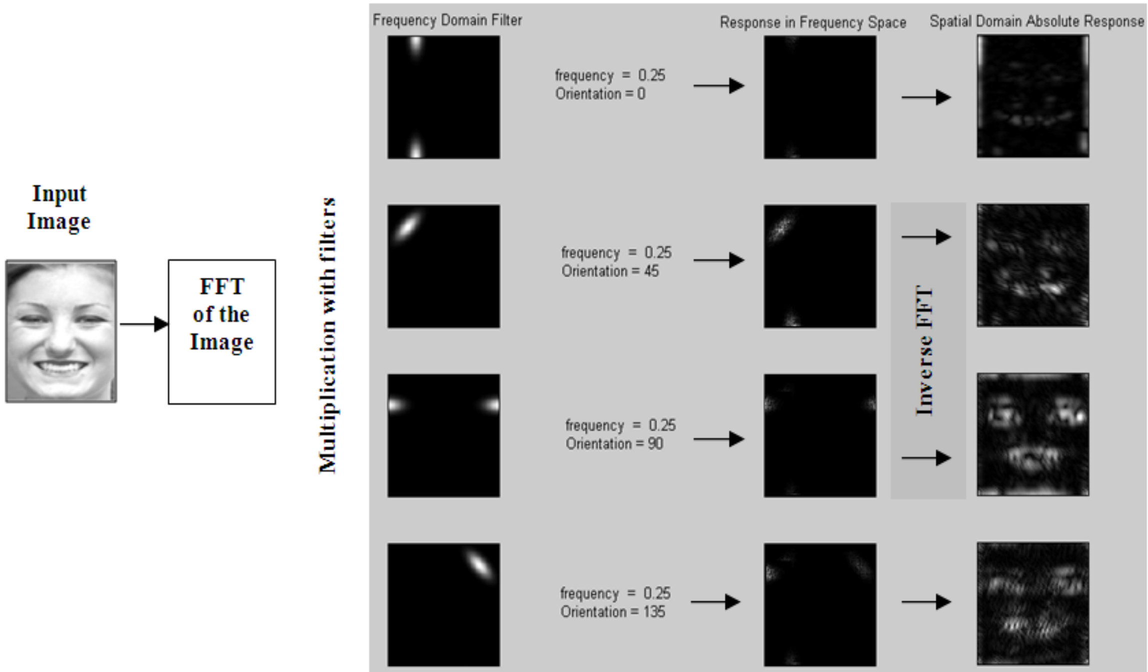


Figure 3-13: Frequency domain filtering process. Face image © Jeffrey Cohn.

3.3 Post-Processing

The feature extraction process for the audio analysis system provides a feature vector of length 85 as mentioned in the Section 3.1, whereas the length of the feature vector obtained from the visual information processing is 20480 which is the combined output from 5 spatial frequencies and 4 orientations Gabor responses of size (32x32). The large feature vectors obtained from the two modalities contain plenty of redundant features which may not be useful for the classification process. Hence, we need to obtain the best set of features such that the emotion recognition accuracy is improved and the computation time is reduced. The next chapter introduces an efficient feature selection technique which can overcome the “curse of dimensionality”.

Chapter 4

Feature Selection and Classification

This chapter is a continuation of the feature extraction processes described in Chapter 3. The high dimensional feature vectors obtained during the feature extraction process are reduced to lower dimensionality by applying a feature selection technique. We evaluate four feature selection techniques and select the one with the maximum cross-validation accuracy. The reduced set of the most discriminative features is classified using a multi-class Support Vector Machine (SVM). We discuss in detail the theory and parameter selection process for the feature selection and classification methods respectively. The feature selection method used for our application is a wrapper method which uses the SVM classifier for an optimal feature set selection. Hence, we begin our discussion with the SVM classification method followed by the feature selection process.

4.1 Support Vector Machines (SVMs)

SVM is a supervised learning technique for classification (pattern recognition) and regression (function approximation) problems. In our application we use SVMs for classification of the audio-visual data for emotion recognition. The general classification problem using any supervised learning technique requires two sets of data for training and testing the classifier (SVM). The training data set is comprised of feature vectors and their respective class labels. The goal of SVM is to create a model using the training data set to predict the labels of the test data.

SVMs are binary classifiers which construct a separating plane during the training phase, to optimally divide the training samples into the two respective classes. The optimal separating plane is defined as the plane which maximizes its distance to the nearest training examples on either side (as in Figure 4-1) such that the generalization error is reduced and the overall classification is improved. The distance between the separating plane and the nearest training samples on either side is measured using two parallel planes on the corresponding sides as represented in the Figure 4-1. The distance between the two separating planes is called the SVM *margin* and the training samples which lie on the two parallel planes are the support vectors. The separating plane is halfway between the two parallel planes. A mathematical formulation of the problem is given below.

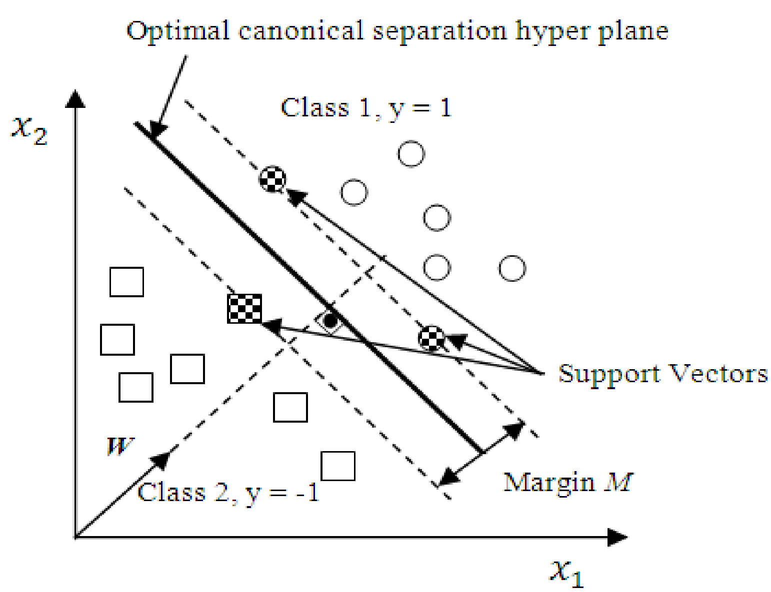


Figure 4-1: Illustration of two dimensional SVM model with separating plane represented by bold line and the two parallel planes with support vectors in the dotted lines. (Following [44]).

Consider l pairs of training examples consisting of feature vector x and the corresponding label y to be represented by $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)$, $x \in R^n$, $y \in \{-1,1\}$. The two classes ($y \in \{-1,1\}$) can be separated by many different separating planes. In fact any plane which satisfies Equation (4.1) can be used as a decision boundary (also called the separating plane) for classification.

$$d(X, W, b) = \sum_{i=1}^{n} w_i x_i + b = W^T X + b = 0, \tag{4.1}$$

where, $d(X, W, b)$ is the decision function, $W = [w_1, w_2, \dots, w_n]^T$ is the weight vector perpendicular to the decision boundary and b is the bias term indicating the distance of the decision boundary from the origin. The decision boundary selected for classification is obtained using the parameters W and b which are chosen (based on the training samples) such that they maximize the distance between the two parallel planes. The equations of the two parallel planes are:

$$W.X - b = \quad 1$$

$$W.X - b = -1$$

$$\Rightarrow y_j\left[W^T X_j + b\right] = 1. \tag{4.2}$$

The above equation is satisfied by only the support vectors ($j = 1, \dots, N_{sv}$), i.e. the training samples that lie on the two parallel planes. For the non-support vector training samples the equation is $y_j\left[W^T X_j + b\right] > 1$. The combination of the two equations for all the training samples is given by:

$$y_i[W^T X_i + b] \geq 1, \; i = 1, \dots, l \tag{4.3}$$

The distance between the two parallel planes can be estimated from geometric analysis as $\frac{2}{||W||}$. Hence, in order to maximize this distance we have to minimize the norm of the weight vector $||W||$. The minimization of the $||W||$ is equivalent to minimization of $W^T W = \sum_{i=1}^{n} w_i^2$. Therefore, the learning problem for the SVM can be summarized as a quadratic programming optimization problem:

$$\begin{aligned} &minimize \quad && \frac{1}{2} W^T W \\ &subject\ to \quad && y_i[W^T X_i + b] \geq 1, \qquad i = 1, \ldots, l\,. \end{aligned} \tag{4.4}$$

4.1.1 Linear Inseparable SVM

The learning process discussed above is for training data which are linearly separable, and there are no overlapping training samples. However, such classification problems are rare in practice and it is often observed that the training samples overlap and fall within the SVM margin. In such a scenario, the overlapped training samples which are present within the margin cannot be correctly classified and the above optimization process tends to over-fit by using all the training samples as the support vectors for classification. Therefore, to obtain a linear classifier with the maximum margin for an inseparable training data set, we have to leave some data points misclassified. In other words, a soft margin has to be created and all the data within this margin should be neglected. The width of the soft margin is controlled by a penalty parameter which balances the training error and generalization ability of the model.

The degree of misclassification thus obtained is measured by the sum of distances of the misclassified points from the corresponding parallel planes. The variables representing the misclassified point distances are known as slack variables ξ. The slack variables along with the penalty parameter C are incorporated in the optimization equations as:

$$minimize \quad \frac{1}{2}W^TW + C\sum_{i=1}^{l}\xi_i \ , \tag{4.5}$$

$$subject\ to \quad y_i[W^TX_i + b] \geq 1 - \xi_i\ ,\ i = 1, \dots l,\ \xi_i > 0.$$

The greater values of C lead to smaller numbers of misclassifications which minimize the overlapping error. The value of $C = \infty$ implies that there are no misclassified points which is not possible for the inseparable training data set. Hence, the value of the penalty term should be $C < \infty$. The value of the slack variables for the correctly classified data points is zero.

4.1.2 Non-Linear SVMs

Another modification of the original quadratic optimization problem for the SVMs was proposed to deal with non-linear training data. These training examples are overlapping but can be classified using a non-linear separating plane. The basic idea for non-linear SVM classification is to map the training data from the input space to a high dimensional space called the ‘feature space’ where the data can be classified linearly. Therefore, though the separating plane is non-linear in the original input space, it is linear in the high dimensional feature space. A non-linear SVM in the original input space is presented in Figure 4-2. The non-linear separation boundary is represented by solid line curve and the linear separation boundary is represented by dotted line.

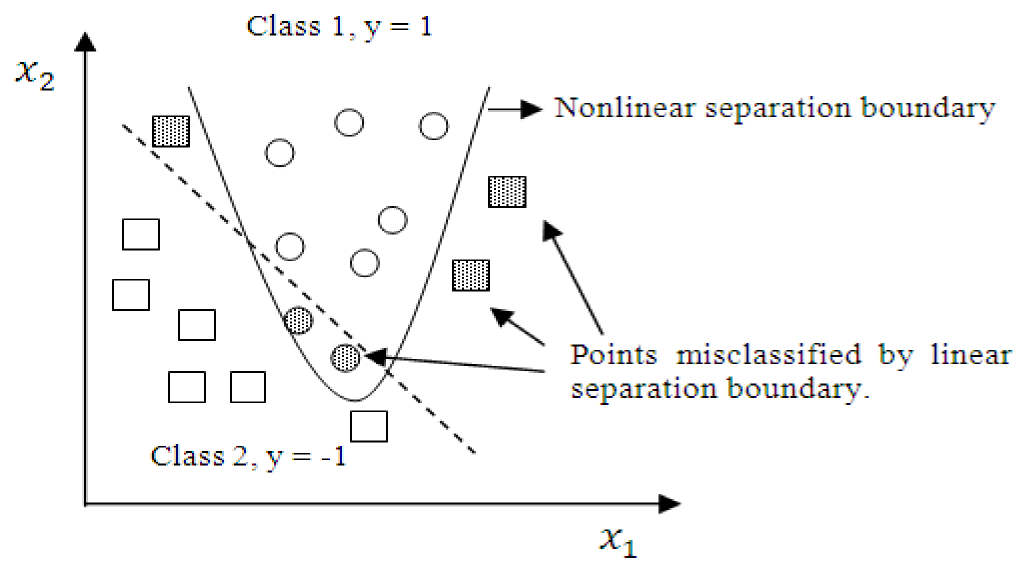


Figure 4-2: Non-linear SVM (Following [44]).

The mapping of the input data from the input space to the high-dimensional feature space is performed using a kernel function $K(X_i, X_j) = \phi(X_i)^T \phi(X_j)$. The most commonly used kernel functions are:

I. Linear:

$$K(X_i, X_j) = (X_i^T X_j) \tag{4.6}$$

II. Polynomial:

$$K(X_i, X_j) = (\gamma X_i^T X_j + r)^d, \ \gamma > 0, \tag{4.7}$$

III. Radial Basis Function (RBF):

$$K(X_i, X_j) = \exp\left(-\gamma \left\|X_i - X_j\right\|^2\right), \ \gamma > 0. \tag{4.8}$$

IV. Sigmoid:

$$K(X_i, X_j) = \tanh\left(\gamma X_i^T X_j + r\right) . \tag{4.9}$$

The variables γ, r, d in Equations (4.7) - (4.9) are the kernel parameters.

4.1.3 Multi-class SVM

SVMs were originally designed for binary classification. Hence, in order to classify more than two classes there are two methods widely used: one-versus-one and one-versus-all. The one-versus-one technique is applied by comparing every pair of classes which results in a total of $C\binom{n}{2}$ combinations where n is the total number of classes. The test examples in this case are classified by using a 'maximum-wins' voting strategy where each classifier votes for one class, the votes are counted for all the pairs and the class which wins the maximum votes is the final classification decision. The one-versus-all technique is applied by comparing each class to all the other classes. The classification for this technique is obtained by 'winner-takes-all' strategy, by which the class with the highest output function value wins.

4.1.4 Probability Estimation

The output of SVM classification is given by the class labels, but there is no measure for evaluation of the confidence level of the classification. In our application, we intend to obtain the confidence level or a probably measure for the test instances based on the training examples such that we can use it for fusing the outputs of the audio and visual modalities respectively. The output probability of classification for each class given an unseen test instance $p(y = i|X)$ is obtained by using Platt's method for probability estimation. The process involves the estimation of multi-class probabilities for classification using the one-versus-one approach discussed above. The pair-wise probabilities are calculated by:

$$r_{ij} = p(y = i|y = i \text{ or } j, X) = \frac{1}{1+exp(A\hat{d}+B)}, \quad (4.10)$$

where, the parameters A and B are estimated using the training data along with their respective decision values $\hat{d}$. The output probability $p_i = p(y = i|X)$ is obtained by solving the following optimization problem:

$$\min_p \frac{1}{2}\sum_{i=1}^{k}\sum_{j:j\neq i}\left(r_{ji}p_i - r_{ij}p_j\right)^2, \quad (4.11)$$

$$subject\ to\ \sum_{i=1}^{k} p_i = 1, p_i \geq 0, \forall\, i.$$

The solution for the above optimization problem is discussed in [11].

4.2 Feature Selection

The feature selection process is an important step for any learning algorithm since it provides a reduced set of most discriminative features which enhances the generalization ability of the learned model, and improves the classification accuracy while reducing the computation time. The algorithms applied for the feature selection process are broadly discussed in three categories: wrapper, filter and embedded methods. The wrapper method uses a classifier for evaluating the utility of the features and obtains the classification accuracy as a feedback for the feature selection method. The filter method performs feature selection by ranking the input features based on a metric and eliminates the features with the lowest ranks. The filter method is independent of any classifier and it works directly on the data. The third feature selection method is embedded in the classifier and performs simultaneous feature selection and classification tasks. In the following section, we discuss four wrapper methods for feature selection all of which use SVM for performance evaluation of the selected features. A brief description of each method is presented for feature selection of two classes which can be extended for multiple classes using one-versus-all multi-class strategy. These methods include: feature selection via concave minimization (FSV) [48], dual zero-norm minimization (L0) [46], mutual information based feature selection (MUTINF) [49], and recursive feature elimination (RFE) [50]. These methods are based on their implementation in the machine learning toolbox, 'SPIDER' [45].

I. Feature Selection via concave minimization (FSV)

This feature selection method is similar to the SVM formulation, in which the separating plane is generated by minimizing the weighted sum of distance of misclassified points i.e. weighted sum of the slack variables along with minimization of the number of dimensions of the feature space which are used to determine the plane. The formulation is given by:

$$\min_{w,\gamma,y,z}(1-\lambda)\left(\frac{e^T y}{m}+\frac{e^T z}{k}\right)+\lambda e^T|w|_*\,, \tag{4.12}$$

$$\begin{aligned} subject\ to \quad & -Aw+e\gamma+e\le y,\\ & Bw-e\gamma+e\le z,\\ & y\ge 0, z\ge 0. \end{aligned}$$

Where, e is a vector of ones, y, z are the slack variables with respect to the two parallel planes and m, k are the total number of examples per class. The parameter $|w|_*$ keeps the count of the selected features as indicated by its components which are equal to 1 if the corresponding component of the weight vector w is selected. The features corresponding to zero components of w are removed. The value of γ in Equation (4.12) is given by $\gamma = x^T w$, where x is the separating plane and A, B are the bias terms for the two parallel planes respectively. The solution of the optimization problem gives a sparse weight vector w which provides a generalized set of features. The feature space is searched for a solution by the iterative gradient descent approach, which iteratively solves an updated linear programming problem similar to Equation (4.12). A major drawback of this method is the computation time. The MATLAB implementation of the above problem runs overnight to select the generalized feature set for classification.

II. Dual Zero-norm minimization (L0)

This method of feature selection is also a modification of the SVM formulation. The algorithm is a simple and fast implementation of the modified SVM where the feature selection is performed by iterative multiplicative rescaling of the training data. The feature selection using this method is summarized below and derived from [46].

(i) Let $z = (1, \ldots 1)$ i.e. vector of ones.

(ii) Solve:

$$\min \sum_{j=1}^{n} |w_j|, \tag{4.13}$$

$$\text{subject to} \quad y_i(w(x_i . z) + b) \geq 1 .$$

(iii) Let $\bar{w}$ be the solution of the problem in step (ii). Update $z . \bar{w} \rightarrow z$.

(iv) Go to step (ii) until $|w| \leq r$, $r \rightarrow$ desired number of features.

Hence, at each step of iteration if one of the features obtains lesser weight relative to others then in the next iteration step this feature will get even lesser weight due to rescaling the data based on the previous iteration. The multiplicative weighting process reduces the scaling factor rapidly to zero which increases the computation speed and still obtain a generalized set of features for classification.

III. Mutual Information (MUTINF)

This method of feature selection is based on ranking the features using mutual information as the metric for obtaining the rank. The mutual information is measured between the test features and the expected target features.

IV. Recursive Feature Elimination (RFE)

The recursive feature elimination based feature selection method attempts to search for the best subset of a pre-defined number of features, say r, which gives the largest SVM margin. The best subset of features is obtained by using a greedy search algorithm which eliminates the input dimensions that minimally reduce the SVM margin width. The process is repeated until only r dimensions are left. This process is widely known as 'backward selection'. The algorithm minimizes the square function of weight vector $W^2(\alpha) = \sum \alpha_i \alpha_j y_i y_j k(x_i, x_j)$, which is inversely proportional to the SVM margin width. The elimination process is performed by testing each input feature and eliminating the one with the minimum value of $|W^2(\alpha) - W^2_{-p}(\alpha)|$, where W^2_{-p} indicates the value obtained after removing the p^{th} feature. The computation speed of the algorithm is increased by removing half of the features at each step of iteration.

4.3 Comparison of the Feature Selection Methods

The performance of the four feature selection methods discussed above is evaluated based on 5 folds cross-validation results. The goal of this comparison test was to obtain the best feature selection method and also obtain the optimal number of features to be selected for classification. The Figures 4-2 (a) and (b) present the results of the comparison test for both audio and visual features respectively.

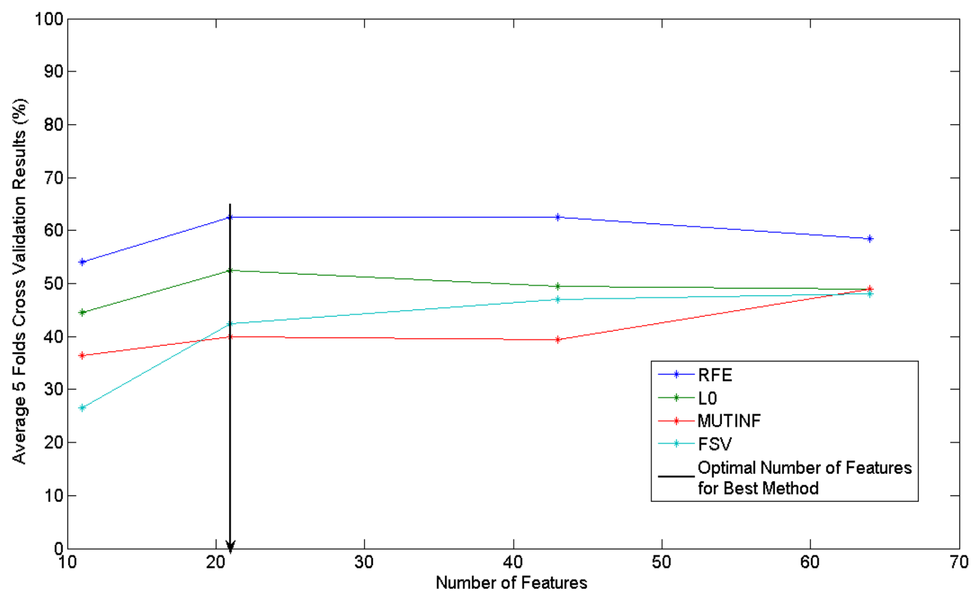


Figure 4-3 (a): Evaluation of best audio feature selection method with optimal number of features.

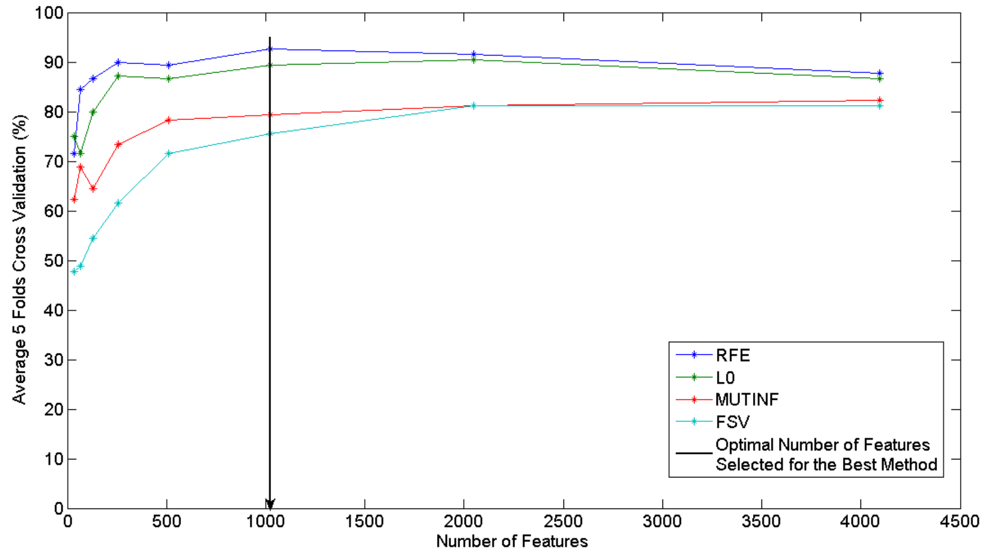


Figure 4-3(b): Evaluation of best visual feature selection method with optimal number of features.

The outcome of this experiment clearly suggests that the recursive feature elimination method performs the best for both audio and visual feature selection methods. The best cross-validation accuracy obtained for the visual features is 93% for selecting 1024 visual features from a total of 20480 features and 67% for selecting 21 audio features out of 85 features. An intuitive interpretation of these results can be obtained by examining which audio and visual features were selected by the RFE method. Table 4-1 presents a list of audio features selected by this method for emotion classification.

Table 4-1: List of acoustic features selected by RFE

Feature numbers	**Feature description**
F3	Pitch Mean Absolute Slope
F17	Signal Mean
F19	Spectral Energy below 650
F24	MFCC 1
F26, F28, F30	MFCC 2
F38	MFCC 4
F43 F44	MFCC 5
F47, F48, F49, F50	MFCC 6
F62	MFCC 9
F73, F75	MFCC 11
F76, F78, F79	MFCC 12
F81, F83, F85	MFCC 13

The visual features selected by RFE are presented based on the spatial location, frequency and orientation of the Gabor filters. Figure 4-3(a) represents the spatial location of the Gabor features with respect to individual emotion classes. These features selected for individual emotion classes were merged to obtain a common list of features for representing all emotions using RFE in conjunction with multiclass SVM. Figure 4-3(b) depicts this common set of selected features. Only the common set of features was used for emotion classification. The distribution of the selected Gabor features based on their frequency and orientation is presented in Figure 4-4.

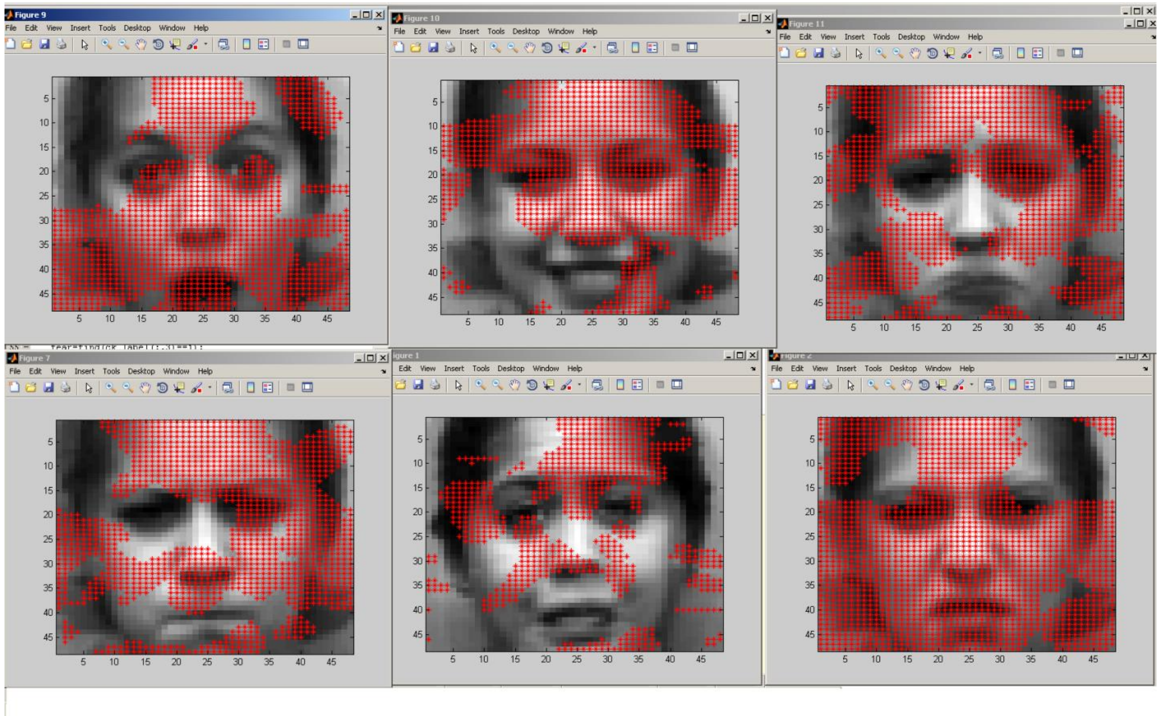


Figure 4-3(a): Spatial locations of Gabor features selected using RFE for individual emotion classes

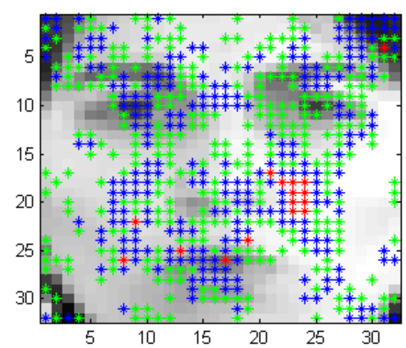
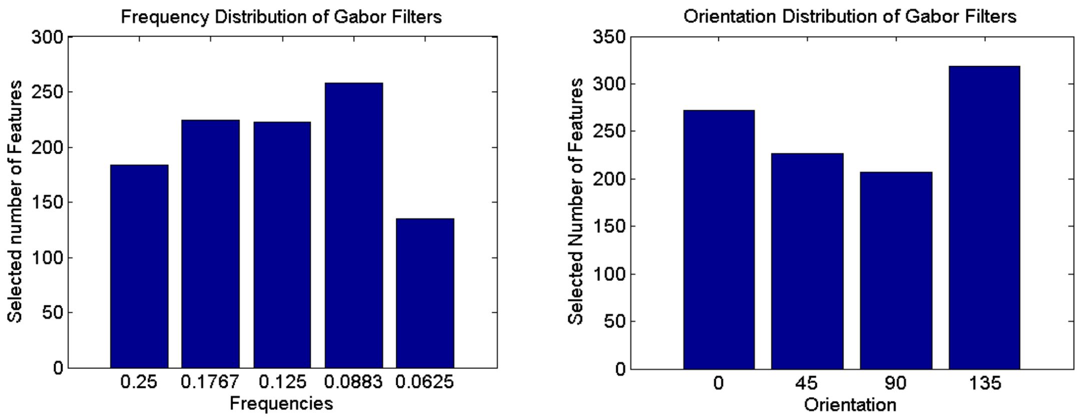


Figure 4-4(b): Spatial locations of Gabor features selected using RFE for all emotion classes

Legend: (* Green) – Selected at least once

(* Blue) – Selected more than once

(* Red) – Selected maximum number of times



(a) (b)

Figure 4-5: Distribution of the Gabor features selected in terms of frequencies and orientations

The frequency $f_4 = 0.088$ (or) $\lambda_4 \cong 11\, pixels$ in Figure 4-4(a) was selected maximum number of times. The size of the input images and the frequency domain filters used to generate the distribution plot was (128x128) pixels. Hence, it can be concluded that a minimum number of 11 pixels window length is required to optimally capture the information for our data. A mathematical relation between the image size and selection of maximum frequency for textured images can be referred from [38]. Chapter 5 will present the experimental results obtained using recursive feature elimination method based on the optimal number of audio and visual features calculated above.

Chapter 5

Experimental Results

This chapter integrates the individual processing steps of our bimodal emotion recognition system. It begins with the description of the databases used for validating our approach followed by the details of the training and testing processes. We specifically discuss three different methods implemented for training the visual component of our system. The merits and demerits of each training method are evaluated based on the experiments performed on the test data. The performance of the system is compared with respect to the results obtained by applying different fusion techniques and a temporal analysis scheme.

5.1 Database

We evaluate our approach on two types of databases. The first database [28] comprises posed audio-visual sequences in a lab environment with the subjects mostly facing the camera. The second database is a subset of the 'Belfast Naturalistic Database' [20] which consists of video recordings of natural conversations between participants and interviewer with unconstrained lighting and head movements. A sample of the two databases is presented in Figure 5-1.

The posed database consists of 9 subjects, expressing 5 emotions ('Angry', 'Disgust', 'Happiness', 'Sadness' and 'Surprise'), in 5 different situations. We test our approach on each subject for all emotions which provides a total number of 45 audio-visual test sequences and 180 training sequences. The natural database consists of 3 subjects with 6 test sequences for 2 emotion classes (‘Happy’ and ‘Sad’).

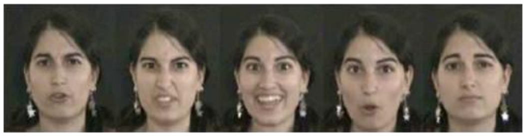


Figure 5-1 (a): Posed audio-visual database selected from ‘eNTERFACE 2005’.

Figure 5-1(b): Spontaneous audio-visual database selected from ‘Belfast Naturalistic Database’.

5.2 Training

We train the visual component of our system using a key frame approach, where the task is to choose exemplars from the video sequence that reflect peak intensities in each emotion class. The initial attempt for the key frame selection is done manually, to ensure the presence of useful information in the chosen frame. We then implement three different methods for automatic key frame selection. They are:

I. A heuristic based approach

II. A semi-supervised approach and

III. An unsupervised approach

I. A heuristic based approach:

This method for frame selection was proposed by Yongjin et al. [5]. The heuristic is based on the intuition that the peak facial expressions occur at the maximum audio intensity instances. The visual component of our system is trained using the corresponding maximum audio intensity frames.

II. A semi-supervised approach:

In this approach, we use a combination of image database along with a visual sequence database. The image database has photographs of subjects expressing the required facial expressions with full emotional intensity [24]. The visual sequence database consists of subjects deliberately posing the facial expressions beginning with the neutral and ending with the peak intensity [25]. We use only the last frame i.e. the frame comprising the peak intensity, facial expression for the training purpose. Examples of the two databases are presented in the Figure 5-2.

Figure 5-2 (a): Japanese Female Facial Expression Database [24]

Figure 5-2 (b): Cohn-Kanade database [25]. © Jeffrey Cohn.

The two labeled databases shown above are used to compute the centre of clusters for 5 emotion classes. The labeled audio-visual sequences are then processed to determine the distance of each frame in the sequence to the centre of the corresponding class cluster. The frame with the minimum distance from the centre is selected for the training process. The distance of each frame in the sequence is measured using two distance metrics:

(i) Linear distance metric

$$d(x,y) = \left||x-y|\right|^2 \tag{5.1}$$

where, $d(x,y)$ is the distance between vectors *x* and *y* which represent the feature vector of the frame at the centre of the cluster and the feature vector of the frame from the audio-visual sequence respectively.

(ii) Entropy based distance metric

$$d(x,y) = \frac{[H(x,y) - I(x,y)]}{H(x,y)} \tag{5.2}$$

where, $H(x,y)$ and $\mathrm{I}(x,y)$ are the joint entropy and the mutual information of vectors *x* and *y* respectively, having the same representations as in Equation (5.1)

III. An unsupervised approach:

This method is useful when there is no prior information about the emotions present in a given training sequence, and also if there is no database available for selecting the training examples. We assume that the training sequences are pre-segmented and they consist of only one emotion class per sequence. A clustering algorithm is then applied to separate the frames in the sequence into emotion-containing frames and non-emotion-containing frames. These frames will be referred to as 'emotion frames' and 'non-emotion frames' respectively. The cluster containing the largest number of continuous frames is considered to be the emotion frame cluster, since the audio-visual sequences are pre-segmented into five emotion classes with each sequence comprising the majority of the frames from the corresponding emotion class. The frame which is at the minimum distance to the centre of this cluster is selected for training the visual component of our bimodal system. In order to deal with the outliers, e.g. only one frame in the cluster, a re-clustering technique is used where the outlier frames are removed and the clustering process is performed iteratively until a minimum pre-defined number of continuous frames are obtained in each of the clusters. The pre-defined number of continuous frames is based on the observation that the emotions in the database lasted on an average between 0.25 and 2 seconds. The frame rate of the audio-visual sequences is 25 fps and hence the number of continuous frames is set to the lower threshold of 0.25 second i.e. a minimum of 5 continuous frames per cluster. A pseudo code for the unsupervised approach for training the visual component of the system is presented in Figure 5-3.

```
% VISUAL ANALYSIS
Cluster (Emotion/Non-Emotion Frames): C1,C2

while ((C1 < pre-defined number) && (C2 < pre-defined number))
   Remove outlier frames
   Re-cluster
end

for (k = 1:n_frames)
   Set counters for continuous number of frames in each cluster
   Bubble sort maximum number of continuous frames belonging to one cluster
end

Pick minimum distance frame from continuous longest cluster
Train visual SVM
```

Figure 5-3: Pseudo code for unsupervised training approach

Examples of results obtained from the clustering algorithm are presented in the Figure 5-4. The pair of images shown in this figure represents the information and the non-information frames which are at the minimum distance from the centre of emotion and non-emotion clusters respectively.

Figure 5-4: Information (Emotion) and Non-Information (Non-emotion) Frames

A comparison of the frames selected using the three training methods discussed above along with the manually selected frames is illustrated in Figure 5-5.

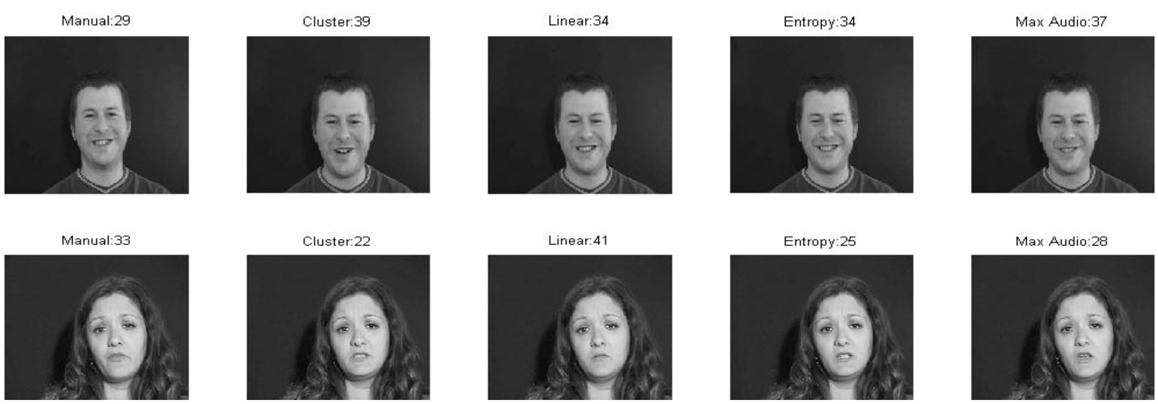


Figure 5-5: Comparison of Frame Selection Methods using: Manual, Unsupervised (Clustering), Semi-supervised (Linear and Entropy) and Heuristic (Maximum Audio Intensity) approaches.

The standard deviation of absolute differences between the frame numbers selected by our proposed training methods and the manually selected frames are: 14.7, 12.9, 13.5 and 12.7 for unsupervised (clustering), semi-supervised – linear metric, semi-supervised – entropy metric and heuristic – maximum audio intensity methods respectively.

The audio component of our system is trained based on global statistics of acoustic features obtained from the speech signal as discussed in Chapter 3. The pre-processing step removes the leading and trailing silent edges of the auditory signal and all the features are normalized with respect to each subject in the database.

5.3 Testing

The testing process for the recognition using visual information is carried out in a similar way to the training process. We evaluate the three approaches described in the previous section. The first test is performed using the heuristic-based approach. For this method, the key frame is selected from the test sequence based on the maximum audio intensity criterion.

The second method for testing is the semi-supervised approach. The training frames obtained based on this approach is used to determine the centre of clusters for different emotion classes. The test frames are selected by measuring their distance from the centre of these clusters. The only difference in the test process when compared to the training process is that, the test sequences are not labelled. Hence, the distance of all the frames in the test sequence is calculated with respect to all the cluster centres. The frames which are at the minimum distance to each cluster are selected and the frame which has the minimum distance to the selected frames is chosen for final testing.

The last test is based on the unsupervised approach. In this method, the clustering process is repeated for the test sequences to divide the frames into emotion frames and non-emotion frames. The minimum distance frame belonging to the longest continuous cluster is considered to be the emotion-frame which is selected for testing.

The audio-based emotion recognition is performed in the same way as the training process and the acoustic features obtained for the test sequences are also normalized for each subject in the database. The audio and visual data are combined using two different fusion techniques: (i) Feature-Level fusion and (ii) Score-Level fusion.

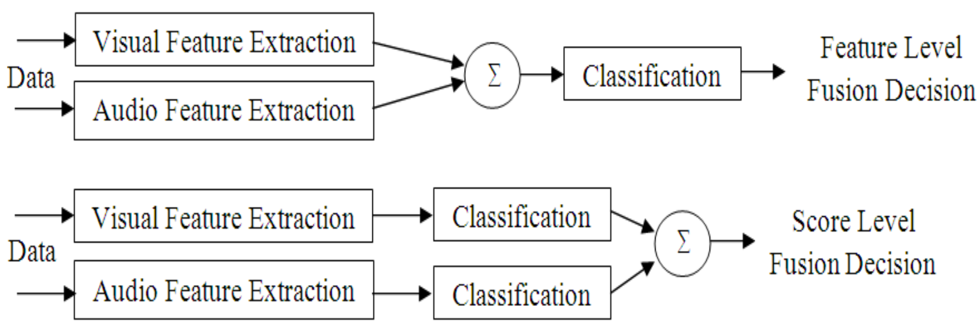


Figure 5-6: Feature-Level (above) and Score-Level Fusion (below) Techniques

The feature-level fusion is performed by concatenating the visual feature vector of the key frame with the global audio feature vector. This large feature vector is reduced in dimensionality using the RFE method, and classified using a multi-class SVM. The second method for combining the two modalities is based on the scores or the probability estimates obtained after individually classifying the two modalities using SVM. We combined the scores of the two modalities using the product rule based on the assumption that the features obtained from the two modalities are independent of each other. The combined score S for a class d using product rule is given by the following equation:

$$S(d) = \frac{\prod_m p_m(d)}{\sum_c \prod_m p_m(c)}, \tag{5.3}$$

where, $p_m(d)$ is the probability of the test example belonging to class d for mode m and class c. The number of modes in the present case is 2 (audio and visual) and the number of classes are 5 ('Angry', 'Disgust', 'Happiness', 'Sadness' and 'Surprise'). The results obtained based on the two fusion techniques for the three approaches, discussed above are presented in the Table 5-1. The percentage of correct recognition rate for audio-based emotion recognition is 53% which is common to all the results presented below.

Table 5-1: (% Correct) Recognition Rates for Comparing Fusion Techniques

	Unsupervised Clustering	**Semi-supervised (a) Linear**	**Semi-supervised (b) Entropy**	**Heuristic (Max Audio Intensity)**	**Manual**
Visual Only	80	56	67	64	76
Feature Level	80	60	71	80	76
Score Level	84	58	76	80	82

The score-level fusion performs better than the feature level fusion for all the training methods except for the semi-supervised training method based on the linear distance metric. The best (correct) recognition rate is obtained in case of the clustering algorithm with 84% for score-level fusion which is marginally better than the 82% (correct) recognition rate for the manual process. The only disadvantage of the unsupervised clustering approach is that the seed of the cluster centre is randomly selected, which generates different training examples at each run of the algorithm. The next best result is obtained using the heuristic based approach followed by the semi-supervised approaches.

5.4 Temporal Analysis

The single-frame based visual emotion recognition process discussed above, involves the frame selection process even during the testing phase. These computations are avoided by using a temporal analysis scheme where all the frames of the test sequence are considered for visual based emotion classification. This process involves the aggregation of the scores of the visual frames after individually classifying them using a multi-class SVM. The scores are aggregated using the sum rule based on the assumption that the consecutive frames in the sequence are related to each other, unlike the independence assumption used to justify the product rule. The sum rule for a class *d* is given by the equation:

$$S(d) = \frac{\sum_m p_m(d)}{\sum_c \sum_m p_m(c)}, \quad (5.4)$$

where, the notations remain the same as in Equation (5.3), i.e. *m* represents the mode for fusion and *c* represents the emotion class.

A comparison of four cases is presented in Figure 5-7: (i) emotion recognition based on single frame analysis, (ii) single frame analysis using score level fusion, (iii) temporal analysis using visual data only, and (iv) temporal analysis using score-level fusion. The combined audio-visual emotion recognition rates obtained using the single-frame based approach outperforms the joint recognition rates obtained using the temporal analysis of the visual data for all the methods except the semi-supervised method using the linear distance metric.

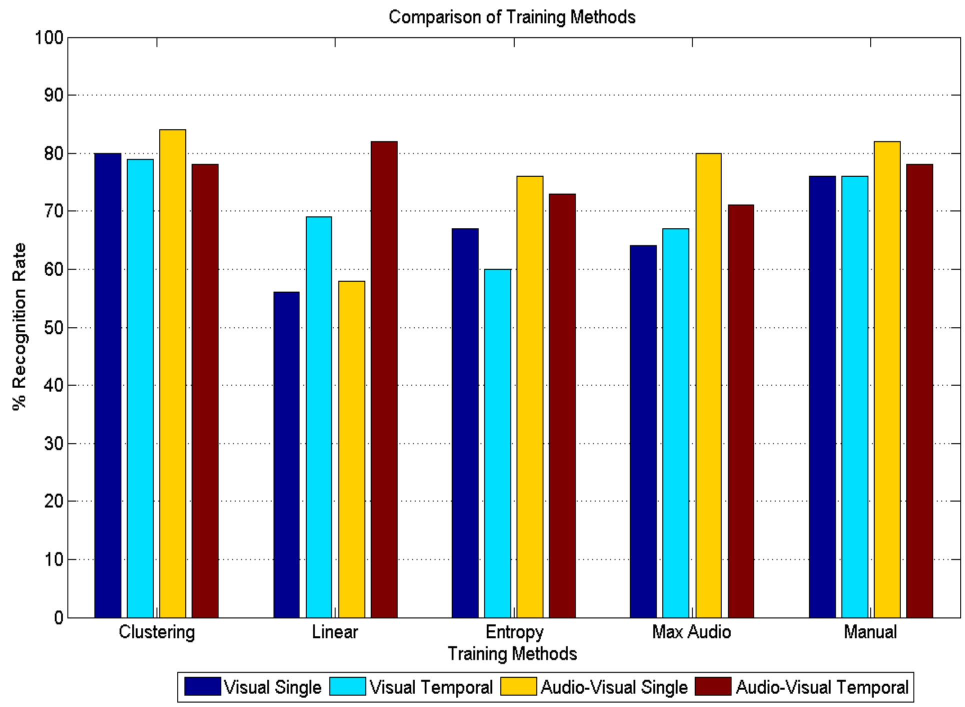
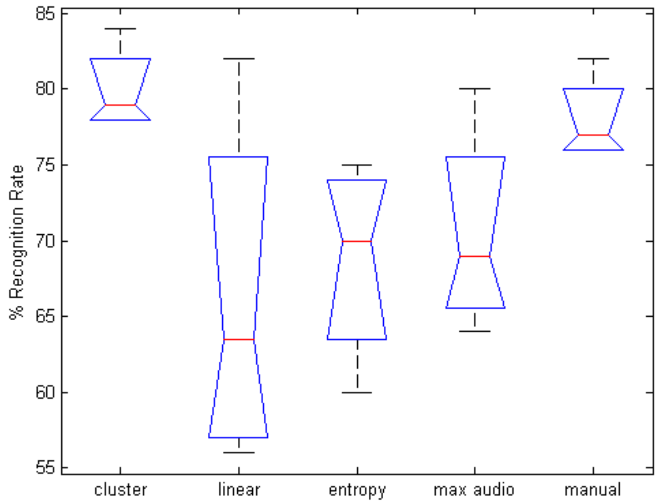


Figure 5-7: Comparison of Training Methods

A quantitative analysis of the different training methods was performed using one way Analysis of Variance statistics (ANOVA). The result of ANOVA test is presented in Figure 5-8 in form of a box plot with $p = 0.06$. This suggests that the four methods differ only marginally. However, it is evident from the plot that the clustering method is closest to manual method of sampling frames from the video sequences.



ANOVA TABLE					
Source	SS	Df	MS	F	Prob>F
Columns	573.7	4	143.42	2.84	0.061
Error	758.5	15	50.56		
Total	1332.2	19			

Figure 5-8: Statistical significance test using one way ANOVA (p=0.06)

The manual training method is considered as the reference and the pair-wise significance of each method was obtained and these results are presented in Table 5-2. Hence it can be concluded that entropy method was significantly different from the manual method and not capable of reproducing the same results as obtained by manual sampling.

Table 5-2: Pair-wise comparison of 4 training methods against the manual method

Methods	**Clustering**	**Linear**	**Entropy**	**Maximum Audio**
p values	0.35	0.10	0.04	0.09

5.5 Results

The generalization ability of the four methods used for the visual component of our system is estimated in terms of the number of support vectors required for training purposes. These support vectors are obtained from a total number of 180 training examples used for classifying 5 emotion classes. The semi-supervised training method based on the entropy distance metric has the smallest number of support vectors (163/180) and the largest number of support vectors are used by the heuristic based approach (175/180) for multi-class Support Vector Machine. The percentages of correct recognition rates obtained from all the methods discussed above are summarized in Table 5-2. The number of support vectors selected for each method is listed in column 2 of this table.

Table 5-3: Summary of (% Correct) Recognition Rates

	NSV	**Mode**	**Visual Single**	**Visual Temp**	**AV Score Single**	**AV Score Temp**	**AV Feature**
Unsupervised (Clustering)	169	80	80	78	84	78	80
Semi-supervised (Linear)	167	78	56	69	58	82	60
Semi-supervised (Entropy)	163	69	67	60	76	73	71
Heuristic (Max Audio Intensity)	175	80	64	67	80	71	80
Manual	170	71	76	76	82	78	76

The percentage of correct recognition rates for the manual case from the above Table are represented in the graphical form in Figure 5-8.

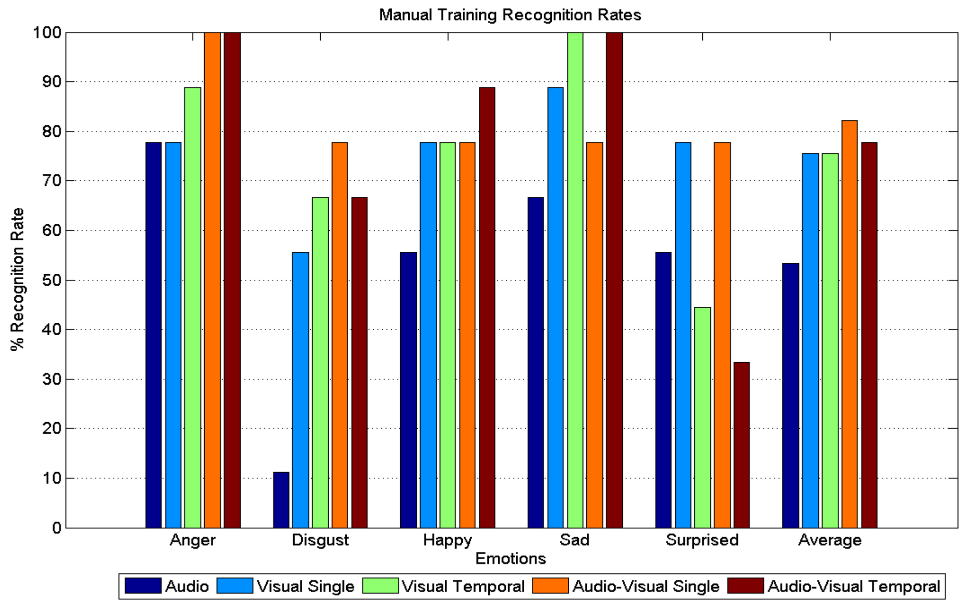


Figure 5-9: (% Correct) Recognition Rates for Manual Training Approach

The graphical representation of the confusion matrix for the single frame manual training approach along with the audio analysis based confusion matrix is shown in Figure 5-9.

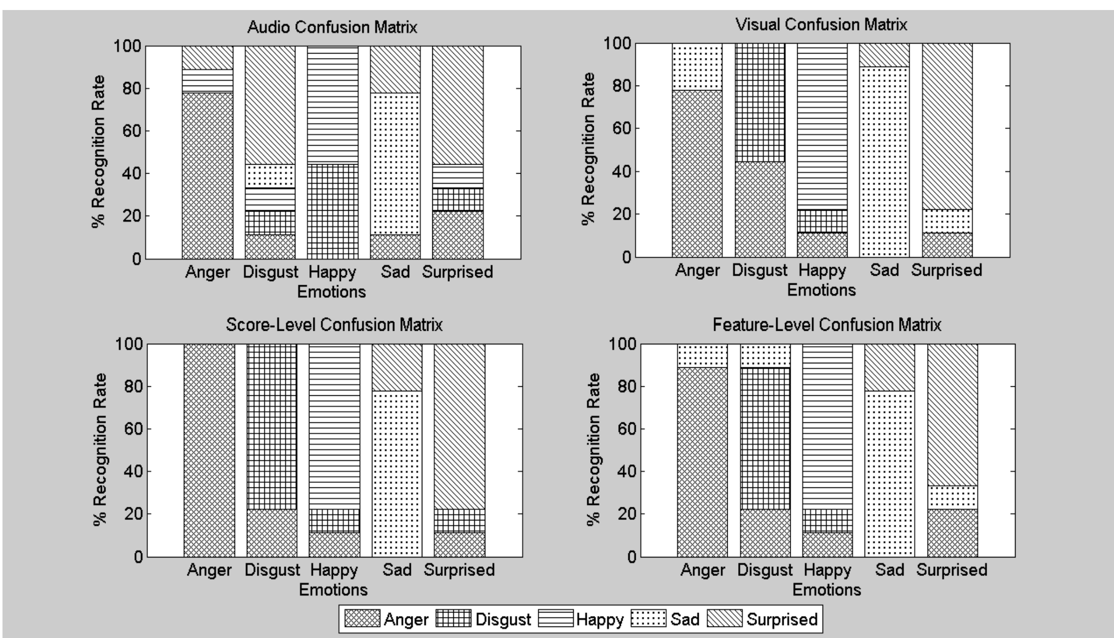


Figure 5-10: Confusion Matrix for the Manual Training Approach

5.6 Natural Database Results

The second set of experiments is performed on the natural database. The number of examples per emotion class in this database is not uniformly distributed. Hence, we use the image databases [24], [25] and the manual frames selected from the posed audio-visual database for training the visual component of our system based on the semi-supervised training method. Our audio system is trained using the acoustic global features of the posed audio-visual database. We do not use any data from the natural sequences for neither training the audio nor the visual systems. Therefore, the test performed on the natural database, are completely person-independent. We however, train the system for identifying only two emotions: 'Happy' and 'Sad' in the natural sequences due to the limited access to the database. The individual visual and audio correct recognition rates are 50% and 67% respectively. The combined audio-visual accuracy at the score level for two-class person-independent emotion recognition is 83%. The confusion matrix for these results in terms of percentage of correct recognition rates is presented in Figure 5-10.

	Audio		Visual		Audio-Visual (Score-level)	
	H	S	H	S	H	S
H	67	33	33	67	67	33
S	33	67	33	67	0	100

Figure 5-11: Confusion Matrix for 2 natural emotions: Happy (H) and Sad (S)

Chapter 6

Conclusion and Future Work

6.1 Overview

This thesis presented a bimodal emotion recognition system using audio and visual information channels in a video stream. We proposed and implemented a naïve approach for the analysis of the audio-visual sequences in order to reduce the computation complexity of the system. The audio component of the system was simplified by obtaining global features instead of local noisy measurements. The visual component of the system was based on the features extracted from a single frame rather than extracting continuous features from all the frames using traditional tracking algorithms. We proposed three different methods for frame selection from the audio-visual sequences, namely, a heuristic approach, a semi-supervised approach and an unsupervised approach. A sequential analysis scheme for the visual component of the system was implemented to compensate for the loss of the temporal information at the feature level. The experimental results suggested that the temporal analysis improved the recognition rates only for the semi-supervised training approach. The two modalities were combined based on two standard approaches: feature-level fusion and score-level fusion techniques.

6.2 Visual Analysis

We implemented a single frame based approach for analyzing the facial expressions of the subjects in an audio-visual sequence. We proposed and demonstrated that the short time facial expressions can be captured in a single frame for training purposes. We implemented three different training approaches for the visual component of our system. These were: a heuristic approach, a semi-supervised and an unsupervised approach. The heuristic approach performed reasonably well for the posed audio-visual sequences, but showed poor performance for the natural sequences. This suggested that the natural facial expressions are subtle and may not be present at the maximum audio intensity instances. The unsupervised approach for obtaining the information and non-information frames from the audio-visual sequence was based on a clustering algorithm for which a random seed was generated at each attempt. This method provided the best recognition rate however the results were unstable. The semi-supervised method proved to be the most consistent and effective solution for selecting the training examples and its recognition rates were at par with the manual process.

We obtained the visual features from the selected frames using a bank of Gabor filters. The advantage of using the Gabor filters for feature extraction for facial expression recognition is that they preserve the local spatial relations between facial features and eliminate the need for explicitly tracking each facial point. The cost of this feature extraction method is the high dimensionality of the feature vectors obtained at each instance of time. The dimensionality of the feature vector is reduced using a feature selection method. We evaluated four different feature selection methods by comparing their cross-validation results and obtained the optimal number of features and the best feature selection method for classification. This process suggested that the accuracy as measured by cross-validation does not improve by increasing the number of features. Therefore, it is necessary to select optimal number of features for classification.

We recognized the emotions in the test sequences by classifying each frame and obtaining a final result based on different criteria such as, the maximum vote wins, a sum rule based aggregation of scores and single frame-based classification. The single frame approach outperformed the other two methods. The temporal analysis scheme using the sum rule did not perform well due to the fact the emotions fade away with time and aggregating the classification score of all the frames can thus reduce the recognition rate. Hence, we instead tried a method where we sequentially combined the classification score of the frames around the key representative frames. This method performed better for the natural sequences where the emotions are non-uniformly spread over longer periods of time.

6.3 Audio Analysis

The audio component of the system used the global statistics of the acoustic features for emotion recognition. This analysis helped in dealing with the variable length of the input speech signal. The features obtained from the pitch and intensity of the speech signal contributed the most to the emotion recognition rates. It was also observed that the MFCCs had an important role in improving the recognition rates. The key improvement was observed when the features were normalized with respect to the subjects. This made the process person-dependent in case of the posed database. The effect of the normalization process was not evident in the natural database as there were only two emotions classes (‘Happy’ and ‘Sad’) which could be easily distinguished.

6.4 Future Work

The missing link in the present system is the need for an improved method for aggregating the instantaneous visual classification results. This can be achieved by using algorithms such as Adaboost for obtaining the weights for each visual frame for the temporal analysis. The temporal relation between the classification scores for the visual frame in a given sequence can also be learned using a Hidden Markov Model (HMM) as implemented in [10]. The authors of this work, however, analyzed only the facial expression and not the audio-visual information. Hence, it would be interesting to obtain the combined scores from the two modalities and learn the temporal relations using an HMM.

As mentioned earlier in the thesis, the motivation for this work is derived from the Tele-Health care application where we would like to automatically analyze the emotional states of a patient in response to the type of interventions provided by a nurse practitioner. The automatic analysis thus obtained can be used to provide an offline feedback to the nurse in order to improve the nursing intervention protocols. It can also be useful in automatic generation of interventions learned from the patient-nurse interactive conversations.

Finally, in conclusion, it is observed that the emotion recognition is substantially facilitated by the combination of audio and visual data. This result suggests that the Tele-Health care application discussed in this thesis may be feasible in the near future.

BIBLOGRAPHY

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor. "Emotion Recognition in Human-Computer Interaction". *Signal Processing Magazine*, volume 18, pages 32-80, January 2001.

[2] B. Schuller, G. Rigoll and M. Lang. "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture". *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, no.1, pages I-577-80, May 2004.

[3] I. Murray and J. Arnott. "Synthesizing Emotions in Speech: Is it time to get excited?" *In Proceedings of 4th Int. Conf. Spoken Language Processing*, pages 1816-1819, 1996.

[4] B. Schuller, G. Rigoll and M. Lang. "Hidden Markov Model-based Speech Emotion Recognition". *IEEE International Conference on Multimedia and Expo*, volume 1, no.1 pages I-401-4, July 2003.

[5] Y. Wang and L. Guan, "Recognizing Human Emotional State from Audio-Visual Signals". *IEEE Transactions on Multimedia*, volume 10, no.4, pages 659-668, June 2008.

[6] V. Sethu, E. Ambikairajah, J. Epps, "Speaker Normalisation for Speech-Based Emotion Detection". *IEEE International Conference on Digital Signal Processing*, pages. 611-614, July 2007.

[7] P. Ekman and W. Friesen. “Facial Action Coding System: A Technique for the Measurement of Facial Movement”. *Consulting Psychologists Press*, Palo Alto, CA, 1978.

[8] M. S. Bartlett, G. Littlewort, I. Fasel and J. R. Movellan. “Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction”. *In Proceedings of Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction,* June 2003.

[9] Cohen, A. Garg and T. S. Huang. “Emotion Recognition from Facial Expressions using Multilevel HMM”. *In Neural Information Processing Systems*, 2000.

[10] M. Valstar and M. Pantic. “Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics”. *In IEEE International Workshop on Human Computer Interactions,* LNCS, pages 118-127, 2007.

[11] J. Platt. “Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods”. *In A. Smola, P. Bartlett, B. Schölkopf and D. Schuurmans (Eds.), Advances in large margin classifiers. Cambridge: MIT Press,* 2000.

[12] G. Castellano, L. Kessous and G. Caridakis. “Multimodal Emotion Recognition from Expressive Faces, Body Gestures and Speech”. *In Proceedings of the Doctoral Consortium of 2nd International Conference on Affective Computing and Intelligent Interaction,* 2007.

[13] A. Corradini, M. Mehta, N. O. Bernsen, and J. C. Martin. “Multimodal Input Fusion in Human Computer Interaction on the Example of the On-Going Nice Project”. *In*

Proceedings of the NATO-ASI conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management, Yerevan, Armenia, 2003.

[14] L.C. De Silva, T. Miyasato and R. Nakatsu. "Facial Emotion Recognition using Multi-Modal Information". *Proceedings of International Conference on Information, Communications and Signal Processing*, 1997.

[15] L. S. Chen, T. S. Huang, T. Miyasato and R. Nakatsu. "Multimodal Emotion/Expression Recognition". *In Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, 1998.

[16] D. Datcu and L.J.M. Rothkrantz. "Semantic Audio-Visual Data Fusion for Automatic Emotion Recognition". *Euromedia'2008 Porto*, ISBN 978-9077381-38-0, pages 58-65, April 2008.

[17] M. Song, C. Chen and M. You. "Audio-Visual based Emotion Recognition using Tripled Hidden Markov Model". *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V-877-80, May 2004.

[18] M. Paleari and B. Huet. "Toward Emotion Indexing of Multimedia Excerpts". *In Proceedings of International Workshop on Content Based Multimedia Indexing*, June, 2008.

[19] C. Busso, Z. Deng, and S. Yildirim. "Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information". *In Proceedings of ACM 6th International Conference on Multimodal Interfaces*, 2004.

[20] E. Douglas-Cowie, R. Cowie and M. Schröder. "A New Emotion Database: Considerations, Sources and Scope". *Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[21] B. Kort, R. Reilly, R. W. Picard. “An Affective Model of Interplay between Emotions and Learning: Reengineering Educational Pedagogy-Building a Learning Companion”. *IEEE International Conference on Advanced Learning Technologies*, page 43, 2001.

[22] Z. Zeng, Y. Hu, G.I. Roisman, Z. Wen, Y. Fu, and T.S. Huang, “Audio-visual Emotion Recognition in Adult Attachment Interview”. *International Conference on Multimodal Interfaces,* pages 139-145, 2006.

[23] C. Lisetti, F. Nasoz, C. LeRouge, O. Ozyer and K. Alvarez. “Developing Multimodal Intelligent Affective Interfaces for Tele-Home Health Care”. *International Journal of Human-Computer Studies*, 2003.

[24] M. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba. “Coding Facial Expressions with Gabor Wavelets”. *In Proceedings, IEEE International Conference on Automatic Face and Gesture Recognition*, April 1998.

[25] T. Kanade, J.F. Cohn and Y. Tian. “Comprehensive database for facial expression analysis. *In Proceedings of, IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46-53, 2000.

[26] MMI Face Database: http://www.mmifacedb.com/

[27] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss. “A Database of German Emotional Speech”. *Interspeech*, 2005.

[28] O. Martin, I. Kotsia, B. Macq and I. Pitas. “The eNTERFACE’05 Audio-Visual Emotion”. *IEEE Workshop on Multimedia Database Management,* April 2006.

[29] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang. “Facial expression recognition from video sequences: temporal and static modeling”. *Computer Vision Image Understanding*, pages 160-187, July 2003.

[30] C. Padgett, G. Cottrell, and B. Adolps. “Categorical perception in facial emotion classification”. *In Proc. Cognitive Science Conference*, volume 18, pages 249-253, 1996.

[31] C. Hu, Y. Chang, R. Feris, and M. Turk. “Manifold based analysis of facial expression”. *In Proceedings of Computer Vision and Pattern Recognition Workshop*, Volume 5, page 81, 2004.

[32] G. Littlewort, M. Bartlett, I. Fasel, J. Chenu, and J. Movellan. “Analysis of machine learning methods for real-time recognition of facial expressions from video”. *Computer Vision and Pattern Recognition*, 2004.

[33] F. Torre, J. Campoy, Z. Ambadar and J. F. Cohn. “Temporal Segmentation of Facial Behavior”. *IEEE International Conference on Computer Vision*, pages 1-8. 2007.

[34] P. Boersma and D. Weenink (2009). Praat: doing phonetics by computer (Version 5.1.12) [Computer program]. Retrieved August 4, 2009, from http://www.praat.org/

[35] X. Huang, A. Acero and H.W. Hon. “Spoken language processing: A guide to theory, algorithm, and system development”. - Prentice Hall PTR Upper Saddle River, NJ, USA. 2001.

[36] P. Viola and M.J. Jones, “Robust real-time face detection”. *International Journal of Computer Vision*, pages 137- 154, 2004.

[37] Intel, “Open CV: Open source Computer Vision Library”, http://www.intel.com/research/mrl/research/opencv/.

[38] A. K. Jain and F. Farrokhnia. "Unsupervised texture segmentation using Gabor filters". *IEEE International Conference on Systems, Man and Cybernetics*, pages 14-19, 1990.

[39] J. Ilonen, J. K. Kamarainen and H. Kalviainen. “Efficient Computation of Gabor Features”. *Research Report 100, Lappeenranta University of Technology, Department of Information Technology*, 2005.

[40] A. Kapoor, W. Burleson and R. W. Picard. “Automatic Prediction of Frustration”. *International Journal of Human Computer Studies*, vol. 65, no. 8, pages 724-736, 2007.

[41] Z. Zeng, M. Pantic, G. Roisman and T. Huang. “A Survey of Affect Recognition Methods: Audio, Visual and Spontaneous Expressions”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, January, 2009.

[42] H. B. Deng, L. W. Jin, L. X Zhen and J. C. Huang. “A New Facial Expression Recognition Method Based on Local Gabor Filter Bank and PCA plus LDA”. *International Journal of Information Technology*, vol. 11, no.11, pages 86-96, 2005.

[43] M. J. Lyons, J. Budynek, A. Plantey and S. Akamatsu. “Classifying Facial Attributes Using a 2-D Gabor Wavelet and Discriminant Analysis”. *IEEE International Conference on Automatic Face and Gesture Recognition*, page 202, 2000.

[44] L. Wang. “Support Vector Machines: Theory and Applications”. ISBN: 3540243887, Springer-Verlag, New-York, Inc. 2005.

[45] SPIDER, Machine Learning Toolbox:
http://www.kyb.tuebingen.mpg.de/bs/people/spider/main.html

[46] J. Weston, A. Elisseeff and B. Scholkop. “Use of the Zero-Norm with Linear Models and Kernel Methods”. *The Journal of Machine Learning Research,* vol. 3, pages 1439 - 1461, 2003.

[47] O. Martin, J. Adell, A. Huerta, I. Kotsia, A. Savran and R. Sebbe. “Multimodal Caricatural Mirror”. *eINTERFACE'05 - Summer Workshop on Multimodal Interfaces,* 2005.

[48] P.S. Bardley, O. L. Mangasarian. “Feature Selection via Concave Minimization and Support Vector Machines”. *International Conference on Machine Learning*, pages 82 – 90, 1998.

[49] M Zaffalon and M Hutter.“Robust Feature Selection by Mutual Distributions”. *International Conference on Uncertainty in Artificial Intelligence*, pages 577 – 584, 2002.

[50] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. “Gene selection for cancer classification using support vector machines”. *The Journal of Machine Learning*, Springer, vol. 1, no. 3, pages 389 – 422, 2002.

KEY TO ABBREVATIONS

COPD: Chronic Obstructive Pulmonary Disease

MFCC: Mel Frequency Cepstrum Coefficient

SVM: Support Vector Machines

GMM: Gaussian Mixture Model

HMM: Hidden Markov Model

AU: Action Unit

AdaSVM: Adaboost SVM

FFT: Fast Fourier Transform

IFFT: Inverse Fast Fourier Transform

RBF: Radial Basis Function

FSV: Feature Selection via Concave minimization

L0: dual Zero-norm minimization

MUTINF: Mutual Information

RFE: Recursive Feature Elimination

ANOVA: Analysis of Variance