

**BEYOND TRADITIONAL EMOTION
RECOGNITION**

RUCHIR SRIVASTAVA

(B.Tech., IIT Roorkee, India)

**A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF
PHILOSOPHY
DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE**

2012

Acknowledgments

First and foremost I thank the Supreme Lord for giving me the intelligence to do research. Without the mental and physical tenacity that He has given me, I could not have even thought of doing a PhD.

I am deeply grateful to my teacher in IIT, whom we lovingly call 'Sir', for giving me a vision of using my education for the welfare of society. Doing a PhD is a major milestone towards that vision. His guidance at every step of my life has instilled a sense of responsibility in me.

I would like to express my sincere thanks to my supervisors Dr. Yan Shuicheng and Dr. Terence Sim for accepting me as their PhD student. They've always encouraged me to think along newer research areas, which has helped me to develop my research aptitude.

I am indebted to Dr. Sujoy Roy for his guidance in understanding what research is all about. At a difficult juncture in my research, without his guidance and support, I would have lost my motivation to continue with the PhD. Throughout my research; he has kept me focused on my work through helpful discussions which have shaped most of my research work.

I am grateful to Dr. Surendra Ranganath for taking me as his student at a very crucial stage and giving the much needed support to me. Discussions with him helped me prepare for my qualifying examination.

I am thankful to my lab-mates, friends and lab officers Jack and Francis for providing me facilities for conduct my research. Last but not the least, I

wish to express my gratitude for my parents and brother who have stood by me through all the ups and downs in my life so far.

Contents

1	Introduction	1
1.1	About Emotion Recognition (ER)	1
1.2	How to Define Emotions?	2
1.3	Applications of Emotion Recognition	3
1.3.1	Biometrics	3
1.3.2	Medicine	4
1.3.3	User Response	5
1.3.4	Robotics	5
1.3.5	Surveillance	6
1.4	Proposed Application of ER on Difficult Data	6
1.5	Contributions of the Thesis	7
1.6	Publications	8
2	Literature Review	9
2.1	Databases for Emotion Recognition	9
2.1.1	Visual Databases	10
2.1.2	Audio Databases	11
2.1.3	Audio-Visual Databases	13
2.2	Emotion Recognition	14
2.2.1	Using Visual Clues	15
2.2.2	Using Clues from Speech	17

2.2.3	Using Multimodal Clues	19
2.2.4	Detailed Categorization of Works on ER	21
2.2.5	Type of the Data Used	22
2.2.6	Way to Define Expressions	26
2.2.7	Directions Beyond Traditional Emotion Recognition	31
2.3	Automated Personality Assessment	33
3	Utilizing 3D information for Facial Expression Recognition	35
3.1	Database Used	36
3.2	3D Residues for Subject Dependent FER	37
3.2.1	Problem Formulation	37
3.2.2	Feature Extraction	45
3.2.3	Classification	47
3.2.4	Experimental Results and Discussions	48
3.2.5	Limitations	58
3.3	Deformation Modeling: Subject Independent FER	59
3.3.1	Background	59
3.3.2	Feature Extraction	60
3.3.3	Experiments	62
4	Bimodal Spontaneous ER Applied to Multi-actor ER	67
4.1	Emotion Recognition in Movies	67
4.1.1	Facial Expression Recognition	71
4.1.2	Lexical Analysis	73
4.1.3	Fusing Clues from FER and Lexical Analysis	77
4.1.4	Dataset	80
4.1.5	Experiments	81
4.2	Multi-actor ER vs. Single Actor ER	87
4.2.1	Facial Expression Recognition (FER)	89

4.2.2	Lexical Analysis of Dialogs	91
4.2.3	Fusing Visual and Lexical Cues	92
4.2.4	Experimental Results and Discussions	94
5	ER Applied to Automated Personality Assessment	98
5.1	Background	101
5.1.1	Models of Personality	101
5.1.2	Five Factor Model (FFM) & the BFI	102
5.2	Multimodal Feature Extraction	103
5.2.1	Feature Extraction	104
5.2.2	Emotion Feature Estimation	107
5.3	Automating Answering of Personality Questionnaires	109
5.3.1	Features for Regression	109
5.3.2	F2A by Sparse and Low-rank Transformation	111
5.3.3	A2P using BFI scoring scheme	118
5.4	Model for scoring the Questionnaire	119
5.4.1	CRF Model to Predict Personality Scores	119
5.5	Experiments	121
5.5.1	Dataset	121
5.5.2	Predicting Answers using F2A Transformation	123
5.5.3	Accuracy for Personality Prediction	125
5.5.4	Personality Prediction Using Learnt CRF model	127
6	Conclusion and Future Works	129
	Bibliography	134

Summary

This thesis investigates the problem of recognizing human emotions experienced in real-life which has two aspects, (1) facial appearance can be affected by low face resolution, non-frontal pose, significant head motion etc. and (2) low intensity of emotions. An attempt to deal with these difficulties takes the proposed work beyond the traditional works on Emotion Recognition (ER) which usually do not consider these problems. Although the data we have used are not captured from real-life but they are closer to real life as compared to lab-recorded data.

In order to deal with the above mentioned difficulties, additional information about emotions is acquired in the following ways: (1) using 3D instead of conventional 2D information for recognizing facial expressions which is one of the modalities contributing to ER, and (2) fusing information from multiple modalities. Figure 1 shows the flow of this thesis.

As compared to 2D Facial Expression Recognition (FER), researchers have identified practical advantages of 3D FER [211]. In this thesis, the idea of optical flow in 2D has been extended to 3D and the resultant features have been called residues. The proposed method is found to perform better than similar state-of-the-art approaches. However, computing residues requires a neutral face model which may not be always available. Another approach is presented to deal with this problem. Evaluation has also been made on low intensity expressions which characterize a difficult data. Experimental results show that

Problem: Dealing with difficult data

Proposed solution: Using more information about the expression.
3D instead of 2D

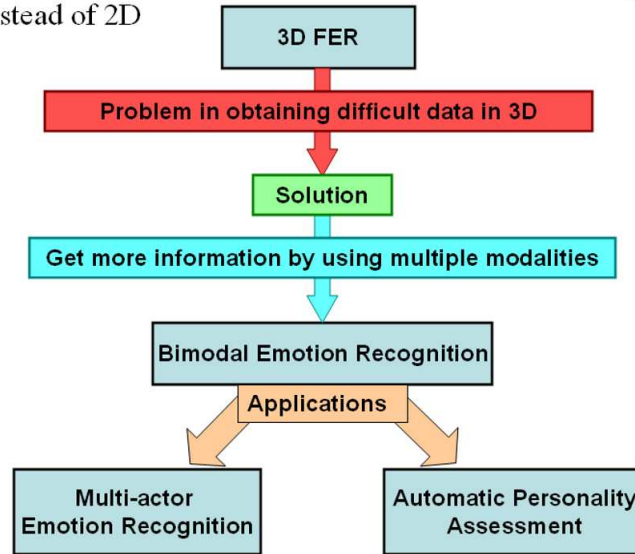


Figure 1: Flow of the work covered in this thesis

adding the additional depth information (characteristic of 3D) improves the recognition rate. Since such difficult data is not easily available in 3D, 2D data in the form of movie clips has been used for further research. The approaches developed in 2D, form a basis for extension to 3D data.

Another way of getting additional information is to use multiple modalities, each of which is a source of information about the emotions of the person. In this part of the thesis, clues from facial expressions and spoken words have been fused together for recognizing emotions of movie characters. An algorithm for sentiment analysis from movie reviews has been improved to deal with movie dialogs. Fusion is performed by a novel dynamic weighting methodology which improves the ER performance, as compared to using each of the two clues alone. The bimodal approach for ER in movies has been applied for fusing emotions of multiple characters to predict emotion conveyed by the movie scene.

Another application of ER in movies has been proposed in the form of automatic personality assessment by first answering a personality questionnaire,

the shorter version of the Big-Five Inventory (BFI-10 or BFI). BFI is answered using psychology based multimodal features including features extracted using emotion recognition. The features are mapped to BFI answers using a novel regression formulation based on sparse and low rank transformation.

To automatically predict personality scores from answers, a novel approach is proposed which is applicable to a wide range of questionnaires instead of being limited to only one questionnaire.

List of Tables

2.1	Classification of the works on ER	30
3.1	Details about training and test datasets	47
3.2	Confusion matrices for Set 1	49
3.3	Confusion matrices for Set 2	50
3.4	Set 1: Comparison of our method with [172]	52
3.5	Set 2: Comparison of our method with [172]	52
3.6	Comparing SVM and LDA: Confusion matrices	54
3.7	Comparing SVM and LDA: Recognition Rates	55
3.8	Efficacy of using 3D flow than just using 2D optical flow.	55
3.9	Comparison of performances of the related works	56
4.1	Performance of ER using SO.	74
4.2	Patterns of tags for extracting two-word or single word phrases.	75
4.3	An instance showing the effectiveness of fusion.	81
4.4	Effectiveness of fusing visual and lexical clues.	83
4.5	Performance of acoustic and lexical features	87
4.6	Multi-actor Emotion Recognition results	95
4.7	Classification results after fusion	96
5.1	The Big-Five Dimensions and the associated traits	102
5.2	Notations used in defining the features.	104

5.3	Explanation of the features used.	105
5.4	Emotion Recognition performance.	109
5.5	Performance of predicting BFI answers.	111
5.6	BFI-10 scoring: Questions associated with each factor.	118
5.7	Comparison of the results of predicting the personality scores.	123
5.8	Accuracy in scoring the BFI: Comparing CRF and BFI.	128

List of Figures

1	Flow of the work covered in this thesis	vii
3.1	Sample images of the 3D models from the BU-3DFE database	37
3.2	3D face models provided in the BU-3DFE database	38
3.3	Distance features used in related works	39
3.4	Ambiguities due to measurement of absolute distances.	40
3.5	Comparing residues with absolute distance features for mouth width	42
3.6	Comparing residues with absolute distance features for mouth height	43
3.7	Shortcomings in taking difference of distances as features	44
3.8	Effect of training data size	53
3.9	Variation of Avg. Recognition Rate with number of features	57
3.10	Some misclassified images using the method in [172]	58
3.11	Movement of the lip corner in happiness.	59
3.12	Classification scheme for the proposed algorithm.	60
3.13	Histograms for $\theta_{p_0}^i(n)$ and $\varphi_{p_0}^i(n)$	63
3.14	ROC curves with the areas under them	65
3.15	Confusion matrices of the classification results.	66
4.1	Difficulties in the movie data	69
4.2	Framework for the proposed system.	70

4.3	Speaker detection using motion of FFPs on lips.	80
4.4	Examples of failures using visual clues.	84
4.5	Variation of visual weight with head pose.	85
4.6	An example of failure of the proposed algorithm.	86
4.7	Fusing emotional clues from multiple actors.	88
4.8	Dealing with pose variations using two pose parameters	89
4.9	Examples of failures in using the multi-actor approach.	96
5.1	Proposed personality assessment framework	99
5.2	Algorithm for personality assessment of movie characters	100
5.3	Effectiveness of using Sparse and Low-Rank regularization . . .	112
5.4	Variation of personality scores for Jack in <i>Titanic</i>	121
5.5	Comparison of Feature Selection Techniques	124
5.6	Predicted personality scores for a sample shot of Jack.	127

Chapter 1

Introduction

1.1 About Emotion Recognition (ER)

The face is the index of the mind. Combined with that, what we speak and how we speak or react to external stimuli are all helpful clues in making in-roads into understanding the complex world of human emotions. Humans display a wide range of emotions from subtle emotions like confusion to very intense emotions like jubilation.

Studying human emotions helps to understand human psychology and refine interpersonal interactions. Moreover, such studies will also benefit intelligent devices which are expected to assist us in our day to day activities. For example, when you come back *'tired'* from the office work, your music player will play your favorite musical program and if you are *'drowsy'* and want to take a nap, lights will dim automatically.

Such intelligent devices need to become socially intelligent as well, apart from their computational intelligence [189]. An important aspect of that social intelligence is the ability to recognize human emotions and respond accordingly. But emotion is much of an internal state and sometimes even human beings find it difficult to understand internal feelings of a person. This makes

human emotion recognition even much more difficult for computers, motivating the researchers to delve into the field of human Emotion Recognition using computers (ER).

1.2 How to Define Emotions?

For the task of ER, it needs to be clearly defined what are we trying to recognize. Are we trying to recognize just a movement of facial muscles such as raising of eyebrows or are we trying to recognize the anger behind such facial actions? Answer to this question is provided by the psychologists through *message judgment* and *sign judgment* approaches [41]. Using message judgment representation, the aim is to recognize the emotion underlying a particular facial expression. On the other hand sign judgment approach does not directly recognize emotion but it recognizes certain external facial movements. It tries to code all possible *perceptible* changes occurring on a face due to expressions without going into the mental state of the person. Further analysis is needed to recognize emotions.

One of the popular definitions of emotions under the message judgment approach is in terms of six basic universal emotions proposed by Ekman and Friesen [57], which are *Anger*, *Disgust*, *Fear*, *Happiness*, *Sadness* and *Surprise*. Most of the existing works on ER have recognized these six basic emotions. Labelling emotions is easier using this description of emotions since it has an intuitive understanding for humans. But a disadvantage of this representation is that many emotions that we encounter in our day to day life cannot be categorized into these six classes. These include fatigue, pain, thinking, embarrassment etc. These emotions can be referred to as *subtle emotions*.

Another way to represent emotions under message judgment approach is using the dimensional approach where emotions are characterized by two di-

mensions viz. evaluation and activation. Evaluation determines whether the emotion is positive or negative while activation determines the intensity of the emotion. A wide range of emotions can be represented using this system. However, in this system, expressions for fear and anger cannot be distinguished [218]. Also, trained labelers are required for labeling emotions using the dimensional approach.

Using sign judgment approach, emotions can be defined in terms of a set of facial actions such as raising eyebrows, pulling lips apart and so on. These facial actions (called as Action Units or AUs) are defined in the Facial Action Coding System (FACS) introduced by Ekman and Friesen [59]. An advantage of AUs is that they can be combined to define a wide range of facial expressions corresponding to emotions beyond the six basic emotions. However due to interpersonal variations in display of emotions, it is hard to associate a specific set of AUs with a particular emotion. Consequently most of the works just recognize AUs without recognizing the underlying emotions.

1.3 Applications of Emotion Recognition

Emotions can be recognized through different modalities such as facial expressions, body gestures, voice intonation, spoken words, physiological signals such as rate of heart beat, blood pressure and so on. A particular modality is used based on the feasibility of data capture. Some of the applications of ER using either one or more of these modalities, are explained below.

1.3.1 Biometrics

‘Biometrics’ refers to methods for uniquely recognizing humans based upon one or more intrinsic physiological or behavioral trait. Physiological traits are related to the shape of the body. For example, fingerprint, face, DNA, hand

and palm geometry, iris, and odor/scent. Behavioral traits are related to the behavior of a person. For example, typing rhythm, gait, signature and voice.

While the robustness of behavioral biometrics has often been questioned with regard to their sensitivity to human emotional conditions, physiological biometrics also exhibit limitations, being either too expensive and intrusive (iris, DNA), or not accurate enough for high security applications [23]. For example, face recognition is always sensitive to facial expressions. This has led the researchers to propose facial expressions as an alternative biometric [176].

1.3.2 Medicine

Emotions have been widely used in clinical research to study the affective and cognitive states, and psychopathology of an individual. ER has played a major role in the study of schizophrenia, which is a neuropsychiatric disorder characterized by deficits in emotional expressiveness. Patients with schizophrenia are known to have impaired performance in emotion processing, both in terms of recognizing and expressing emotions. Patients with schizophrenia often demonstrate either or both types of impairment in expressing emotions: “flat affect” (a severe reduction in emotional expressiveness) and “inappropriate affect” (inappropriate expression to intended emotions). Techniques of ER have been used to quantify the facial expression abnormality of patients with schizophrenia [195].

Research has also been performed to recognize emotions of patients during epileptic seizures [114] which helps to understand the cerebral organization of seizures.

1.3.3 User Response

ER is used for understanding how well a student is receiving a lecture delivered via an automated tutoring system. Through these systems the students can provide feedback about the lecture, even subconsciously through their facial expressions and physiological signals. For example, an emotion mouse [12] developed by IBM could recognize emotions of the user through physiological signals captured when he/she touches the computer mouse. Based on such feedback, an interactive tutoring system could dynamically adjust the speed of instruction, allowing the speed to increase when the student's understanding is solid and to slow down during an unfamiliar topic [199].

In the field of advertising, ER is used for understanding the emotional responses of consumers towards television advertisements [46]. Recent applications of ER has come up in the form of development of technologies such as targeted advertisement where the advertisement on the billboards adapt to the emotions of the viewer.

Smart call centers can detect dissatisfaction in customers by recognizing frustration through their voice and the spoken words. Such customers can be provided assistance or reconciliation [49].

1.3.4 Robotics

The term "Social Robotics" refers to an area of robotics where robots are designed such that they are able to recognize humans and engage in social interactions with them. The capability of ER from facial expression, speech and hand gestures, helps social robots to understand the intent of humans and respond accordingly. For humans, the face to face communication is a real-time process operating in times of the order of 40 milliseconds. At this time scale the level of uncertainty is considerable if a robot recognizes emotions . So,

automated real time facial expression techniques are needed [16].

1.3.5 Surveillance

Generally face recognition algorithms will be effective in surveillance when the subject doesn't exhibit any emotion. This is not easy to achieve especially in camera surveillance systems where people are unaware of being tracked. Some systems can model human emotions based on ER techniques and remove the facial expressions [131]. Another security application is to hide the identity of an internet user during a chat without hiding his expressions. In such a case, the person's expressions are communicated using a virtual mask whose shape can be like that of a tiger or spider-man [95].

1.4 Proposed Application of ER on Difficult Data

Emotions of an individual have been found to be correlated to his/her personality [77]. For example, a positive correlation exists between exhibiting fear and tendency to become easily nervous (a characteristic of neuroticism) [113]. By observing an individual's emotions over a period of time, clues can be obtained about his/her personality. For example, one of the personality traits is being reserved. Psychological studies indicate that reserved persons are less happy as compared to those not so reserved [137]. Consequently, the fraction of the time a person is happy indicates to what degree he/she is reserved. Apart from being reserved, there are other personality traits addressed through specially designed personality questionnaires. In the proposed application, results of emotion recognition are used to devise features which can automatically fill up personality questionnaire for that individual.

Personality scores are obtained from the questionnaire using another novel scoring methodology. The proposed scoring methodology can be used for scor-

ing a wide range of personality questionnaires as opposed to the usual scenario when each questionnaire has to have a separate scoring scheme. Details about personality assessment are presented in Chapter 5.

1.5 Contributions of the Thesis

The contributions of the thesis are listed below:

1. Novel algorithm has been developed for 3D Facial Expression Recognition which is around 9% more accurate than the state of the art approach in dealing with *low intensity (subtle) expressions* in addition to the higher intensity expressions.
2. A novel methodology of fusion has been developed which, for the first time, *combines visual clues with lexical clues* for emotion recognition in movies. Fusion improves the accuracy by 6% as compared to using the visual or lexical cues individually.
3. A near real-life *data set for emotion recognition* (based on movie characters) is generated.
4. Emotion Recognition has been applied for the first time to *automate the answering of personality questionnaires*. It is observed that predicting personality scores through answering personality questionnaire performs much better than directly predicting personality scores from the extracted features.
5. We proposed a methodology for scoring the personality questionnaires which is not specific to a questionnaire but *can be applied to a wide range of personality questionnaires*.

1.6 Publications

The work covered in this thesis has been published in the following papers:

1. Ruchir Srivastava, Sujoy Roy, Utilizing 3D Flow of points for Facial Expression Recognition, journal manuscript under review, *Multimedia Tools and Applications*.
2. Ruchir Srivastava et al, Don't Ask Me What I'm Like, Just Watch and Listen, accepted in *ACM Multimedia 2012 (Full paper)*.
3. Ruchir Srivastava, Shuicheng Yan, Terence Sim and Sujoy Roy, Recognizing Emotions Of Characters In Movies, in proc. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.
4. Ruchir Srivastava, Shuicheng Yan, Terence Sim & Surendra Ranganath, Subject Independent Facial Expression Recognition from 3D Face Models using Deformation Modeling. In A. Malik, T. Choi, & H. Nisar (Eds.), *Depth Map and 3D Imaging Applications: Algorithms and Technologies*, IGI Global, 2012, pp. 574-595. doi:10.4018/978-1-61350-326-3.ch030
5. Ruchir Srivastava, Sujoy Roy, Shuicheng Yan, and Terence Sim, Multi-actor Emotion Recognition in Movies Using a Bimodal Approach, in proc. *International Conference on Multimedia Modeling (MMM)*, 2011.
6. Ruchir Srivastava, Terence Sim, Shuicheng Yan, Surendra Ranganath, Feature selection for facial expression recognition using deformation modeling, in proc. *International Conference on Digital Image Processing (ICDIP)* 2010.
7. Ruchir Srivastava, Sujoy Roy, 3D Facial Expression Recognition Using Residues, in proc. *IEEE TENCON*, 2009.

Chapter 2

Literature Review

The two main components of this thesis are *Emotion Recognition* and its application for the task of *automatic answering of personality questionnaires*. This chapter presents a brief review of the related works in these two areas. Moreover, since we are dealing with difficulties (section 2.2) with the data, a review of the available datasets for ER is also presented.

2.1 Databases for Emotion Recognition

Construction of databases for ER has been guided by the trend of research in this area. Consequently, most of the databases for ER have been captured under controlled conditions i.e. sufficient lighting, frontal face pose and significant facial resolution. Also the expressions are exaggerated and limited to the six basic expressions. For the audio databases, background noise has been minimized by recording in a quiet environment with recording devices close to the speakers. Audio databases are also faced with the problem of posed emotional utterances. Consequently, there is lack of databases capturing difficult data.

The databases for ER can be categorized into visual, audio and audiovisual as follows. Please note that this section may not contain all the relevant

emotion databases. For a wider coverage please refer to [54], [61] and [218].

2.1.1 Visual Databases

One of the most widely used databases for facial expression recognition is the Cohn-Kanade (CK) database [171]. The CK database consists of approximately 500 image sequences from 97 subjects. The subjects were asked to perform displays consisting of single or a combination of AUs beginning from a neutral face. The last frame in each sequence is labeled with one or more Action Units (AUs, Section 1.2). These combinations can also be converted to the corresponding facial expressions using the guidelines in [58]. Images in the sequences are grayscale images of size 640×480 pixels with the faces almost frontal and of good resolution.

There may be some challenge in the data in the form of insufficient illumination and presence of facial hair. Still since the displays are exaggerated and acted out and the environment is mostly controlled, the CK database may not be suitable for testing algorithms developed for difficult data.

Apart from the databases of images or image sequences, with the evolution of Facial Expression Recognition using 3D data, Yin et al. [211] at Binghamton University constructed a 3D facial expression database (BU-3DFE database) for facial behavior research. The BU-3DFE database contains triangulated 3D mesh models and 2D facial textures (512×512 pixels) for 100 subjects. Each subject has 3D models for a neutral and 4 intensities of 6 expressions, making a total of 2500 3D models. Intensity of an expression refers to different stages of development of an expression. A low intensity level is closer to a neutral face and intensity increases as the expression progresses in time towards the peak. The spatial coordinates (x, y, z) and color (R, G, B) values are provided for all vertices in each facial model. Apart from this, the database also provides

the spatial positions and color for 83 corresponding facial points on each facial model. These 83 points correspond to prominent facial features such as corners and contours. Few sample images for 3D models in the database are given in Figure 3.1 (section 3.1). Figure 3.2 (section 3.1) shows the 83 points marked on the face.

An advantage of 3D data is having the extra depth information. BU-3DFE contains static face models only and it has been extended to the BU-4DFE database [210] which contains dynamic 3D data i.e. 3D videos of facial expressions. However, these expressions are still acted out in an exaggerated manner and the videos are shot under sufficient illumination and with good facial resolution. Due to these constraints, the data in BU-3DFE and BU-4DFE fall in the category of easy data.

There have been only a few databases of natural facial expressions. The MMI facial database [135] can be considered as the most comprehensive database of facial behavior recordings containing both posed and natural facial expressions [218]. Recordings are in the form of both static images and videos and most of them have both profile and frontal views. The database is searchable and downloadable via the internet. The data is labeled with 6 basic emotions and/or AUs. The MMI database also provides challenges in the form of naturalness of expressions, facial hair, glasses and profile view of the face. It can be helpful for testing FER algorithms addressing these challenges. However, the faces are of good resolution and with sufficient illumination.

2.1.2 Audio Databases

For ER from speech, the most widely used database is the Banse-Scherer vocal affect database [15]. The database consists of audio recordings of 12 professional stage actors depicting 15 emotion categories; hot anger, cold anger,

panic, fear, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust and contempt. In order to elicit the desired emotions from the actors, they were presented a written text describing a scenario along with the intended emotion. They were asked to imagine the scenario and start acting when they felt the intended emotion. The acting was repeated unless the actors felt satisfied. Audio was recorded using high quality microphones. The expressions were acted out so may not represent natural emotions. Moreover, recordings are in a controlled environment.

The Danish Emotional Speech Database (DES) [64] consists of speech of 2 male and 2 female actors. The actors read texts in five emotional states: angry, happy, sad, surprised, and neutral. The text itself is not emotionally colored but only the way of reading it is emotional. This is also an acted database.

Apart from the above mentioned databases, few databases cover fully natural *emotion-related states* which are different from an emotion in a strict sense. Some of these emotion-related states are depression and suicidal states (The Vanderbilt II database [70]), stress (The SUSAS database [82]). Even mother-child interactions [160] can be related to emotions [182]. The ISL meeting corpus [29] contains recordings of 18 meetings with 5 persons per meeting on an average. In a meeting scenario, we may not expect a wide range of emotions to be present. Consequently, the data is labeled with positive, negative and neutral emotions. The CSC corpus [83] contains audio recordings of 32 subjects who were asked to tell the truth and also to deceive the interviewers in two different tasks. The corpus is labeled with deceptive and non-deceptive speech. The Automatic Call Center (ACC) database [103] contains human-computer dialogs at a commercial call system with emotions labeled as negative and non-negative. Similarly Bank and Stock Service database [48] contains human to human dialog recorded at a call center. The emotion categories considered are fear, anger and stress. The AIBO database [20] contains audio recordings of children in-

teracting with robots. This database has a wider range of emotions, namely joyful, emphatic, surprised, ironic, helpless, touchy, angry, bored, motherese, reprimanding and rest.

2.1.3 Audio-Visual Databases

Among the audio-visual emotion databases, the Chen-Huang audio-visual database [38] contains posed facial and vocal displays of the six basic emotions along with other states namely interest, puzzlement, frustration and boredom. The data was recorded for 50 males and 50 females. The subjects were asked to speak sentences expressing a specified emotion. The subjects were not instructed on how to enact the emotions. Still, the emotions are acted out and are not natural.

The Belfast Database [54] is an attempt to create a natural audio-visual emotion database. It consists of video clips taken from television and realistic interviews with the research team. Instead the labeling of the clips has been done based on the dimensional labeling (Section 1.2). Other databases which are labeled using dimensional labeling are the SAL database [55] and its extension in the form of the SEMAINE database (SEMAINE-DB) [5]. In both of these databases spontaneous (induced) audiovisual data was recorded from conversations between subjects and artificial listeners with different personalities [78]. However, in the case of dimensional labeling, it is possible that not all of the basic emotions are covered [218].

The Adult Attachment Interview (AAI) database [146] contains recordings of interviews where 60 adults were asked to describe their childhood experiences. Data is labeled using Facial Action Coding System (FACS). Apart from the 6 basic emotions, the emotions contained are embarrassment, contempt, shame, general positive and negative. Researchers at Rutgers University col-

lected the RU-FACS database [17] in which subjects were asked to convince the interviewers about their being truthful. The database is labeled with FACS and contains 33 AUs. The AAI and RU-FACS databases are not publicly available. Another database of natural emotions is the Geneva groups recording of travelers [152] who had lost their luggage. However, it is very specific to the situation and does not contain all the basic emotions.

From the review of databases for emotion recognition, it can be concluded that the databases are limited either in terms of the emotions covered or they are too specific to a situation. Moreover, they are usually acted out and are different from day to day emotions. There is a need for databases of a wider range of naturally expressed emotions especially one covering at least all the six basic emotions expressed naturally in a close to real life environment.

2.2 Emotion Recognition

Most of the researchers have been trying to recognize the six basic emotions i.e. Anger, Disgust, Fear, Happiness, Sadness and Surprise since they are found to be sufficiently distinct from each other. This is true when these emotions are acted out in an exaggerated manner (*Posed Emotions*). However, the same six emotions can lose much of their distinctness if they are expressed naturally (*Spontaneous Emotions*). Attempt to recognize spontaneous emotions is expected to be more helpful in order to realize real-life implementation of ER systems [218].

Apart from the way the expressions are acted out, an important factor determining the usefulness of the research on ER is the environment in which the test data is captured. Most of the databases used for ER have been captured in highly controlled environment with sufficient lighting, nearly frontal face, almost no occlusion and so on along with exaggerated expressions (such

data is referred henceforth as ‘*easy data*’). However, real life data is not bound by these constraints and is referred to as ‘*difficult data*’. As explained later in Section 4.1, real-life challenges make ER a difficult task. This section covers related work on ER using both kinds of data. First those using easy data have been covered in sections 2.2.1 to 2.2.3 and then works on other data have been covered in section 2.2.7.

Another aspect about the data used is the modality (source of information) which can be facial expression, voice, body gestures etc. Human ER using computers has mostly been performed by recognizing facial expressions. However, other clues such as voice intonation, bodily gestures, spoken words etc. have also been found to be helpful in this task [218]. Consequently, the trend has been shifting to multimodal emotion recognition which fuses these different clues (modalities). Since we have used Acoustic, Visual and Lexical (multimodal) clues for ER, in this section, approaches using these three clues either alone or in combination are discussed. Works using modalities other than these three (Such as [79], [94] and [86]) are not discussed here. Please note that the works presented in the next section deal with easy data and are categorized under *Traditional Emotion Recognition*. Works going beyond traditional emotion recognition have been dealt with separately in section 2.2.7.

2.2.1 Using Visual Clues

Human ER from visual clues has been mostly performed by analyzing facial expressions beginning with the work of Suwa et al. [170]. Most of the researchers have been working on easy data. Two approaches have been used for this task; geometric and appearance based depending of the type of features extracted.

In the geometric approach, facial feature points are identified on prominent

facial landmarks such as the pupils, inner/outer corners of the eyes, nose and mouth. These points are tracked and their motion serves as a feature for facial expression recognition. Among these approaches, [207] used optical flow in the regions of mouth, eyebrows and eyes for modeling non-rigid facial motions in image sequences. A planar face model was used for modeling rigid facial movements. Extracted motion parameters were used in a rule based framework to predict 6 basic facial expressions and the neutral face. [98] fitted a Candide grid on video sequences for tracking facial feature points. [191] also used a similar approach. [157], [132] and [134] used motion of facial feature points around eyebrows, eyes, nose, lips and chin for facial expression recognition with [134] using both frontal and profile views of the face. [209] proposed a spatio-temporal approach for recognizing six basic expressions from video and also for computing levels of interest i.e. the intensity of the expression. Optical flow vectors projected onto a lower dimension using Principal Component Analysis (PCA) were used as basic features.

In the appearance-based approach, facial region as a whole is used for extracting motion, change in texture etc. [51] compared different appearance based approaches to recognize 12 AUs and best accuracy of 96% was achieved using both Gabor filters and Independent Component Analysis, individually. One of the major works using appearance based approach was in [18] who applied Gabor filters on the frames of input video and used the output magnitudes for recognizing 17 AUs. Different recognition engines such as Adaboost, Support Vector Machines (SVMs) and Linear Discriminant Analysis (LDA) were compared. Feature selection was explored using Adaboost so as to reduce the dimensionality of the feature vectors before feeding to SVM or LDA classifiers. Best results were obtained using Adaboost followed by SVM. Apart from Gabor filters, Haar filters were used in [200] in conjunction with Adaboost for AU recognition and showed better performance as compared to Gabor filter

based approach. [187] used the concept of multilevel motion history images (MMHI). MMHI representation is an extension of temporal templates which are 2D images showing motion history i.e. where and when motion occurred in an image sequence. 21 AUs were recognized comparing two classification schemes: (i) a two-stage classifier combining a k-Nearest Neighbor (k-NN) based and a rule-based classifier, and (ii) a SNoW classifier.

Considering individual limitations of appearance based and geometry based approaches, researchers have also used a combination of appearance and geometric features. [177] recognized 16 AUs and the neutral face by analyzing both permanent facial features (related to eyes, mouth, nose etc.) and transient facial features such as wrinkles, furrows etc in videos. They proposed multi-state face and facial component models for this task. Wang et al. [195] performed FER experiments to validate their feature representation which consisted of 2D and 3D geometric features, and the moment invariants combining both 3D geometry and 2D textures. Using their features they recognized four expressions; Anger, Fear, Happiness and Sadness; with a better accuracy than using either 2D geometric or texture features alone. [178] modeled the interactions between rigid and non-rigid facial motions using Dynamic Bayesian Networks (DBN). Features extracted from Active Shape Models (ASMs) and Gabor filters are combined to recognize 14 AUs.

2.2.2 Using Clues from Speech

There can be two clues related to speech; from the acoustic property of speech and from the content of the speech. ER from speech has been mostly performed by analyzing the acoustic properties of speech [14] [220] using features such as pitch-related measures (range, mean, median, and variability), intensity, and duration. Spectral features such as Mel Frequency Cepstral Co-

efficients (MFCCs) and cepstral features have also been successfully used in many recent studies. It has been reported that MFCC models the human perception to speech quite well [143] and is therefore used for many speech applications. The most important acoustic features which capture emotional content of speech are pitch and energy [31]. While other significant features are Zero-Crossing-Rate (ZCR), Teagor Energy Operator (TEO) based parameters; and frequency-centroid, formant and pitch vibration [203].

Scherer et al. [151] performed a study in which 29 acoustic features were extracted from voice recordings of 14 professional actors. It was found that Sadness and Anger are the best recognized followed by Fear and Joy. While Disgust is the worst recognized emotion if purely vocal signals are used. [15] used fundamental frequency (F0), the energy, the articulation rate, and the spectral information in voiced and unvoiced segments of speech to recognize 14 emotions. Petrushin et al. [139] compared the abilities of humans and computers in recognizing emotions from speech. The emotions included in their work were Happiness, Anger, Sadness, Fear, and Normal (no emotion) and these emotions were acted out by 30 non-professional actors. They obtained similar recognition rates of around 65% for both humans and computers. The features used were selected from fundamental frequency F0, energy, speaking rate, first three formants (F1, F2, and F3) and their bandwidths (BW1, BW2, and BW3). Chen et al. [38] used pitch, intensity, and pitch contours as acoustic features to recognize Happiness, Sadness, Fear, Anger, Surprise, and Dislike. Test data was for 2 speakers who were asked to speak 6 different sentences for each emotion. The sentences were in a language foreign to the speakers in order to avoid their getting affected by the linguistic content.

Nwe et al. [126] compared Log-frequency power coefficients (LFPC), and two cepstral-based features, namely Linear Predictive Cepstrum Coefficient (LPCC) and MFCC. Austermann et al. [14] extracted the values of the fun-

damental frequency, energy, jitter and shimmer as well as the power spectrum and the speech/pause time as acoustic features from each frame of the audio signal. These features were used to recognize Anger, Fear, Happiness, Sadness and Neutral. In [62], 12 MFCCs, 12 delta coefficients, 0th cepstral coefficient, and the speech energy were used as features to recognize six emotions grouped into three sets: high-arousal emotions (Anger, Fear, and Happiness), low-arousal emotions (Boredom and Sadness) and the Neutral emotion.

Apart from the acoustic features, content of the speech also indicates emotions. Lexical features have been used for the task of ER [103] as well as for analysis of movie reviews [185]. Lee and Narayanan [103] used the concept of emotional intelligence for emotion extraction from dialogues. Emotionally salient words were selected from a test dialogue based on a salient word dictionary. Lexical features were extracted using the selected words and the dialogue was classified into having positive or negative emotion. Such analysis of each word separately may not give sufficient clue about the context in which the word was spoken. For example, consider the phrase ‘Great blunder’ which is negatively oriented but individually both words have opposing indications about the emotion in the phrase.

Turney [185] used the concept of Semantic Orientation (SO) for classifying movie reviews into positive or negative. Positive value of SO indicates a positive emotion while negative value of SO indicates a negative emotion. There are other works which utilize lexical clues in combination with other clues for ER. Some of them have been discussed in the next section.

2.2.3 Using Multimodal Clues

There are works on ER combining multiple clues. [39] present one of the earliest multimodal approaches for recognizing Anger, Dislike, Fear, Happiness,

Sadness and Surprise. The acoustic features included pitch statistics, rms energy of the speech waveform and maximum and minimum of the derivative of pitch contour. From the video, optical flow vectors were computed on which Fourier Transform was applied and the resultant coefficients were used as visual features. In addition, some manually recorded facial movements are used as visual features. [217] fused visual features in the form of 12 action units with prosodic features i.e. logarithm of energy, syllable rate, and two pitch candidates and corresponding scores. They recognize posed six basic emotions from videos. [30] used motion of 102 facial markers and global-level prosodic features such as the statistics of the pitch and the intensity. Four emotions namely Anger, Happiness, Neutrality and Sadness were recognized from posed video sequences.

Wu and Liang [204] combined acoustic and prosodic features with semantic labels derived from a Chinese knowledge base to recognize four emotions; Neutrality, Happiness, Anger and Sadness. Song et al. [162] utilized Facial Animation Parameters obtained after Active Appearance Model based tracking of facial points to extract visual features from videos. Pitch and energy features extracted from audio were combined with visual features to recognize 6 basic emotions and neutrality. Bimodal fusion was found to perform better than the individual modalities. Hoch et al. [84] extracted prosodic parameters namely pitch, power, formants and duration of voiced segments and fused them with Haar based features to recognize positive, negative and neutral emotions. The video sequences were acted out in an automotive environment. Wang and Guan [196] recognized the 6 basic emotions using Gabor wavelet based visual features and prosodic, MFCC and formant frequency features. It was found that Disgust, Sadness and Fear were the most confused emotions and especially the Gabor based features could not distinguish between these emotions.

Lexical clues have been used very rarely with the visual clues. [154] de-

tected human interest level combining different clues including the lexical features from speech and Active Appearance Model (AAM) based features for recognizing facial expressions. Apart from these features, eye-activity and acoustic features was also explored. For linguistic analysis, a vocabulary of terms of interest was established which included non-linguistic vocalizations such as coughing, yawning, laughter, consent, hesitation and words such as *if*, *oh*, *yeah* and so on. For a test document, the frequency of occurrence of each of the vocabulary terms was used as the feature.

In the context of interest level detection, the choice of these terms is justifiable. But in recognizing emotions from movie dialogues, information about what words are spoken is essential. E.g. for two phrases, “*Oh! Nice painting*” and “*Oh! It is very troublesome.*”, occurrence of the word ‘*Oh!*’ indicates interest of the speaker but emotion is conveyed by the words ‘*Nice*’ and ‘*troublesome*’. This shows that spoken words are more important than just non-linguistic vocalizations for ER in movies.

2.2.4 Detailed Categorization of Works on ER

The works on ER can be categorized into different classes based on the following criteria:

1. Type of data used
2. Way to define expressions

This categorization does not involve the methodology used but it deals with the problem domain of the works. Categorization based on the above criterion is presented below:

2.2.5 Type of the Data Used

ER algorithms are meant for a wide range of applications depending on which, there can be different types of data which are to be analyzed. The different types of data can be:

- **D1: Single image in a lab environment**

Many works deal with recognizing facial expressions in a static image of a person. The image is captured in a controlled indoor environment. In [134], static images depicting both frontal and profile views of faces were used. Facial Feature Points (FFPs) were extracted separately in both the views. Displacement of these FFPs from neutral face to expressive face was used to recognize 22 AUs from frontal images and 24 AUs from profile images. When information from both views was combined, 32 different AUs were recognized.

- **D2: Single image in a real world environment**

It is equally important to build algorithms which would meet the challenges of FER in real world settings. This may involve a complex background, highly uncontrolled motion of the person and other such situations. An attempt was made in this direction in [85] using personalized gallery of images. Test image was matched to the images in a gallery using the Elastic Graph Matching technique for Face Recognition. The personalized gallery of this person was used for recognizing the expression on the test face. The FER system was tested on varying illumination and background structure for achieving robustness in real life environment.

- **D3: Video in a lab environment**

Many works have used videos captured in a controlled environment for ER. An example of this is in [74]. It was a good beginning to FER

and identified major facial actions of opening and closing the mouth and eyebrow raising apart from neutral and smile expressions. The face was modeled using 19 landmark points on the face. It was assumed that the deformation of the face from neutral can be expressed as a linear combination of a small number of known basis vectors. The basis vectors were computed using Singular Value Decomposition(SVD) of the 3D shape trajectory matrix. The coefficients of the linear combination were used as features for classification.

- **D4: Real world videos**

These are challenging due to pose, illumination variations. Other challenges are complex background, presence of speech etc. An attempt was made by Black and Yacoob back in 1996 [207]. In their work, optical flow was used in the regions of mouth, eyebrows and eyes for modeling non-rigid facial motions while a planar face model was used for modeling rigid facial movements. A rule based system used the extracted motion parameters to predict facial expressions corresponding to the 6 basic emotions. Performance of their algorithm was reported on video clips depicting real world situations.

- **D5: Multiple images**

Multiple images means that the images are taken over a significant period of time say with a gap of a few months. No work could be found in this direction.

- **D6: Group images or videos**

The images or videos contain groups of persons rather than one person only. Any attempt on this is yet to be done, to the best of our knowledge.

- **D7: 3D static models**

This type of data contains the 3D models of faces which are used for FER purpose. Only a single face model is used which bears a facial expression. In some cases a 3D model of the neutral expression is also required. Example of an approach using 3D static models is that used in [193] where primitive geometric surface features were used for FER. There are twelve primitive surface types e.g. ridge, valley, peak, saddle, concave hill etc. Each of the vertices was labeled with one of these surfaces. The face was divided into seven regions. Each region had vertices labeled with one of the 12 labels corresponding to each surface type. The fraction of each of the labels in each region, served as features for classification. The assignment of a surface type to a vertex was performed based on the surface curvatures and their principal directions.

- **D8: 3D dynamic model**

Just as videos are sequences of images, recently, sequence of 3D models are also being used as data for FER. In [167], a spatio-temporal approach was proposed using 3D dynamic geometric facial model sequences, to tackle FER problems. The approach integrated a 3D facial surface descriptor and Hidden Markov Models (HMM) to recognize facial expressions. The facial surface descriptors were the primitive geometric features as used in [193].

- **D9: Both images and 3D models**

Information from both images and 3D models was combined in [195] to get relevant features for FER. 3D geometric features were extracted in a fashion similar to that in [193]. Apart from 3D geometric features, 2D geometric features were also extracted. Face was divided into 28 regions and was marked with 58 landmark points. The area of these regions and distances between some of the fiducial points served as 2D geometric

features. The moment invariants served as textural features. A combination of 2D geometric features, 3D geometric features and textural features was used for recognizing the query expression.

- **D10: Audio in a lab environment**

This kind of data refers to audio only which is recorded in a lab environment. Such an environment involves minimal noise with good quality microphones located close to the speaker. This ensures better audio quality and such a data is easy to process. An example of usage of such data is in [47]. The data was collected by showing a sentence to the subjects and asking them to speak it with the labeled emotion. Only the pitch information was used to recognize Anger, Fear, Happiness and Sadness.

- **D11: Audio in real-life environment**

Such data involves audio collected in an uncontrolled or less constrained environment. An example is in [50] which performed real-life emotion detection in real agent-client spoken dialogs from a medical emergency call center.

- **D12: Text only**

This data refers to the spoken content of speech i.e. ‘*what*’ is spoken, rather than ‘*how*’ is it spoken. [203] recognized happiness, unhappiness and neutrality from text using Emotion Generation Rules from psychology.

- **D13: Multiple modalities in lab-environment**

Instead of a single modality such as facial expression or audio or spoken text, a combination of these modalities leads to a multimodal data input. [192] combined information from facial expressions, bodily gestures and speech to recognize positive high, positive low, negative high

and negative low emotions. The multimodal data is captured in a controlled environment.

- **D14: Multiple modalities in real or near real life environment**

As opposed to *D13* above, the multiple modalities can be obtained from a real life or close to real life environment. Currently, there is no existing work to the best of our knowledge which uses data falling in this category.

2.2.6 Way to Define Expressions

As mentioned above, facial expressions can be defined in terms of the six basic emotions, using FACS or using the dimensional representation. Based on this, the works on ER can be categorized as dealing with either of the following:

- **E1:Posed basic expressions**

In many of the works, a database of posed six basic expressions is used. In forming such a database, subjects were asked to pose the expressions. Most of the works described above fall in this category.

- **E2:Spontaneous basic expressions**

Spontaneously exhibited six basic expressions are captured by putting the subjects in a situation which would elicit expressions on their faces. In [209], six basic expressions were elicited by showing relevant video clips to the subjects. Their algorithm also assigned a level of interest to the expressions which is equivalent to determination of the intensity of expression.

- **E3:Subtle expressions**

These expressions are more close to the day to day life and are not just

limited to the six basic expressions. Few of the works recognizing subtle expressions are discussed later in section 2.2.7.

- **E4:Posed Action Units**

In this class of works, Action Units are recognized from posed visual displays. The work in [134] mentioned in this section above, recognized AUs using rule based methods. Under a set of rules, each AU was associated with a specific motion of one or more Facial Feature Points.

- **E5:Spontaneous Action Units**

To get spontaneous visual displays for AU identification, the work in [17] used the ‘false opinion’ paradigm in which the subjects were asked to speak the truth about a concept or to tell a lie. Moreover, they needed to convince interviewers that they were speaking the truth. In their work, video frames were passed through Gabor filters and filter coefficients were used for classification.

- **E6:Dimensional Representation**

In this representation, emotions are described in terms of dimensions rather than discrete emotion categories. Most popular representation in this category uses 2 dimensions namely valence and arousal [81] to represent emotions. Valence refers to the level of pleasure while arousal refers to the intensity of the affective state.

Table 2.1 shows the classification of some of the prominent works on ER, based on the above scheme. In the table, each column corresponds to a specific type of data used and each row corresponds to the way to define expressions. For example, the works referenced under column *D1* use single static image for FER and the works referenced in the row *E1* classify facial expressions posed by actors into the six basic expressions. From an analysis of the table,

we can find out the well researched areas and those areas where there is a lot of scope for research.

If the type of data is considered, it can be seen from Table 2.1 that most of the works on ER deal with videos shot in a lab environment (column *D3*). One of the reasons for using videos is the importance of temporal dynamics for the analysis of facial expressions. Many researchers have proposed that the dynamics of expressions is more important than just the static image. The NSF workshops also address this issue [67]. Experiments were also conducted in [213] to show the importance of motion in identifying subtle facial expressions. Through the experiments it was established that the inherent dynamic property in motion is beneficial for reducing the ambiguity in recognizing facial expressions. But the possibility of video sequences improving the perception of facial expressions by providing larger samples of the expressions has been ruled out. Timing of expressions has also been used in differentiating between posed and spontaneous expressions[186] [107].

Temporal information can be utilized in 3D as well [167]. But the problem in using 3D models is that there are huge computational and storage requirements that make it impractical for real time. But there is a good scope for exploring this area as well, and overcoming this limitation.

As far as expressions are concerned, most of the work is done to recognize posed six basic expressions or AUs (rows *E1* and *E4*). However it has been identified by researchers [150] [63] that the emotions exhibited by human beings in day to day life accompanies facial expressions which are beyond just the six basic expressions. On the other hand, FACS on their own do not give much information about the facial expressions of a person. The FACS need to be integrated in order to recognize emotions. The area which needs attention is the recognition of subtle expressions (row *E3*) and it is encouraging to see works emerging in this direction.

Other conclusions that can be made from Table 2.1 are as follows:

- There is a need to experiment on real world data (columns *D2*, *D4*, *D11* and *D14*) which includes many practical difficulties such as pose and illumination variation, occlusion etc.
- It would also be interesting to experiment on group data either video or images (column *D6*) for ER. This can be useful to find the overall response of an audience involving many people.
- Multiple images of a person have not been explored for FER (column *D5*). The multiple images in this context refers to the images of a person over a long time period. There is a possibility that the way a person exhibits facial expressions varies with time. So, it is worth studying the effect of aging on facial expression display.
- There are differences in opinion over the point whether 2D image based approaches are better or the 3D model based approaches. But a combination of 2D images and 3D models can be used for FER. Very few works have addressed this issue (column *D9*).
- Dimensional representation of emotions is still less popular due to its limitations as mentioned in section 1.2.

E	D	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14
E1		[44][222] [109][130] [89][212] [69][108] [133][224] [157][225] [119][88]	[85]	[207][25][171] [97][148][120] [96][67][66] [40][177][209] [191][194][129] [99]	[25]			[193][211] [195][80] [114][121] [173][122] [163][206] 3D Residues (Sec. 3.2) Deformation Modeling (Sec. 3.3)	[74][168] [72][37] [167][53] [52][23]	[195][168] [169]	[220][14][204] [38][126][151] [139]	[50]	[203]	[30][204] [162][84] [196][39] [217]	
E2					[209]										
E3		[136]		[13][8][90] [45][63][71] [107]			Multi- actor ER (Sec. 4.2)				[15][31][62] [161]	[175][102] [27][100]	[103] [185] [153]	[154][216] [215][155] [192]	Bimodal ER (Sec. 4.1)
E4		[200][134] [133][178] [208]		[187][43][186] [221][132]					[22][23]						
E5		[108]		[17][186][18] [201][42] [158]											
E6											[73]				

Table 2.1: Classification of the works on ER based on the type of data used and the way to define expressions. The works covered in this thesis are also shown. Refer to section 2.2.4 for the symbols used.

2.2.7 Directions Beyond Traditional Emotion Recognition

Traditional emotion recognition involves recognizing the 6 basic emotions posed by actors in a controlled environment. However real life data is far beyond these constraints. Considering this, the trend is to go beyond traditional emotion recognition. This section briefly mentions some of those directions.

Considering Facial Expression Recognition, most of the early works were for frontal faces but recently few works have dealt specifically with non-frontal facial expressions [88] [99] [119] [224] [225]. [88] proposed to use view based classifiers based on appearance based features such as Linear Binary Patterns (LBPs), Histogram of Oriented Gradients (HOG) and Scale Invariant Feature Transform (SIFT).

Apart from the basic emotions, researchers are attempting to recognize more subtle expressions such as pain [13] [107], depression [60], fatigue [90] and other cognitive states like agreeing, concentrated, disagreeing, interested, thinking, confused and frustration [215] [155] [63] [93] [209] [161]. [27] detected stress in speech signal using acoustic features. It was shown that features based on cepstral analysis such as LPCC, One-Sided Autocorrelation Linear Predictive Coding Cepstral Coefficients (OSALPCC) and MFCC clearly outperform the performance of the linear-based features of Linear Predictive Coding (LPC) and One-Sided Autocorrelation Linear Predictive Coding (OSALPC).

[215] went beyond the six basic emotions to recognize 4 other cognitive states namely interest, boredom, puzzlement and frustration using a bimodal approach. 12 facial features were tracked to obtain visual features while pitch and energy were used as audio features. However these cognitive states were acted out. Sebe et al. [155] used motion of facial features and prosody features to recognize 6 basic emotions along with confusion, interest, boredom and

frustration from video sequences.

There are works which deal with less constrained environments. [73] used a dimensional approach (Valence-Arousal representation) to recognize emotions from movies. They used 10 audio features based on Mel-frequency cepstral coefficients (MFCC), Fast Fourier Transforms (FFT), Zero Crossing Rate (ZCR), pitch information and variation of chroma elements. [100] used pitch, log energy, formant, mel-band energies, MFCCs as the base features, along with velocity/acceleration of pitch and MFCCs to form feature streams. Pitch and energy were experimentally found to be the most important features. Experiments were performed on AIBO database to recognize 5 emotions Anger, Boredom, Happiness, Neutrality and Sadness.

Schuller et al. [153] integrated Part-of-Speech and higher semantic tagging in addition to using the spoken words directly to recognize Motherese, Emphatic, Neutral and Angry classes. The data used was in the form of recordings of children interacting with pet robots. Automatic Speech Recognition (ASR) was used to extract the speech content from the recordings. Lee et al. [102] used low level descriptors namely zero crossing rate, root mean square energy, pitch, harmonics-to-noise ratio, and 12 mel-frequency cepstral coefficients and their deltas. 12 statistical functionals computed for every low level descriptor were used to recognize Anger, Emphasis, Positive emotions, Neutrality and Resting. The data used had emotions elicited through affective scenario prompts.

[216] recognized positive and negative emotions using the Adult Attachment Interview database. Facial textures were used along with total 20 prosodic features for the task. Simon et al. [158] recognized 10 AUs from spontaneous facial displays. SIFT features extracted at facial feature points were used in conjunction with k-segment SVM classifier. Yang et al. [208] proposed to use compositional features to recognize low intensity AUs. These compositional

features were formed by combining local appearance based features from different facial regions. Park and Kim [136] recognized subtle facial expressions by magnifying the motion vectors at prominent facial feature points.

From the distribution of the works on Emotion Recognition, the areas needing attention were identified (Section 2.2.4). Out of the different ways of going beyond traditional emotion recognition, *this thesis focuses on working with a difficult data* as defined in section 2.2.

2.3 Automated Personality Assessment

Most of the research on automatic assessment of personality has attempted to directly predict the personality scores using features extracted from audio and/or video. Personality scores have five dimensions in accordance with the Five Factor Model (FFM). These five dimensions (five factors or traits) are *Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism* and are also referred to as the Big Five personality traits. Details on the FFM are presented in section 5.1.

To the best of our knowledge, automatic assessment of personality was first attempted by Argamon et al. [11] using lexical clues from informally written text. They tried to distinguish high level from low level of Neuroticism and Extraversion in authors of such texts. It was identified that lexical features were effective in predicting Neuroticism. However, improvement was needed in predicting Extraversion. Oberlander and Nowson [127] tried to improve the approach of Argamon et al. by extending to other personality dimensions. They identified four of the five personality traits (excluding Openness) from a corpus of web blogs using n-gram textual features. Their results were comparable across the four personality traits. Zen et al. [214] estimated levels of extraversion and neuroticism of a person using clues from proxemics and visual

attention. Proxemics refers to the interpersonal distances which an individual prefers. Other features used are extracted from the people tracking results and head orientation estimation.

Different from the works which identify only few of the five personality traits mentioned in the FFM, Mairesse et al. [111] [112] introduced acoustic features and used them along with lexical features to assess all the Big-Five personality traits. The acoustic features used were pitch statistics, intensity statistics, voiced time and speech rate (words per second). It was observed that Extraversion was the easiest to predict using spoken language. Mohammadi et al. [118] identified all of the Big-Five personality traits using prosodic features. The performance was above chance except for Openness with a satisfactory recognition rate observed only for Extraversion (76.3%) and Conscientiousness (72.0%).

The accuracy of automatic personality assessment algorithms is still not satisfactory for all the Big-Five traits.

Chapter 3

Utilizing 3D information for Facial Expression Recognition

Facial Expression Recognition (FER) can be performed using either 2D images or 3D face models. Most of the earlier works used images as the data for developing FER algorithms. This approach to FER using images is known as the image based approach. However, the image based approach has its limitations [211]. Especially, in exhibiting an expression, the facial muscles move and the change is reflected in the facial skin. The skin motion is in 3D which cannot be captured accurately with 2D images of the face.

A step towards using full 3D information is using range images which are grayscale images in which the intensity values at a pixel is proportional to the depth of the object at that pixel. Yabui et al. [206] used range images for FER. Analogous to the ‘Eigenfaces’ used in [184], the basis eigenvectors obtained from range images were called ‘3d-classic eigenfaces’. This approach has the advantage of being robust to illumination changes since the texture information is not taken into consideration. However, 3D information about the face can be more accurately obtained with higher resolution using 3D models as compared to range images. The approaches using 3D models of the face for FER can be

called as the 3D model based approaches.

3.1 Database Used

To facilitate research on 3D facial expression analysis, a 3D facial expression database was constructed [211] at Binghamton University which has proved to be very useful for facial behavior research. We have used the 3D models provided in this database (known as the BU-3DFE database). This implies an assumption that the 3D face models are available to us which is a valid assumption considering the fact that as 3D imaging is advancing, 3D face models can be acquired in real time [1]. Devices like Microsoft Kinect [2] can be used to acquire both color and depth information. Currently there may be problems with the resolution of the acquired depth and color images. However it is expected that resolution of Kinect sensor will improve in the near future. Resolution problems can also be addressed by developing FER algorithms to deal with low resolutions.

Another way of getting 3D face models is using face reconstruction techniques. Usually for multimedia applications, data available is in the form of 2D image sequences. 3D face models can be reconstructed from videos if the person moves his face to show different views of his face. This technology known as face reconstruction uses example-based methods, stereo methods, video-based methods and silhouette-based methods to obtain 3D locations of a few facial landmark points with reference to a fixed origin. 3D location of a sufficiently large number of facial points approximates the 3D model for the full face. Hence our premise of availability of 3D data is valid. For a survey on 3D face reconstruction please refer to [166].

The BU-3DFE database, used in this work, contains the triangulated 3D mesh models and 2D facial textures for 100 subjects. The shapes are provided

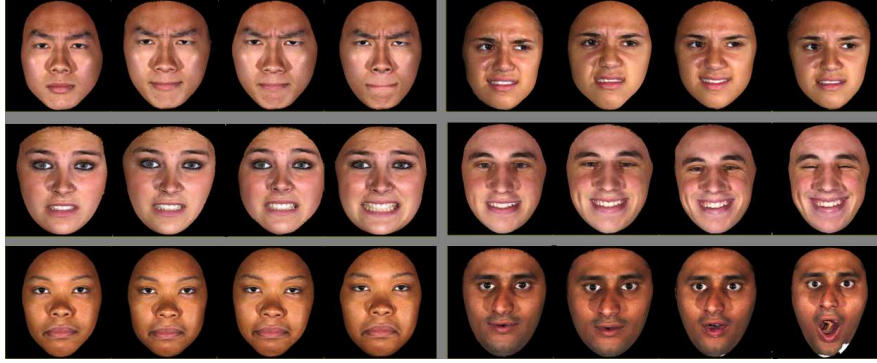


Figure 3.1: Sample images of the 3D models from the BU-3DFE database. Four gradations of expressions are shown. Top row: Anger and Disgust; middle row: Fear and Happiness; bottom row: Sadness and Surprise

in the form of triangulated mesh models (Fig.3.2b). A triangular mesh is a graph, $G = (V, T)$ with V denoting the set of vertices, and T denoting a set of triangles $T = (\bar{p}_i, \bar{p}_j, \bar{p}_k)$, where $\bar{p}_i, \bar{p}_j, \bar{p}_k \in V$ [166]. Along with the geometrical coordinates, color (R,G,B) values are provided for all the vertices in each facial model. Each subject has models for 4 gradations of 6 expressions and a neutral model. There are 2500 face models in total. To help the researchers in analyzing properties of Facial Feature Points (FFPs), the database also provides (x, y, z) coordinates and color (R,G,B) for 83 corresponding facial points on each facial model. Some of the sample images for the 3D models in the database are given in Figure 3.1 with the 83 FFPs shown in Figure 3.2a.

3.2 3D Residues for Subject Dependent FER

3.2.1 Problem Formulation

Few closely related works need to be discussed whose limitations lay down the need for the proposed work. Soyel and Demirel [163] compute five 3D facial distances. These distances are shown in green lines in Figure 3.3a. To deal with varying facial sizes, these facial distances are normalized using the

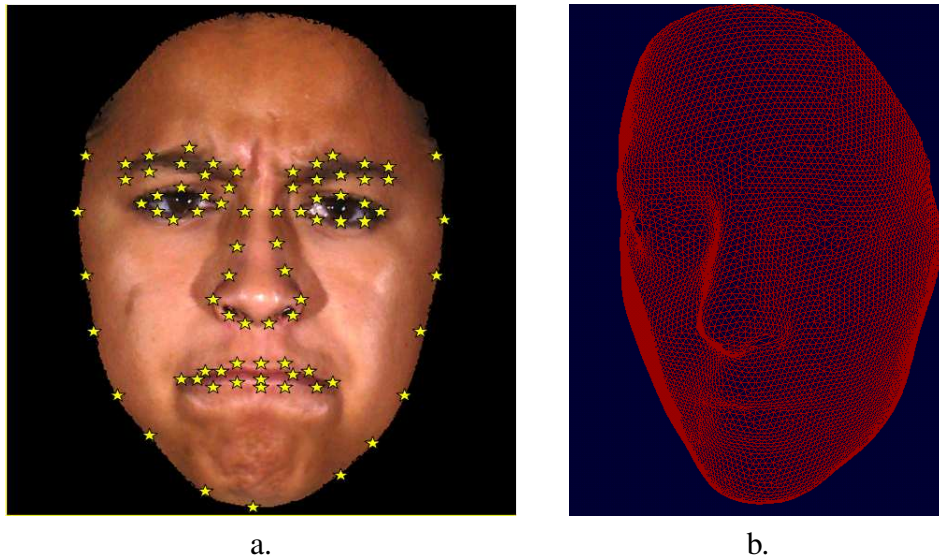


Figure 3.2: a) 83 Facial Feature Points marked on each 3D face model are available in the BU-3DFE database. b) 3D models are in the form of triangulated mesh models.

distance between the outermost point of right-face contour and the outermost point of left-face contour (blue line in Figure 3.3a).

It is proposed that these normalized facial distances characterize facial expressions. E.g. upon smiling, lips are stretched apart which means that a large distance between lip corners indicates a happy expression. But there has to be a reference with respect to which we can say if this distance is large or not. For some persons, this distance may be large even in neutral state while for another person even on smiling this distance may be small. Size of lips may not just depend on the size of the face so normalizing facial distances by face size is insufficient.

Figure 3.4 illustrates that size of facial parts such as lips, mouth etc. may not depend on the size of the face as a whole. If a person X has a bigger face than another person Y , it does not imply that X 's nose is also longer than the nose of Y . Because of these interpersonal variations, we cannot say that facial distances characterize expressions unless there is a reference to these facial distances in the neutral face of the person. If the distance between lip corners

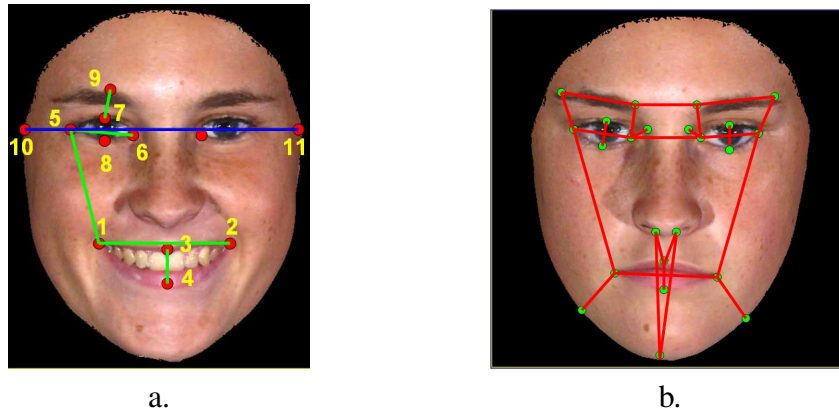
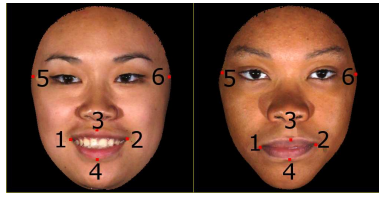


Figure 3.3: Distance features used in related works: a. The distances marked in green; normalized by the distance between points 10 and 11 (blue); are used in [163] as features. b. Distances used for extracting discriminating features in the method proposed by Tang and Huang [172] using manual features.

increases *with respect to that distance in a neutral face*, then it is possible that the person is smiling showing happy expression.

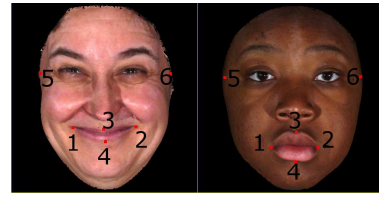
To illustrate the deficiency in taking facial distances as features, an example is presented in the first row of Figure 3.4, where the mouth opening (distance between points 1 and 2 in Figure 3.4a.) for a person with neutral face is more than that for a person smiling fully, even after normalization by facial scale. But this distance is expected to be more in a smiling face as compared to a neutral face. In order to analyse this defect over the full database, histograms showing the distribution of mouth opening over all the subjects for the six expressions are shown in Figure 3.5a. The distributions for different expressions have a significant overlap. For comparison, residue for left lip corner was considered since it indicates an increase in the mouth opening upon exhibiting expression. Figure 3.5b. shows that Happiness; which is characterised by an increase in mouth opening; is better discriminated from other expressions using residue for the left lip corner.

Similarly, in Figure 3.4b., same problem is encountered in measuring the mouth height (distance between points 3 and 4). Distances shown in the figure



Happiness grade4 Neutral
 $d_{12} = 0.41$ $d_{12} = 0.47$

a. Mouth opening



Happiness grade4 Neutral
 $d_{34} = 0.10$ $d_{34} = 0.21$

b. Mouth height

Figure 3.4: Ambiguities due to measurement of absolute distances. Mouth opening is more in neutral than in expression and mouth height is more in neutral than in Happiness. The distances given are normalized using d_{56} .

are the normalized by facial scale i.e. the distance between points 5 and 6 (Figure 3.4). Figure 3.6 compares histogram distributions for mouth height and residues and it is observed that residue for the mid-point of lower lip may be discriminative for surprise.

Because of interpersonal variations, using the 3D facial distances on the expressive facial model can lead to wrong predictions. This distance measure is referred to hereafter as the absolute distance. The problem in using absolute distances will also be encountered in using coordinate positions of the 3D facial feature points as features [223] [110], since features computed using absolute distances are nothing but linear combination of coordinate positions.

Major contribution of the proposed work is to utilize the ‘3D flow’ for FER instead of difference of facial distances or just using 2D optical flow which has been previously used by researchers. ‘Flow’ refers to the displacement of a facial point from neutral face to the expressive face and it utilizes information about the *change* that occurs in the face upon exhibiting an expression.

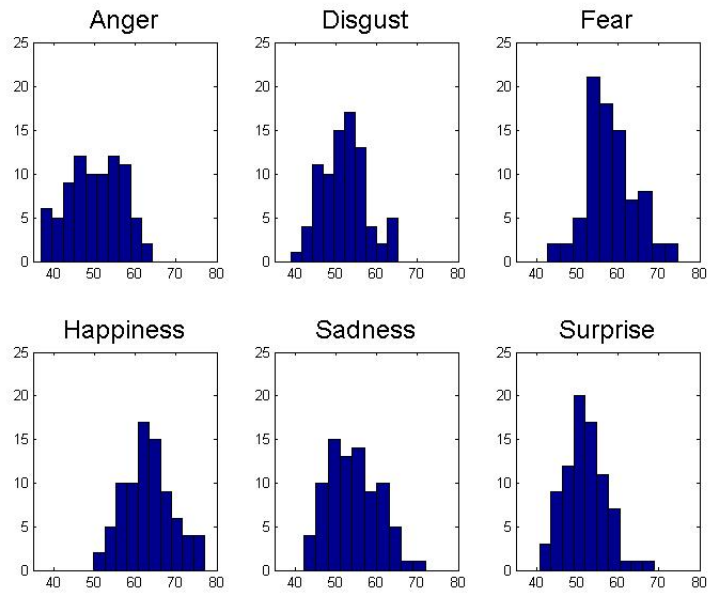
As a second contribution, an important issue has been discussed in this work which has not been addressed previously for 3D FER. In real time situations, a person might not exhibit expressions with full intensity. So, there is a need to refine the existing expression recognition systems to cope up with

the expressions of moderate or less intensity. Only in [122], all the four gradations are used for experimentation. However, not much analysis is done in accessing the performance of their expression recognition systems on the lower intensity gradations. To our knowledge, other works on the BU-3DFE database have used either the two highest gradations or do not explicitly mention about the gradations used for experimentation. We have tested our algorithm on all the four gradations of the expressions in set 1 of the experiments. A deeper analysis has also been performed through more rigorous experimentation.

Similar to the proposed approach, Tang and Huang [172] used the information from the change of the face shape from neutral to expression in the form of 24 normalized Euclidean distances between FFPs. These features are called ‘manual features’ by Tang and Huang (Tang and Huang’s method will be referred to as the ‘other’ method, henceforth). An automatic feature selection method is also proposed by them which selects relevant features from a pool of possible features. These features are called ‘auto features’. First absolute distances between a subset of the feature points are calculated. Then the corresponding distances in the neutral face are subtracted from those in the expressive face. These difference of distances (called henceforth as ‘difference features’) form the feature pool. In this work further references to Tang and Huang’s method indicate the method using manual features.

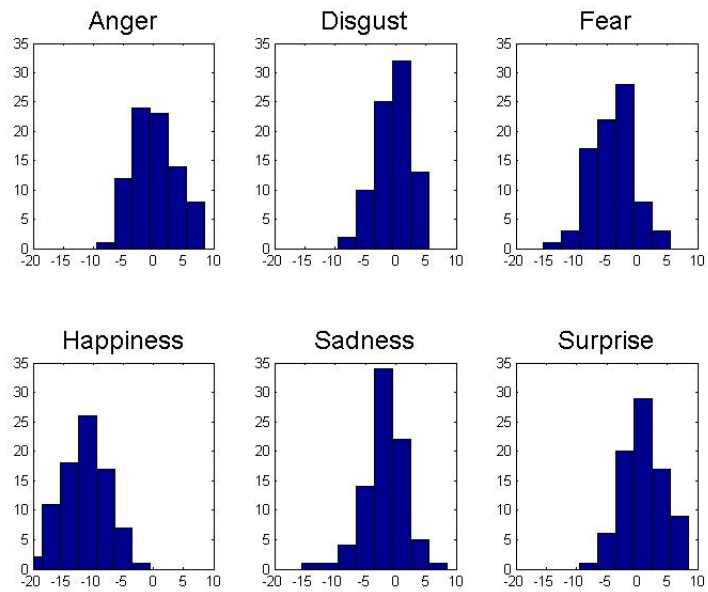
Difference features may not be robust in discriminating between two motions in different directions. Consider a pair of FFPs (Figure 3.7a.), A and B separated by a distance d_{AB} on an expressionless face. Let these points move to A' and B' , respectively when expression is exhibited and the new distance between them be $d_{A'B'}$. The difference feature for this pair of FFPs is given

Distribution of mouth-width features for different emotions



a. Histograms for absolute distance feature

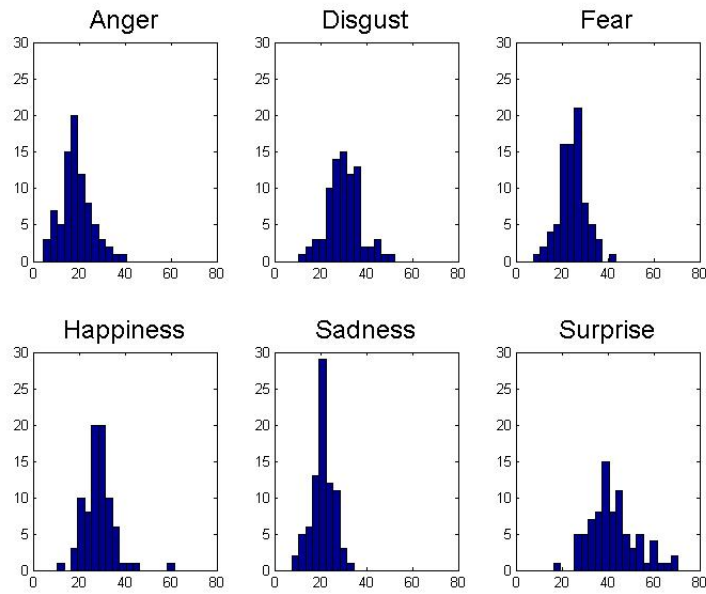
Distribution of residues for the left lip corner for x-coordinate



b. Histograms for residue

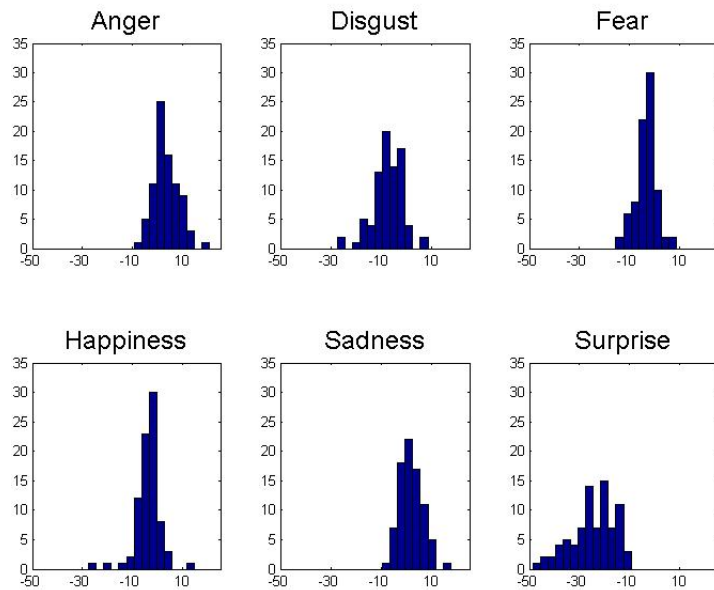
Figure 3.5: Comparing residues with absolute distance features for mouth width (opening). Increase in mouth opening is related to the motion of the left lip corner in the x-direction. Happiness is better discriminated from other expressions using residues than using absolute distance. See section 3.2.1 for detailed explanation.

Distribution of mouth-height features for different emotions



a. Histograms for absolute distance feature

Distribution of residues for mid-point of lower lip for y-coordinate



b. Histograms for residue

Figure 3.6: Comparing residues with absolute distance features for mouth height. Increase in mouth height is related to the motion of the mid-point of the lower lip in the y-direction. Surprise is better discriminated from other expressions using residues than using absolute distance. See section 3.2.1 for detailed explanation.

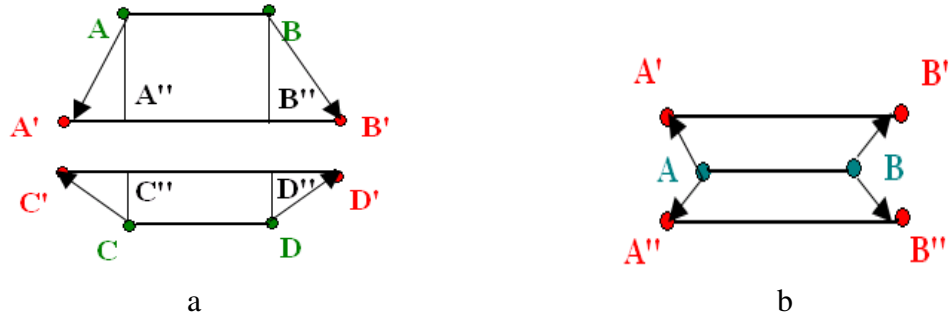


Figure 3.7: Shortcomings in taking difference of distances as features

by:

$$\begin{aligned}
 f_{AB} &= d_{A'B'} - d_{AB} \\
 &= A'A'' + B'B''
 \end{aligned} \tag{3.1}$$

In another case with points C and D , the difference feature is given by

$$\begin{aligned}
 f_{CD} &= d_{C'D'} - d_{CD} \\
 &= C'C'' + D'D''
 \end{aligned} \tag{3.2}$$

Although $f_{AB} = f_{CD}$, the motion of the two pairs of points is quite different. Similarly, in Figure 3.7b., A, B move to A', B' in one case and to A'', B'' in the other. Difference features are unable to discriminate between these movements which are in different directions altogether. This defect will be there in both manual and auto features used by the other method since both feature representations use difference features. Such ambiguities reflected in even a few of the features can lead to misclassification.

3D flow which is proposed in this work captures both the magnitude and direction of the change in the position of Facial Feature Points upon exhibiting an expression. This makes 3D flow more effective than difference features.

3.2.2 Feature Extraction

The 3D shapes are in the form of triangulated mesh models described in section 3.1 (Fig. 3.2b). Let the vertex coordinates of these 3D shapes be stored in matrices represented by $S_1^{e_j}, S_2^{e_j}, S_i^{e_j}, \dots, S_{100}^{e_j}$. Here $i = 1$ to 100 corresponds to the person, e corresponds to the expression and j corresponds to the gradation (intensities) of the expression e.g. gradations 1 to 4 of Happiness expression. Including different intensities of expressions is important because in real life we display both exaggerated and mild expressions. $S_i^{e_j}$ is a $N_v \times 3$ matrix, with N_v being the number of vertices in the model and 3 corresponds to x, y and z coordinate positions. $S_i^{e_j}$ contains vertices spread over the full face. 83 of these vertices correspond to prominent facial feature points (FFPs). Let the matrix containing coordinates of these 83 FFPs be represented as $D_i^{e_j}$.

Training and test data matrices $S_{tr}^n, S_{tes}^n, D_{tr}^{e_j}, D_{tr}^n, D_{tes}^{e_j}$ and D_{tes}^n are given. Here $tr \in TR$ and $tes \in TES$. TR and TES denote sets of characters used for ‘training’ and ‘test’ respectively and superscript n denotes ‘neutral’. Note that the full shape matrices S ’s are required to determine nose tip (See ‘Pre-processing’ below). The problem is to classify the test data into one of the six expressions. We achieve this using the following algorithm.

Pseudocode of the proposed algorithm (See section 3.2.2 for the notations used):

1. Inputs:

- 3D models of the test face, S_{tes}^n (all vertices), with the positions of 83 FFPs given in the form of matrices, D_{tes}^n and $D_{tes}^{e_j}$ corresponding to neutral and expression, respectively.
- N training examples ($S_{tr}^{e_j}, D_{tr}^{e_j}$ and D_{tr}^n) with expression labels.
- A multi-class classifier

2. For D_{tes}^n and $D_{tes}^{e_j}$, do the preprocessing (see below).
3. Find the 3D flow matrix, $F_{tes}^{e_j}$ and convert to feature matrix, $X_{tes}^{e_j}$ (see below).
4. Preprocess the training data in a similar way as the test data and train the classifier.
5. Feed $X_{tes}^{e_j}$ to the classifier
6. Output: The expression to which test data belongs.

Preprocessing

The raw 3D models, do not have the coordinate axes centered at the same point of the face. In the preprocessing, all the face models are aligned in such a way that the coordinate systems of all the face models have their origins at the nose tip. The nose tip is located by finding the highest vertex (i.e. one with the maximum z coordinate value) in the nose region. The nose region is found with reference to the positions of feature points lying near the nose tip. FFP coordinate values are normalized by dividing them by the distance between the outermost point of right-face contour and the outermost point of left-face contour (blue line in Figure 3.3a). This normalization takes care of the variations in face sizes across different persons.

Calculating 3D flow

After preprocessing each 3D model, 3D flow matrices $F_i^{e_j}$ are calculated using Eq. 3.3.

$$F_i^{e_j} = D_i^{e_j} - D_i^n \quad (3.3)$$

Here i, e, n and j refer to the person, expression, neutral and gradation, respectively. $F_i^{e_j}$ is a three dimensional matrix. To form the feature matrix we

Table 3.1: Details about training and test datasets for both sets of experiments

	Training data	Testing data
Set1	30 males 30 females 3 gradations	5 males 5 females The remaining gradation which is not used for training
Set2	25 males 25 females	5 males(Outside the training set) 5 females(Outside the training set) Gradations are same for both training and test datasets

convert $F_i^{e_j}$ into a two dimensional matrix by horizontally concatenating the 3D flow matrix. If the 3D flows are represented by (dx, dy, dz) , then the first one-third columns of the resultant two dimensional matrix correspond to dx , next one-third correspond to dy and the last one-third columns correspond to dz . The resultant matrix is the feature matrix denoted by $X_t^{e_j}$. Both the testing and training matrices have the same 2D structure.

3.2.3 Classification

The Support Vector Machine (SVM) is a supervised learning tool which is used for classification and regression problems. It finds a separating hyperplane in a higher dimension. SVM is used as the classifier for our experiments [35] and its performance is compared with that of Linear Discriminant Analysis (LDA). In training SVM, before feeding the feature matrices, values for each feature (each column) are scaled to the range of [0,1]. The testing data is also scaled using the same method as used for training data. For example, assume that the first feature of training data is scaled from [-10; +10] to [-1; +1]. If the first attribute of testing data is lying in the range [-11; +8], we must scale the testing data to [-1.1; +0.8]. We use RBF kernel for SVM which has two main parameters which need to be tuned viz. C and γ such that the classifier performs accurately on unknown data. A grid search is used to get the best

set of parameters using v -fold cross validation ($v = 5$ in our experiments). For each set of (C, γ) values, cross validation is performed and the set for which cross validation accuracy is maximum is used for classifying the test data. Details can be found in [87].

3.2.4 Experimental Results and Discussions

This work utilizes the spatial positions (x, y, z) for analysis, leaving out the (R,G,B) values. The matrix having these spatial coordinate values is a $p \times f \times 3$ matrix. Here p , f and 3 correspond to number of persons, number of FFPs and x,y and z (3 in number), respectively.

For the experiments, 60 subjects are chosen. 30 of them are female and 30 are male. Two sets of experiments were performed:

1. To check the performance of our method in classifying previously unseen *gradations* of expressions. We train the system on lower gradations of expressions and test on higher gradations.
2. To check the classification performance of our method in classifying expressions of previously unseen *persons* (Person Independent scenario). These experiments involve testing on persons who were not used for training the system.

In each of the two sets of experiments, we also compare our performance with Tang and Huang’s method using manual features, which was briefly described in section 3.2.1. For comparison this method is chosen since it also utilizes the geometric properties of FFPs in 3D space for the task of FER. This work is very close to the proposed work.

Table 3.2: Confusion matrices of the classification results for Set 1 of experiments. An: Anger, Di: Disgust, Fe: Fear, Ha: Happiness, Sa: Sadness, Su: Surprise. Training: 60 subjects, Testing: 10 subjects whose three gradations of expressions are used for training and remaining one gradation for testing.

(a) Our method

	An	Di	Fe	Ha	Sa	Su
An	80	0	10	0	10	0
Di	10	70	20	0	0	0
Fe	0	0	70	20	10	0
Ha	10	10	40	40	0	0
Sa	10	0	10	0	80	0
Su	10	0	0	0	0	90

(b) Tang and Huang [172]

	An	Di	Fe	Ha	Sa	Su
An	50	10	10	0	30	0
Di	10	70	10	0	10	0
Fe	0	0	70	10	20	0
Ha	10	10	40	40	0	0
Sa	10	0	30	0	60	0
Su	20	0	0	0	20	60

Training: Gradations 2,3 and 4

Testing: Gradation 1

(c) Our method

	An	Di	Fe	Ha	Sa	Su
An	80	0	0	10	10	0
Di	0	100	0	0	0	0
Fe	0	0	70	20	0	10
Ha	0	0	0	100	0	0
Sa	20	0	20	0	60	0
Su	0	0	0	0	0	100

(d) Tang and Huang [172]

	An	Di	Fe	Ha	Sa	Su
An	50	0	20	10	20	0
Di	0	60	20	20	0	0
Fe	10	0	40	20	30	0
Ha	0	0	10	90	0	0
Sa	20	0	10	10	60	0
Su	0	10	10	0	0	80

Training: Gradations 1,3 and 4

Testing: Gradation 2

(e) Our method

	An	Di	Fe	Ha	Sa	Su
An	70	10	0	0	20	0
Di	0	100	0	0	0	0
Fe	0	0	90	10	0	0
Ha	0	0	0	100	0	0
Sa	0	0	0	0	100	0
Su	0	0	0	0	0	100

(f) Tang and Huang [172]

	An	Di	Fe	Ha	Sa	Su
An	70	0	0	10	20	0
Di	0	80	0	10	10	0
Fe	0	0	80	20	0	0
Ha	0	0	0	100	0	0
Sa	10	0	20	10	60	0
Su	0	0	0	0	10	90

Training: Gradations 1,2 and 4

Testing: Gradation 3

(g) Our method

	An	Di	Fe	Ha	Sa	Su
An	90	0	0	0	10	0
Di	0	100	0	0	0	0
Fe	0	0	100	0	0	0
Ha	0	0	0	100	0	0
Sa	0	0	10	0	90	0
Su	0	0	0	0	0	100

(h) Tang and Huang [172]

	An	Di	Fe	Ha	Sa	Su
An	80	10	0	10	0	0
Di	0	80	0	20	0	0
Fe	0	0	70	20	10	0
Ha	0	0	0	100	0	0
Sa	10	0	20	10	60	0
Su	0	0	0	0	0	100

Training: Gradations 1,2 and 3

Testing: Gradation 4

Table 3.3: Confusion matrices of the classification results for set 2 of experiments. An: Anger, Di: Disgust, Fe: Fear, Ha: Happiness, Sa: Sadness, Su: Surprise. Training: 50 subjects, Testing: 10 subjects outside training set.

(a) Our method							(b) Tang and Huang [172]						
	An	Di	Fe	Ha	Sa	Su		An	Di	Fe	Ha	Sa	Su
An	80	0	0	5	15	0	An	70	0	5	5	20	0
Di	0	95	0	5	0	0	Di	5	80	0	10	5	0
Fe	0	0	90	10	0	0	Fe	10	0	75	15	0	0
Ha	0	5	0	95	0	0	Ha	0	0	0	100	0	0
Sa	0	0	10	0	90	0	Sa	5	0	20	10	65	0
Su	0	0	0	0	0	100	Su	5	5	5	0	5	80

(c) Our method							(d) Tang and Huang [172]						
Gradations 3 and 4													
	An	Di	Fe	Ha	Sa	Su		An	Di	Fe	Ha	Sa	Su
An	70	0	7.5	5	17.5	0	An	65	5	2.5	7.5	20	0
Di	2.5	87.5	0	10	0	0	Di	7.5	75	5	10	2.5	0
Fe	0	0	67.5	27.5	2.5	2.5	Fe	7.5	2.5	52.5	17.5	15	5
Ha	5	2.5	12.5	80	0	0	Ha	2.5	2.5	0	92.5	0	2.5
Sa	7.5	0	15	2.5	75	0	Sa	17.5	0	22.5	7.5	52.5	0
Su	2.5	2.5	0	0	0	95	Su	5	2.5	0	0	7.5	85

Gradations 1,2,3 and 4

Set 1 of experiments

Set 1 of the experiments aims at assessing the classification performance when previously unseen gradations of expressions are encountered. This situation is possible when our classifier is trained to classify only higher intensity expressions and the test data contains low gradations of expressions or vice-versa. These experiments are important considering the difficulty in training the classifier for all possible gradations of expressions that can be exhibited by different persons. If the performance of classifier is good over previously unseen gradations, it can recognize expressions of various gradations.

In set 1 of experiments, out of the four gradations of expressions, three are used for training and the remaining one gradation is used for testing which will serve as the previously unseen gradation. Even for the same person, the

expression at a lower gradation can be very differently exhibited as the same expression at a higher gradation. Consequently, we can also test on previously unseen gradations of the same person. Therefore, test subjects are included in the training database but testing is done with the gradation for which data is not trained.

Set 2 of experiments

In set 2 of the experiments, it is assumed that we have captured the data for all the gradations of expressions for each of the person in the training database. The aim in this set of experiments is to see how well the proposed method performs on previously unseen *persons* rather than *gradations*. We train the classifier with all four gradations of a specific set of persons but test on different set of persons. In [172] results are reported only for the highest two gradations of expressions. For comparison with their method, experiments are also performed using only the two highest gradations. In this set of experiments the proposed approach has been compared with the standard approach using 2D optical flow. Feature selection has also been explored.

The details about training and testing data for both sets of experiments is given in Table 3.1. Please note that for testing, in addition to the expressive 3D facial model, a model of the subject's neutral face is also needed.

Results and Discussions

Classification results for sets 1 and 2 are given in Table 3.2 and Table 3.3, respectively in the form of confusion matrices. In the confusion matrix an element (i, j) (i.e. row corresponding to expression i and column to expression j) shows the number of test samples which were predicted to belong to expression j when the sample actually belongs to expression i . Results are compared with that of Tang and Huang's method using manual features on the same pair

Table 3.4: Comparison of the average recognition rates of our method and the method in [172] for set 1 of experiments. Training: 60 subjects, Testing: 10 subjects whose three gradations of expressions are used for training and remaining one gradation for testing.

Gradation of Expression		Avg. Recognition Rate (%)	
Training	Testing	Tang and Huang [172]	Our method
2,3,4	1	58.3	71.7
1,3,4	2	63.3	85.0
1,2,4	3	80.0	93.3
1,2,3	4	81.7	96.7

of training and test datasets. Clearly the performance of our method is much better.

Table 3.5: Comparison of the average recognition rates of our method and the method in [172] for set 2 of experiments. Training set: 50 subjects, Testing set: 10 subjects outside training set.

Gradation of expressions	Tang and Huang [172](%)	Our method(%)
3,4	78.3	91.7
1,2,3,4	70.4	79.2

The average recognition rate in recognizing six expressions are also compared for set 1 and set 2 of the experiments in Table 3.4 and Table 3.5, respectively. For set 1 of experiments, it is observed that with an increase in the gradation (from 1 to 4) of the test expression, both methods perform better. This indicates that higher gradations of expressions seem to convey more information about the expression while the lower grades of expressions are much closer to the neutral and also to one another. However our method performs much better even when all four gradations are tested.

In set 2 of experiments, it is observed (Table 3.5) that the average recognition rate is less when all four gradations are taken. The reason for this is similar to what we have just explained. The difference in between neutral and the lower gradations of expressions is not expected to be very evident but the

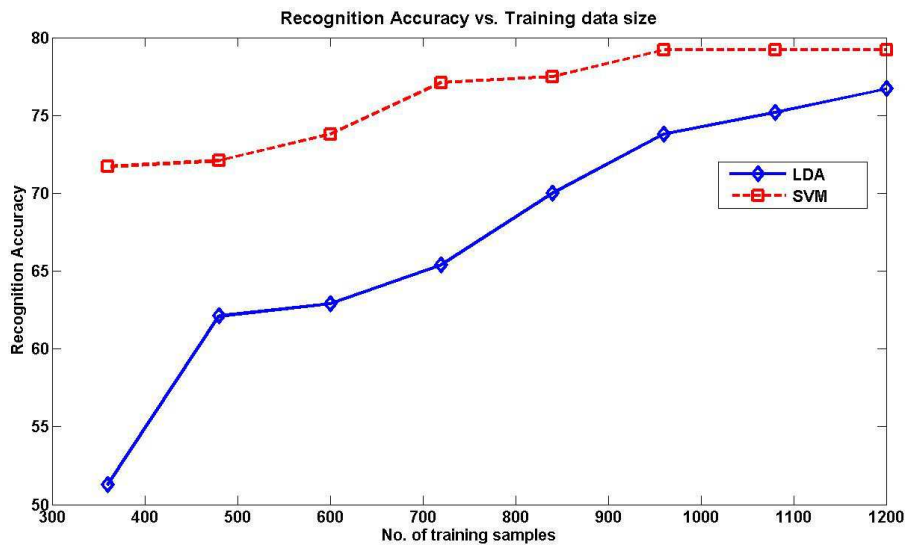


Figure 3.8: Variation of recognition accuracy of SVM and LDA with training data size. Observe that the performance of LDA is affected more by the training data size. Experiments have been conducted using all 4 gradations of expressions, as in set 2 of experiments.

difference is more prominent for higher gradations.

For set 2, the performance of SVM classifier is compared with that of LDA. Results are presented in the form of the confusion matrices in Table 3.6 and the average recognition rates are compared in Table 3.7. From the results, it is found the SVM perform much better than LDA especially when only gradations 3 and 4 are considered. A possible reason can be non-linear separation of classes in feature space. Non-linear separation is indicated by the number of support vectors in training the SVMs [219]. Due to complicated decision boundaries, usually number of support vectors increases with the non-linearity of the data. The number of support vectors were found to be 376 out of the total 600 training samples which comes out to be around 63% which is a large fraction. This indicates the possibility of non-linear separation of data which is better resolved by SVMs than LDA.

In the case of LDA, a slight increase in accuracy was observed when all four gradations are considered. Intuitively, this is unexpected but it should be noted

Table 3.6: Confusion matrices of the classification results for set 2 of experiments with SVM and LDA as classifiers. An: Anger, Di: Disgust, Fe: Fear, Ha: Happiness, Sa: Sadness, Su: Surprise.

	An	Di	Fe	H	Sa	Su
An	80	0	0	5	15	0
Di	0	95	0	5	0	0
Fe	0	0	90	10	0	0
Ha	0	5	0	95	0	0
Sa	0	0	10	0	90	0
Su	0	0	0	0	0	100

SVM

	An	Di	Fe	Ha	Sa	Su
An	95	0	5	0	0	0
Di	15	50	20	5	10	0
Fe	5	15	50	10	15	5
Ha	0	10	5	85	0	0
Sa	15	5	10	0	70	0
Su	0	0	5	0	0	95

LDA

a. Gradations 3 and 4

	An	Di	Fe	Ha	Sa	Su
An	70.0	0.0	7.5	5.0	17.5	0.0
Di	2.5	87.5	0.0	10.0	0.0	0.0
Fe	0.0	0.0	67.5	27.5	2.5	2.5
Ha	5.0	2.5	12.5	80.0	0.0	0.0
Sa	7.5	0.0	15.0	2.5	75.0	0.0
Su	2.5	2.5	0.0	0.0	0.0	95.0

SVM

	An	Di	Fe	Ha	Sa	Su
An	70.0	2.5	12.5	2.5	10.0	2.5
Di	17.5	67.5	10.0	5.0	0.0	0.0
Fe	7.5	0.0	65.0	22.5	5.0	0.0
Ha	0.0	2.5	7.5	90.0	0.0	0.0
Sa	2.5	0.0	22.5	0.0	75.0	0.0
Su	0.0	0.0	7.5	0.0	0.0	92.5

LDA

b. Gradations 1,2,3 and 4

that LDA considers the whole training data set in determining the discriminant function, the performance depending on the size of training dataset [104]. Hence a 2% improvement in the recognition accuracy could be attributed to the increase in training data size which improved recognition accuracy more than the decrease in accuracy due to inclusion of lower intensity expressions. On the other hand SVM looks at the support vectors and not the whole data distribution and its performance doesn't depend a lot on the size of training data. To validate this reasoning, experiments were conducted with different sizes of training dataset using all four gradations. Results are shown in Figure 3.8. It can be seen that LDA's performance worsens with a decrease in training data size while SVM's performance doesn't suffer that much for a smaller training data size.

Table 3.7: Comparison of the average recognition rates using SVM and LDA as classifiers for set 2 of experiments.

Gradation of expressions	LDA (%)	SVM(%)
3,4	74.2	91.7
1,2,3,4	76.7	79.2

	An	Di	Fe	Ha	Sa	Su		An	Di	Fe	Ha	Sa	Su
An	80	0	0	5	15	0	An	80	0	5	0	15	0
Di	0	95	0	5	0	0	Di	5	85	0	10	0	0
Fe	0	0	90	10	0	0	Fe	5	0	60	15	20	0
Ha	0	5	0	95	0	0	Ha	0	5	5	90	0	0
Sa	0	0	10	0	90	0	Sa	5	0	5	5	85	0
Su	0	0	0	0	0	100	Su	0	0	0	0	0	100
Average Recognition Rate with depth information = 91.7%							Average Recognition Rate without depth information = 83.3%						

Table 3.8: Efficacy of using 3D flow than just using 2D optical flow.

In Table 3.9, our performance is compared with that of other recent works on recognizing facial expressions. Only those works have been compared with, which are evaluated on the BU-3DFE database which is a publicly available benchmarked data set. Works using their own databases; many of which are not publicly available; have been excluded to maintain uniformity in comparison.

From Table 3.9, it is observed that [110] reports an Average Recognition Rate (A.R.R.) of 96.1% which appears to be higher than that of our method. However, the A.R.R. reported in [110] is for binary classification tasks namely Anger vs. non-Anger, Disgust vs. non-Disgust and so on. Usually, a two-class classification is considered to be easier than a multi-class classification and the possibility of confusing between different expressions is less with fewer classes. [110] does not report six class classification results.

Comparison with standard 2D optical flow Experiments were also conducted to show that the 3D flow is more informative than just the 2D optical flow. This is achieved by first performing FER using residues of all the three (x, y, z) coordinates and then using residues for only the (x, y) coordinates.

Method	A.R.R.	Remarks (if any)
Wang[193]	83.6	
Venkatesh[188]	81.7	
Yin[167]	80.2	
Tang[172]	78.3	
Tang[174]	87.1	
Soyel[163]	91.3	
Mpiperis[122]	90.5	
Gong[76]	76.2	
Rosato[147]	80.1	
Soyel[164]	88.3	
Maalej[110]	96.1	Experimented on 20 subjects performing only expression vs. non-expression binary classifications
Barretti[24]	77.5	
Zhao[223]	87.2	
Sha[156]	83.5	
Our method	91.7	

Table 3.9: Comparison of the performances of works on the BU-3DFE database. A.R.R.: Average Recognition Rate.

This experiment is a part of the set 2 of the experiments and was performed using only the gradations 3 and 4 of the expressions. Results are shown in Table 3.8 in the form of the confusion matrices for the two experiments. It is clear that the classification accuracy is much higher when the depth information is also included for the classification purpose.

Feature Selection Presently, the feature vector has residues for each of the 83 FFPs along x , y and z directions. This leads to a feature vector of dimension $83 \times 3 = 249$. It might be possible that not all the features corresponding to 249 dimensions are relevant. Therefore, feature selection was tried out using the Fisher ratio test [197]. Dimensionality reduction methods such as the Principal Component Analysis (PCA) transform the features to another space. It is not clear if features which are discriminative in original space will also be discriminative in the reduced space after applying PCA. So we do feature selection in the original data space. It was observed that feature selection does

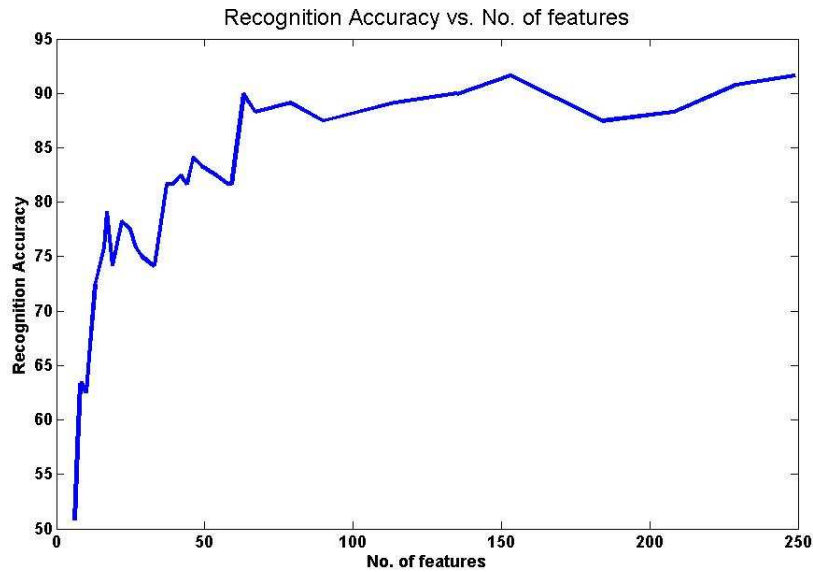


Figure 3.9: Variation of Recognition Rate (average) with number of features. It can be observed that feature selection could not improve the performance. The results are for set 2 of experiments.

not improve the results. The best performance of 91.7% was obtained using 153 features which is the same as using all features. A variation of Average Recognition Rate with number of features is shown in Figure 3.9.

In Figure 3.10, we also give some examples of images where the the method in [172] could not give correct classification. These cases were correctly classified by our method. Similar problems are expected in the automatic method suggested in [172] as well, because of taking distance measures (a problem referred to in section 3.2.1). Moreover, the proposed approach takes into account the 3D flow of each of the 83 points. Thus, variations on all parts of eye, brow etc are captured giving a better idea of the total variation.

In Figure 3.10a a curling of lips is observed which could not be captured by the other method. This is due to taking fewer number of points on the lips for feature calculation. In Figure 3.10c, due to lip stretching, there is an ambiguity for the feature corresponding to lip width. It is classified as happy since happy also has stretched lips. The movement of other parts of the face such as the frown in the eyes could not be differentiated from the movements

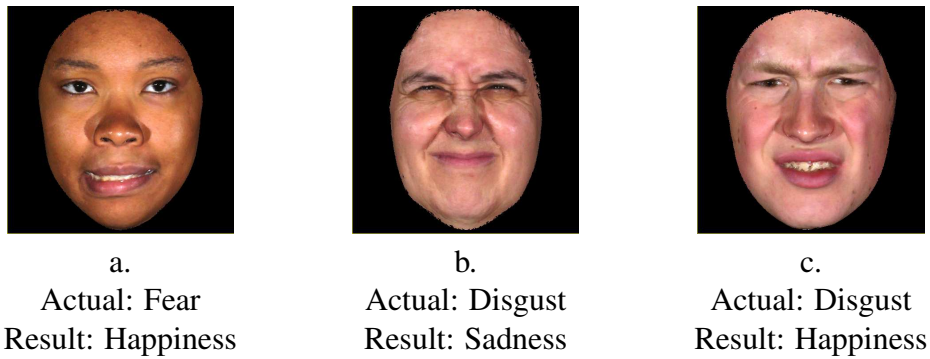


Figure 3.10: Some misclassified images using the method in [172]. These were correctly classified by our method.

in other expressions. But with our method, no ambiguity will be there since the movement of other points gives indication of happy expression. Also, points on the nose change a lot but have not been considered in [172]. Similarly, the expression in Figure 3.10b is misclassified as sad. Here the lip corner move upwards but generally in sad expression, lip corners move downwards. The ambiguity is related to Figure 3.7b. This ambiguity can be the possible cause of misclassification. In our method directional content of the motion avoids this ambiguity.

3.2.5 Limitations

There are a few improvements possible in using 3D residues which are listed as below and motivate us to use Deformation Modeling for FER (Section 3.3):

1. The present algorithm requires a neutral 3D model of the test subject which may not be possible in actual testing conditions. Having the neutral 3D model of the person is only possible for known subjects. To overcome this problem, motion of FFPs can be modeled and observing motion of FFPs between any two grades of expression can give a clue about the expression without actual need of a neutral.
2. Presently the feature dimensionality is 249 which is very high. Prelim-



Figure 3.11: Movement of the right lip corner while exhibiting happy expression.

inary experiments on feature selection did not prove to be helpful and needs to be explored more deeply.

3.3 Deformation Modeling: Subject Independent FER

3.3.1 Background

One of the challenges identified in the method using 3D residues is to obtain subject independence (Section 3.2.5) which enables the algorithm to generalize to novel faces. Apart from the issue of availability of the neutral face model, there are other difficulties as well when a neutral model is needed. E.g. Even if a person is monitored using a video camera; it is difficult to ascertain when the face is actually neutral. *Requirement of a neutral 3D model limits the application of FER methods. The method proposed in this section overcomes this limitation as it does not need any neutral 3D model of the test subject.*

The proposed approach models facial deformations caused by expressions. Deformation of a face is reflected by the movement of various landmark points on the face. To understand how an expression is modeled, consider Figure

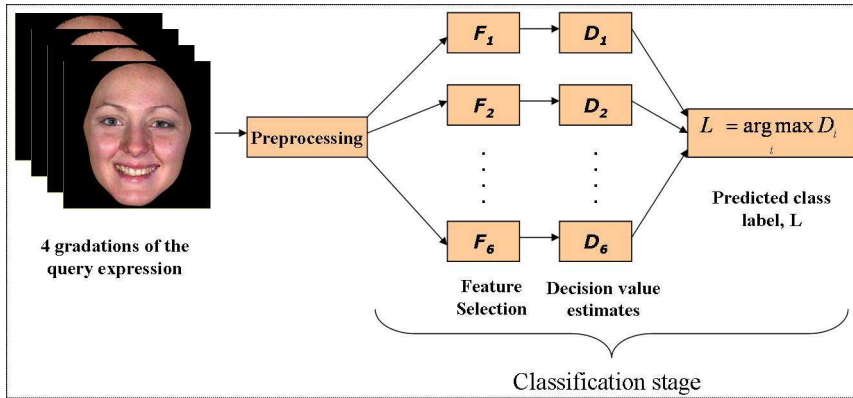


Figure 3.12: Classification scheme for the proposed algorithm. F_i s represent the relevant feature set for each of the six expressions while D_i s represents the decision value estimates of the query expression for each of the six expressions.

3.11 where the right lip corner position moves from point p_1 to p_4 from a low gradation of the ‘Happiness’ expression to the highest gradation of the same expression. Gradation refers to the intensity of the expression. *Our proposition is that movement of this point will be similar across different persons when they exhibit ‘Happiness’ expression.* Deformation model is a representation of that similarity. In this work the temporal variation of 3D spatial position of the landmark points is taken as the measure of similarity.

3.3.2 Feature Extraction

Let $f_p^i(n)$ represent the p^{th} landmark point on the i^{th} person at a temporal instant n . Here $p = 1$ to 83 (corresponding to 83 facial landmark points); $i = 1$ to 100 (corresponding to 100 persons in the database) and $n = 1$ to N_s where N_s is the number of temporal samples found out by interpolation. The pseudocode of the proposed algorithm for feature extraction followed by classification is as follows:

Pseudocode of the proposed algorithm (See sections 3.2.2 for notations not defined in this section):

1. Inputs:
 - 3D models of the test face, S_{tes}^n (all vertices), with the positions of 83 FFPs given in the form of D_{tes}^{ej} .
 - N training examples (S_{tr}^{ej} and D_{tr}^{ej}) with expression labels.
2. For each D_{tr}^{ej} , find 3D position at each facial point in spherical coordinates
 - Set nose tip as the origin.
 - Using Cubic spline interpolation find the (x, y, z) coordinates of the point at intensities intermediate in between gradations j and $j + 1$, $j \in [1, 3]$. Thus (x, y, z) coordinates $f_p^i(n)$ are obtained with $n = 1$ to N_s .
 - Transform the coordinates of $f_p^i(n)$ from Cartesian $((x, y, z))$ to spherical coordinate system (r, θ, φ) .
 - Find $\vec{\theta}_{p_0}^{i_0}$ and $\vec{\varphi}_{p_0}^{i_0}$ as per equation 3.4. $\vec{\theta}_{p_0}^{i_0}$ and $\vec{\varphi}_{p_0}^{i_0}$ represent 3D position of the p_0^{th} landmark point of i_0^{th} person.
3. For i^{th} person, the feature vector is given by eq. 3.5
4. Form the feature matrix X_{tr} by concatenating feature vectors for all the persons in the training set.
5. Select relevant features for each expression, using the significance ratio test.
6. Train the SVM classifier using X_{tr} in one vs. all scheme.
7. Form testing matrix in the same way as X_{tr} was formed.
8. Classify the test data X_{tes} using one vs. all scheme of SVM.

Details of the algorithm are given in section 3.3.3.

3.3.3 Experiments

Interpolation

3D face models from BU3DFE database [211] are selected for 60 subjects in training dataset and 22 subjects in the testing dataset. (x, y, z) positions of 83 landmark points provided for all these models is utilized. It should be noted that there are four gradations of expressions available which correspond to only four temporal samples from neutral to peak of the expression. However, only 4 temporal samples are insufficient to have a reasonable estimation of the motion direction. Intermediate samples are found using cubic spline interpolation. After interpolation at a landmark point; say the left eyebrow corner; we have 61 temporal samples for each of the 60 subjects used for training and 22 subjects used for testing.

After interpolation, the spatial positions of the points are transformed from Cartesian to the spherical (r, θ, φ) coordinate system. This is because the parameters θ and φ directly give the required information about the 3D position of a point with respect to the origin by just two parameters θ and φ instead of three parameters, x, y and z . The parameter r is not considered here since we are concerned with the direction of motion only. Let the θ and φ values for $f_p^i(n)$ be represented by $\theta_p^i(n)$ and $\varphi_p^i(n)$, respectively.

Modeling the motion at a landmark point

For the purpose of modeling the motion direction, 3D position of the points $f_p^i(n)$ for a fixed value of p say, p_0 , are considered. Out of the 60 subjects in the training dataset, a subset I composing of 30 subjects was selected at random. 2D histograms of $\theta_{p_0}^i(n)$ and $\varphi_{p_0}^i(n)$ values were plotted for $f_{p_0}^i(n)$ where $i \in I$. This procedure was repeated 5 times. The histograms thus found are shown as images in Figure 3.13a. The figure shows that the histograms

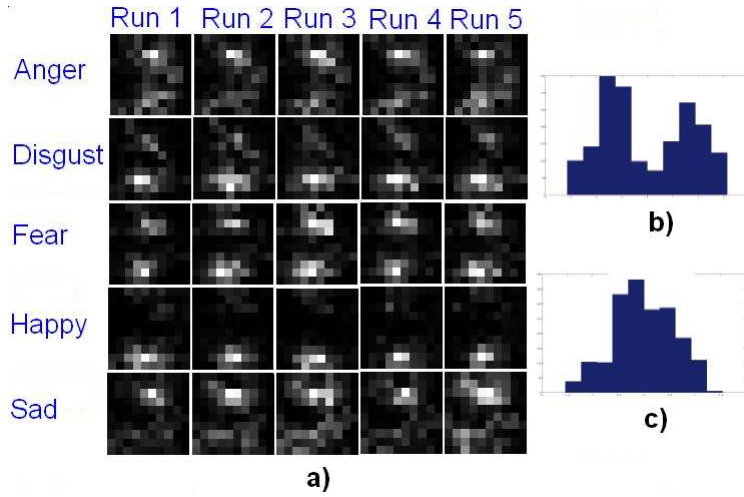


Figure 3.13: a. Histograms for the distribution of $\theta_{p_0}^i(n)$ and $\varphi_{p_0}^i(n)$ over 5 runs b. Separate histogram for $\theta_{p_0}^i(n)$ c. Separate histogram for $\varphi_{p_0}^i(n)$ corresponding to the 2D histogram in run1 and Anger expression. (See ‘*Modeling the motion at a landmark point*’ below for more details.)

are consistent over all the 5 runs. This means that the temporal samples at the landmark point for different subjects, belong to a specific nature of distribution. This supports our proposition about the motion of landmark point being similar in different persons. Consistency in the histograms for the same expression and the difference in between the histograms of different expressions, clearly shows that the motion directions are discriminative and can be used as features for FER.

By projecting the 2D histogram on each of the θ and φ axes, the 2D histogram is decomposed into two 1D histograms showing the distribution of $\theta_{p_0}^i(n)$ and $\varphi_{p_0}^i(n)$, separately (Figure 3.13 b. and c.). The histogram in Figure 3.13 b. is nothing but the histogram of the values of $\theta_{p_0}^i(n)$ alone without considering $\varphi_{p_0}^i(n)$. Similar is the case with Figure 3.13c. Deformations are modeled as these 1D distributions.

For each person, instead of using all of $\theta_{p_0}^{i_0}(n)$ and $\varphi_{p_0}^{i_0}(n)$ we represent them by $\vec{\theta}_{p_0}^{i_0}$ and $\vec{\varphi}_{p_0}^{i_0}$ which are given by:

$$\begin{aligned}\vec{\theta}_{p_0}^{i_0} &= \left[\theta_{(1)p_0}^{i_0} \theta_{(2)p_0}^{i_0} \theta_{(3)p_0}^{i_0} \theta_{(4)p_0}^{i_0} \right] \\ \vec{\varphi}_{p_0}^{i_0} &= \left[\varphi_{(1)p_0}^{i_0} \varphi_{(2)p_0}^{i_0} \varphi_{(3)p_0}^{i_0} \varphi_{(4)p_0}^{i_0} \right]\end{aligned}\quad (3.4)$$

where, $\theta_{(1)p_0}^{i_0}$, $\theta_{(2)p_0}^{i_0}$, $\theta_{(3)p_0}^{i_0}$ and $\theta_{(4)p_0}^{i_0}$ represent the mean, variance, skewness and kurtosis of $\theta_{p_0}^{i_0}(n)$ with n varying from 1 to N_s . Similar notation is used for φ as well.

For the i_0^{th} person, the feature vector is given by

$$\vec{x}_{i_0} = \left[\vec{\theta}_1^{i_0} \quad \vec{\theta}_2^{i_0} \quad \dots \quad \vec{\theta}_p^{i_0} \quad \dots \quad \vec{\theta}_{83}^{i_0} \quad \vec{\varphi}_1^{i_0} \quad \vec{\varphi}_2^{i_0} \quad \dots \quad \vec{\varphi}_p^{i_0} \quad \dots \quad \vec{\varphi}_{83}^{i_0} \right] \quad (3.5)$$

Classification

There are total 83 landmark points used for feature extraction. However, for each expression, only a subset of these 83 feature points is relevant. These subsets are chosen using the significance ratio test [197]. Six individual classifiers are implemented using the one vs. all scheme of SVM. The classification scheme is given in Figure 3.12. For a query expression, each classifier gives a decision value estimate. Predicted expression corresponds to that expression for which classifier gives the maximum decision value.

Results

Results of individual classification are given in the form of the Receiver Operating Characteristic (ROC) curves along with areas under the curves in Figure 3.14. It can be seen that the AUCs are very close to 1 for all the individual binary classifiers. However, in the case of fear vs. other expressions, AUC is not that good reflecting less discrimination between Fear and other expressions. Results of the final classification are presented in Figure 3.15 in the form of a

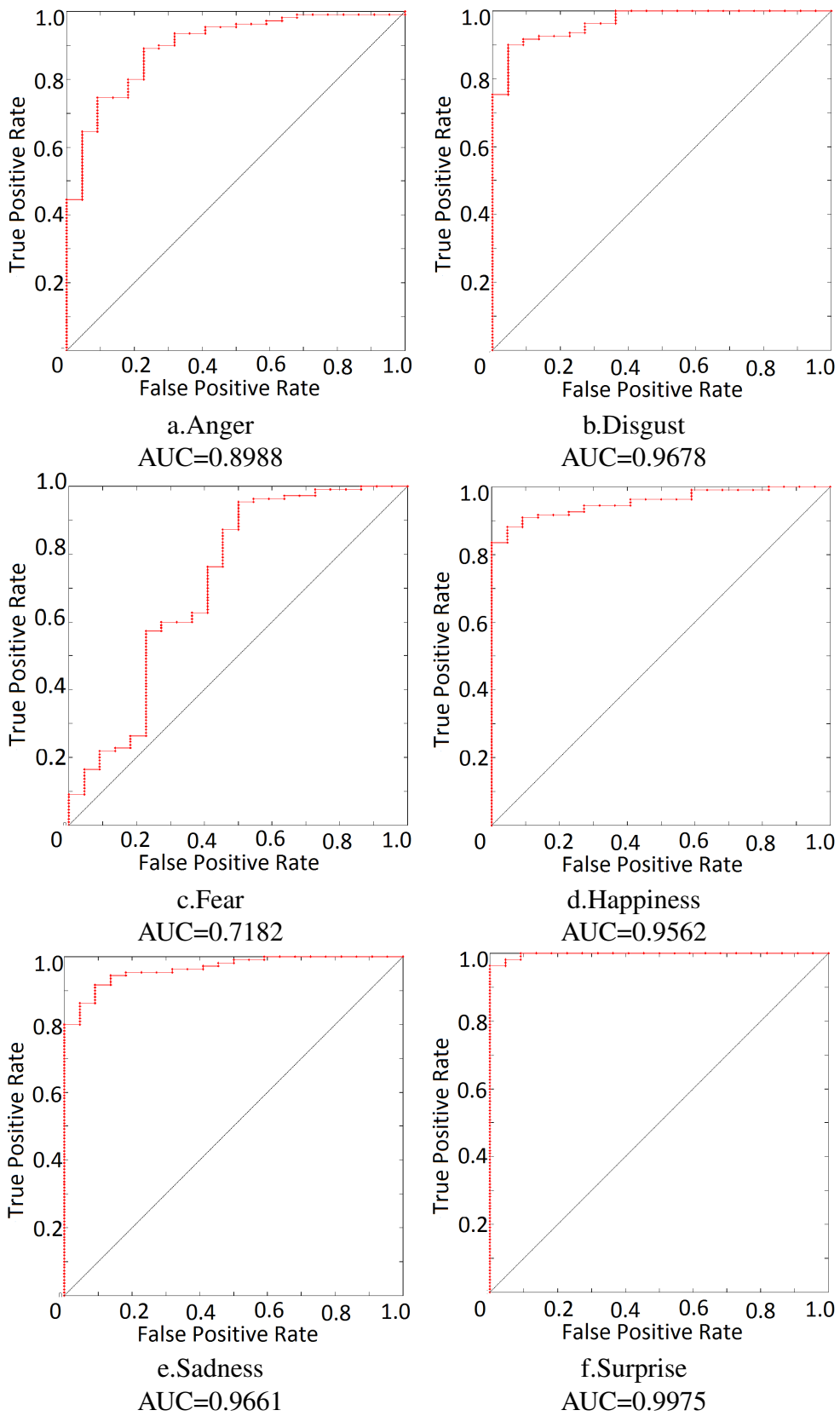


Figure 3.14: a. to f. ROC curves for individual classifiers with the Areas under ROC curves (AUCs).

	AN	DI	FE	HA	SA	SU
AN	72.7	9.1	4.5	0	13.6	0
DI	0	100	0	0	0	0
FE	4.5	9.1	45.5	31.8	9.1	0
HA	0	4.5	0	95.5	0	0
SA	9.1	4.5	4.5	0	77.3	4.5
SU	0	0	9.1	0	0	90.9

Figure 3.15: Confusion matrices of the classification results. An: Anger, Di: Disgust, Fe: Fear, Ha: Happiness, Sa: Sadness, Su: Surprise. Training: 60 subjects, Testing: 22 subjects.

confusion matrix.

Classification results show that the highest recognition rate of 100% has been achieved for Disgust. This is possibly because of a consistent curling of nose for disgust over different subjects. The average recognition rate is 87.3% if we ignore the results for Fear. Even including the results of Fear, the average recognition rate is 80.3%.

Chapter 4

Bimodal Spontaneous Emotion

Recognition Applied to Multi-actor

Emotion Recognition

4.1 Emotion Recognition in Movies

As identified from the literature review (Section 2.2), most of the existing works on ER have looked at interpreting emotions in a laboratory environment which has constraints of sufficient illumination, minimal background auditory noise, frontal or near-frontal facial pose, almost exaggerated facial expressions, minimal head motions etc. However, these constraints cannot be applied to a natural environment which we experience in real life. This brings out the necessity of developing an ER approach suitable for natural environments. Recent studies [189] have also identified this need and the work covered in this chapter aims at recognizing human emotions in a natural environment.

An important feature of a natural environment is the presence of multiple modalities conveying information about the emotion, prominent of which are facial expressions, bodily gestures, speech acoustics and content of dialogs.

Most of the works attempt to recognize human emotions using just the facial expressions. However, psychological studies defining human emotions also say that human emotions cannot be judged based on FER only [142] [128]. Rather representation and recognition of emotion can be enhanced by fusing information from multiple sensory expressions instead of just using one modality [9].

For ER in a natural environment, an important issue is to have representative data. Most of the emotion databases record exaggerated emotions under laboratory conditions. Even the 3D data used in the previous sections of this thesis was lab recorded and could only incorporate one difficulty of a natural environment, namely subtle expressions. 3D data with other difficulties of a natural environment is not easily available due to the capture and storage complexities. Consequently, in order to develop algorithms for near real life data, 2D data in the form of movie clips was chosen for further research. Movies are closer to the natural environment as compared to lab recorded data. Another advantage of movies is that faces, dialogues and the bodily expressions of the movie characters are hopefully well recorded and represented. Because of this advantage detecting events and concepts in movies has been a recent trend in research [68] [159].

A movie contains clues for ER mainly in the form of facial expressions, body gestures, speech acoustics and lexical clues from dialogs. Extracting relevant information from visual clues becomes challenging due to variations in facial pose and scale (Figure 4.1d). If we consider the facial feature point approach for feature extraction, facial feature localization and tracking becomes problematic for a small face size.

Consider an image in Figure 4.1d of resolution 640×480 pixels. In the image, the face region occupies a region of about 50×70 pixels which is just 1% area of the whole image. In such cases, relevant regions of the face such as

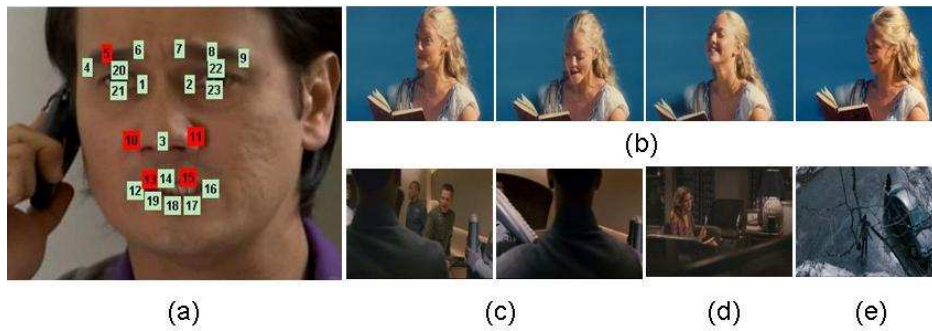


Figure 4.1: a) 23 FFPs marked on the face. Red points (FFPs no. 5, 10, 11, 13, 15) show the five most discriminating points as found out by feature selection. b-e show difficulties encountered in movie data considering the visual clues; b) Head motion c) Face Occluded in frames d) Small face size e) No face detected.

eyes, mouth etc. will be hard to locate. Usually scale changes are dealt with, by normalizing the extracted features using a facial distance such as the distance between two eyes. But, this distance is shortened for a non-frontal face making the choice of a normalizing distance difficult. Problem with pose variations is the loss of feature points that are seen in frontal view.

With appearance based approaches, facial appearance can be extremely different for two different views. Shadows create extraneous edges in the face image causing problems for filtering operations on the image.

Among other challenges, if the whole face becomes occluded (Figure 4.1c.) tracking fails. In such cases, if the face again becomes visible, there is a possibility of reinitializing the tracker. But if the face remains occluded, there is no possible for further analysis using visual clues. Tracking can also fail because of significant head motions (Figure 4.1b.). Sometimes, face might not be visible at all, while the dialogue is going on in the background (Figure 4.1e). This rules out the possibility of facial expression analysis.

Extracting emotional clues from speech acoustics is affected by background noises present in a natural environment. Out of the different possible acoustic features for ER, it is also challenging to identify and automatically extract relevant features in speech signals [218]. These challenges brought out in the

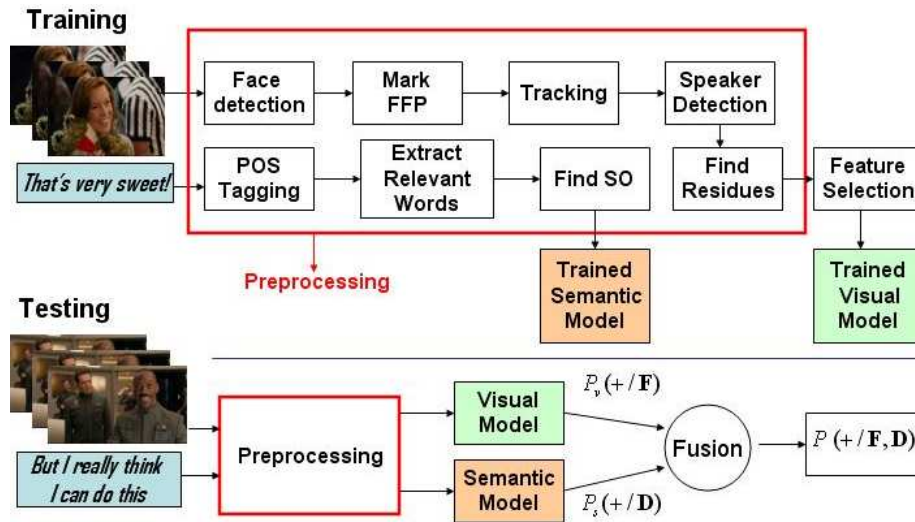


Figure 4.2: Framework for the proposed system.

context of movies are also applicable to any natural environment.

Most of the works on ER using audio related features use speech acoustics as features. However, considering the above mentioned challenges, information extracted from acoustic clues may be insufficient for identifying emotions in a natural scenario. [19] mention that, “the closer we get to a realistic scenario, the less reliable is prosody as an indicator of the speaker’s emotional state”. [49] show that lexical clues can perform better for ER as compared to acoustic clues. However, movie dialogues are very short and succinct making lexical analysis of dialogues difficult.

In this work a bimodal framework is proposed (Figure 4.2) that attempts to address some of the above concerns. The contribution of this work is mainly in fusing facial expressions and emotional dialogues to classify emotions of the movie characters into positive or negative. A novel weighting methodology is presented to perform this fusion. Another contribution is in the form of an improved algorithm for finding emotions from movie dialogs. As far as the two class prediction is concerned, it can be seen as a higher level abstraction which gives some insight into a more challenging problem with many more output emotion categories. The two class case also gives useful statistics for movie

recommendation purposes.

Fusion of modalities makes up for the deficiency in any one of the two modalities. A novel methodology is proposed to determine weights given to each modality while fusing them. An emotion understanding algorithm is also proposed especially suited to movie dialogues. The facial expression recognition technique specifically addresses the difficult problem of facial expression recognition under pose variations to a great extent, although we do report failures under very difficult scenarios. To our knowledge the proposed bimodal framework and the fusion approach is novel and as experimentally shown later, it gives promising results compared to a unimodal approach or combining visual with acoustic clues.

Application of the bimodal ER approach for Multi-actor ER has been presented in section 4.2.

4.1.1 Facial Expression Recognition

Given the input image sequence \mathbb{F} , with the i^{th} frame denoted by F_i , faces are detected using the Viola-Jones multi-view detector [92] and 23 Facial Feature Points (FFPs) are marked manually in the first frame. A frame with the neutral face of each of the characters in the video sequence is manually selected from the movie and FFPs are marked on it. These FFPs are tracked in the sequence using a PPCA based algorithm [124]. Speaker detection is performed and for the speaker (Section 4.1.5), displacements of FFPs from neutral to expression are used as features. These displacements are also referred to as ‘residues’. Most relevant features are selected using the significance ratio test [197] and are used for FER in each frame using SVM classifier. The output of the FER algorithm is the probability of each frame having a positive expression. The expression to which majority of frames belong to, is the predicted expression

for that image sequence. Prediction about the expression in the sequence is not needed for fusion but is required for evaluating the performance of the FER algorithm. The pseudocode for the algorithm is as follows:

Pseudocode of the FER algorithm:

Inputs: Training and test image sequences with a neutral frame (Denoted by F_{ne}) corresponding to each sequence.

Output: $P_v^i(+/F_i) \forall i \in [1, N_f]$: Probability of each frame having positive expression.

Training: *for all training sequences*

- Initialize the training feature matrix, $X_{tr}=[]$
- Detect faces in F_1 using the Viola Jones face detector [92].
- *for each face*
 1. Mark 23 FFPs in F_1 and F_{ne} . Let D_{ne} be the 23×2 matrix storing (x, y) coordinates of 23 FFPs for neutral.
 2. Track FFPs using a PPCA based algorithm [124]. Let D_i store (x, y) coordinates of FFPs for i^{th} frame.
 3. In case of more than one face detected, perform speaker detection.
- *for the speaker*
 1. Find residue R_i , $R_i = D_i - D_{ne}$, normalize them and shift origin to nose tip.
 2. Let residues for i^{th} frame be R_{Si} (23×2 matrix). Form feature

matrix X and concatenate it to X_{tr} as follows:

$$X = \begin{bmatrix} R_{S1x}^T & R_{S1y}^T \\ R_{S2x}^T & R_{S2y}^T \\ \dots & \dots \\ R_{SNfx}^T & R_{SNfy}^T \end{bmatrix}, \quad X_{tr} = \begin{bmatrix} X_{tr} \\ X \end{bmatrix} \quad (4.1)$$

where subscripts x and y refer to residues of x and y coordinates, respectively.

3. Train SVM classifiers using X_{tr} .

Testing:

- For each face in the test sequence, find X_{tes} as X was found in steps 1 to 2 above.
- Select relevant features using significance ratio test [197].
- Use SVM for classification of each frame into having positive or negative expression and find the probability estimates [205].

Details of the algorithm are given in section 4.1.5.

4.1.2 Lexical Analysis

Lexical analysis of dialogues using *Semantic Orientation* (SO) approach [185] can give a clue whether speaker's expression is positive or negative. Positive SO of a phrase indicates a positive emotion; e.g. 'Very sweet!' has SO = +1.68 while the phrase 'Shameless woman!' has SO = -1.37. SOs of some of the dialogues are listed in Table 4.1.

First step in determining SO, is to extract emotional words or phrases. Out of the different parts of speech; adjectives, adverbs and verbs can be used as

Dialogues	Our method	[185]
You shameless woman!	-1.37	NaN
I swear I won't tell anyone.	-1.98	-1.57
You punish them. You clamp them.	-1.17	NaN
Acting like a gangster doesn't scare me.	-0.82	-0.41
What is going on here?	-0.02	NaN
You have very soft hands.	+1.84	+0.15
Dave, that was amazing!	+1.45	NaN
I'll be close by, okay?	+0.14	NaN
That's great.	+1.08	NaN
That's very sweet.	+1.68	+0.62
Failures in calculating SO	15%	71.9%
Recognition accuracy	86.7%	23.2%

Table 4.1: Comparison of proposed method and method in [185] for ER using Semantic Orientation (SO). SO value NaN shows failure in extracting emotional words.

emotional words [185]. E.g. in the phrase ‘*The beautiful girl danced jubilantly*’, the words *beautiful*, *danced* and *jubilantly* are adjective, verb and adverb respectively and indicate a positive mood. For selecting adjectives, adverbs and verbs from a dialogue, the Stanford Part Of Speech (POS) tagger [180] [179] is applied on the dialogues.

From the relevant words, semantic orientation of a dialogue is calculated on the lines of the algorithm proposed in [185]. Original algorithm [185] involves extracting relevant word-pairs based on some set of rules given in Table 4.2. On applying these rules on movie dialogues, it was observed that most of the time, none of the word pairs satisfied any of these rules. This was because of the difference in the structure of dialogues and sentences in a formal review.

Turney’s algorithm was proposed for classifying reviews and works well in that domain. However, construction of sentences in a review is different from that in a dialogue. A review is well constructed using formal style of writing. On the other hand, dialogues are mostly short sentences without much grammatical consideration. For the dialogues in our dataset, the average number of words per dialogue is only around 7 words. Due to this, many word-pairs

Original [185]			
S.No.	First Word	Second Word	Third word
1	JJ	NN or NNS	anything
2	RB, RBR or RBS	JJ	neither NN nor NNS
3	JJ	JJ	neither NN nor NNS
4	NN or NNS	JJ	neither NN nor NNS
5	RB, RBR or RBS	VB, VBD, VBN or VBG	anything
Additional rules			
6	VB, VBD,VBN or VBG	JJ	anything
7	JJ, RB, RBR,RBS, VB, VBD,VBN or VBG	anything	anything

Table 4.2: Patterns of tags for extracting two-word or single word phrases. JJ: Noun; NN: Noun, singular or mass; NNS: Noun, plural; RB: Adverb; RBR: Adverb, comparative; RBS: Adverb, superlative; VB: Verb; VBD: Verb, past tense; VBN: Verb, past participle; VBG: Verb, gerund or present participle. Both original rules and rules added by us, are shown.

could not match the original patterns given in Table 4.2. E.g. Consider the phrase ‘*That’s nice*’ which indicates a positive emotion. POS tagger output will be *That_DT ’s_VBZ nice_JJ* (See the caption of Table 4.2 for an explanation of the notations used). No word-pairs match the patterns. Considering this difference, new rules were added which are listed in the last two rows of Table 4.2 . The restriction of using a pair of words was removed considering the brevity of the dialogues. According the additional rules, the words ‘*s*’ (is) and *nice* will be extracted as a phrase and checked for SO.

Semantic Orientation (SO) of the chosen phrases or words is calculated using PMI-IR algorithm [185]. The basic idea behind the PMI-IR algorithm is that a word with negative connotation is more likely to co-occur with a negatively connotated word rather than a positively connotated word. Co-occurrence for a query word or phrase is calculated using internet search engines.

The Pointwise Mutual Information (PMI) between two words w_1 and w_2 is given by:

$$PMI(w_1, w_2) = \log_2 \left[\frac{p(w_1 \cap w_2)}{p(w_1)p(w_2)} \right] \quad (4.2)$$

Here $p(w_1)$ and $p(w_2)$ are the probabilities of occurrence of w_1 and w_2 respectively and $p(w_1 \cap w_2)$ is the probability of co-occurrence of w_1 and w_2 . w_1 and w_2 are defined to co-occur if in a document they occur within ten words of one another. If w_1 and w_2 are statistically independent, $p(w_1 \cap w_2) = p(w_1) \times p(w_2)$. In this case $PMI(w_1, w_2) = 0$. Thus, $PMI(w_1, w_2)$ indicates the statistical dependence between w_1 and w_2 or in other words, $PMI(w_1, w_2)$ gives an information about presence of one word if we see the other word.

SO of a phrase or a word Ph , is calculated as follows [185]:

$$SO(Ph) = PMI(Ph, E) - PMI(Ph, P) \quad (4.3)$$

where E and P refer to ‘excellent’ and ‘poor’, respectively. To calculate $PMI(Ph, E)$, the PMI-IR algorithm searches the query ‘Ph NEAR excellent’ in a search engine and records the total number of matching documents (or hits). The search engine that is used in this work is Exalead search engine¹ which has the NEAR operator. NEAR operator searches the simultaneous occurrence of two words within a space of ten words.

SO of a phrase is given by :

$$SO(Ph) = \log_2 \left[\frac{\text{hits}(Ph \text{ NEAR } E)\text{hits}(P)}{\text{hits}(Ph \text{ NEAR } P)\text{hits}(E)} \right] \quad (4.4)$$

where $\text{hits}(X)$ represents the number of hits returned when X is queried in the search engine.

¹<http://www.exalead.com/search/>

4.1.3 Fusing Clues from FER and Lexical Analysis

FER gives the probability of each frame of the image sequence having a positive expression (say $P_v^i(+/F_i)$). SO is the clue from Lexical Analysis.

Fusion is achieved in a weighted fashion using two different approaches: *linear opinion pool* approach and *logarithmic opinion pool* approach [117]. The combined probability of a video frame having positive emotion is given as:

$$P_i(+/F_i, \mathbb{D}) = w_v^i P_v^i(+/F_i) + w_s^i P_s(+/\mathbb{D}) \quad \text{Linear opinion pool} \quad (4.5)$$

$$= P_v^i(+/F_i)^{w_v^i} P_s(+/\mathbb{D})^{w_s^i} \quad \text{Logarithmic opinion pool} \quad (4.6)$$

where $P_s(+/\mathbb{D})$ is the probability of the video having positive emotion as determined by lexical clues. F_i and \mathbb{D} represent the i^{th} video frames and dialogue, respectively. Note that for all the frames, the dialogue is the same. w_v^i and w_s^i are the corresponding weights for visual and lexical clues which add up to one. The probability of the video sequence \mathbb{F} , having positive emotions is given as the mean of probability values over all frames; i.e.

$$P(+/\mathbb{F}, \mathbb{D}) = \frac{\sum_i P_i(+/F_i, \mathbb{D})}{N_f} \quad (4.7)$$

Finding Probabilities

The value of SO does not directly gives the probability of the dialogue being positive or negative. However, magnitude of SO gives a measure of the probability of the dialogue being positive or negative. E.g. A dialogue with an SO value of +2.5 is more probable to be positive as compared to another dialogue with SO value of +1.0. In order to deduce probability values from SO, the most positive and most negative SO values were found out from the training data.

Let these values be denoted by SO_+ and SO_- . A dialogue with positive SO equal to or greater than SO_+ in magnitude, is given the probability of being positive as 1. In general, the probability of a dialogue with SO value S being positive is given by:

$$P_s(+/\mathbb{D}) = \quad S/SO_+ \quad \text{if } S \geq 0 \quad (4.8)$$

$$= 1 - S/SO_- \quad \text{if } S < 0 \quad (4.9)$$

Finding Weights

Weights given to the individual clues are proportional to the confidence on that clue. For lexical clue, a high SO value indicates a greater confidence level. Therefore, the weight for lexical clues is chosen as:

$$w_s^i = \frac{\alpha_s}{\frac{\alpha_v^i}{\alpha_{max}} + \alpha_s}, \quad \text{where } \alpha_s = \quad P_s(+/\mathbb{D}) \quad \text{if } S \geq 0 \quad (4.10)$$

$$= 1 - P_s(+/\mathbb{D}) \quad \text{if } S < 0 \quad (4.11)$$

and α_v^i and α_{max} are defined later in eq. 4.13.

To calculate weights for visual clues, factors were identified which can affect the performance of ER from visual clues. These factors are pose and scale of the face, intensity of expression and head motion. For each of these factors an associated parameter was extracted from frame i of an image sequence as follows:

- **Pose (associated parameters are p_1^i and p_2^i):** Let (x_j^i, y_j^i) be the coordinates of j^{th} FFP in i^{th} frame with the origin at the nose tip. FFPs are numbered as per Figure 4.1. One of the parameters to estimate pose is taken as the ratio of horizontal distance between points 3 and 4 to the horizontal distance between points 3 and 9. If this ratio is less than 1, inverse of this ratio is considered. Second parameter is the slope of line

connecting points 4 and 9. Mathematically these parameters are given as:

$$p_1^i = \max \left(\frac{x_3^i - x_4^i}{x_9^i - x_3^i}, \frac{x_9^i - x_3^i}{x_3^i - x_4^i} \right) \quad p_2^i = \frac{y_9^i - y_4^i}{x_9^i - x_4^i} \quad (4.12)$$

For a frontal face p_1^i should be almost 1 and any deviation from 1 indicates rotation of the face in yaw direction. p_2^i for a frontal face is zero and deviation from zero shows rotation of the face in roll direction.

- **Scale (p_3^i):** Normalization distance is taken as a measure of the scale of the face and is represented by the parameter p_3^i . Section 4.1.5 defines the normalization distance. This distance should be large for a good recognition performance.
- **Intensity of the Expression (p_4^i):** An expression will be more intense if the normalized residues will be more in magnitude. For a frame, intensity is given by the sum of the magnitudes of residues for that frame, normalized by the scale of the face. This parameter is represented by p_4^i . Higher value of p_4^i promises a better recognition.
- **Head Motion (p_5^i):** Head motion is taken to be captured mostly by nose motion since there is minimal non-rigid motion for the point on the nose tip. Displacement of the nose tip from its position in the first frame is taken to be a measure of head motion for that frame. Let this parameter be p_5^i . For a good recognition performance, p_5^i should be low.

The weight given to the visual clue for a frame i is calculated as

$$w_v^i = \frac{\frac{\alpha_v^i}{\alpha_{max}}}{\left(\frac{\alpha_v^i}{\alpha_{max}} + \alpha_s\right)}, \quad \alpha_v^i = \frac{p_3^i p_4^i}{(1 + a_1^i |p_1^i - 1|)(1 + a_2^i |p_2^i|)(1 + a_3^i p_5^i)} \quad (4.13)$$

and α_{max} is the maximum value of α_v^i for the training data. Parameters p_1^i to p_5^i with their physical meanings are defined above. a_1^i, a_2^i and a_3^i are

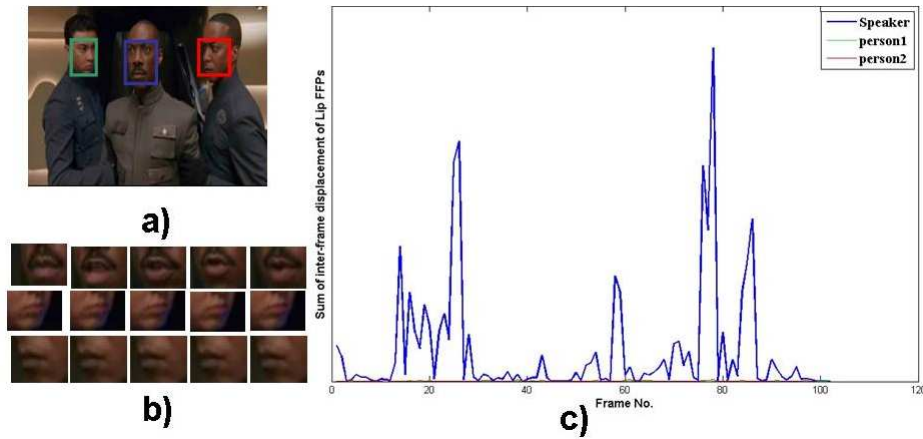



Figure 4.3: Speaker detection using motion of FFPs on lips. a) First frame of the sequence b) Lip images for frame nos. 1, 16, 31, 54 and 70. Top row is for the character in the middle, 2nd and 3rd row for the characters on left and right respectively. Each column corresponds to one frame. c) Graph of sum of magnitudes of inter-frame displacements of lip FFPs for the three characters. Magnitude of gradient for the person in the middle (blue curve, which is the only clearly visible curve) is very large as compared to curves for other two persons (green and red curves, barely visible).

chosen manually from the training data. For example, keeping a_2^i and a_3^i fixed, the value of a_1^i was tuned such that the variation of p_1^i is found as desired (mentioned above). a_2^i and a_3^i are also obtained similarly.

4.1.4 Dataset

To, the best of our knowledge, there is no publicly available dataset of movie scenes along with the dialogs. Consequently, we had to develop our own dataset, the remarkable characteristic of the which, is the visual difficulty level of the scenes. Some of those variations are depicted in Figure 4.1 b-e, Figure 4.5, and Table 4.3. Experimental data consists of 700 movie clips collected from 17 movies belonging to 5 genres viz. Comedy(6), Action(2), Adventure (3), Drama(1), Horror(5). The number in bracket is the number of movies for each genre. Each movie clip is accompanied by the corresponding dialog. There were 350 clips belonging to each of the positive and negative classes.

Table 4.3: An instance showing the effectiveness of fusion. Lexical weight is high due to high SO magnitude. Lower weight is assigned to FER result due to small size of the face, and low facial expression intensity. Upon fusion, the correct prediction of LA dominates the incorrect prediction by FER.

Sample Frame	
FER	Negative
Visual weight	0.21
Text	Captain! Thank goodness you're back.
LA	Positive(SO=+0.32)
Lexical weight	0.79
FER+LA	Positive

Positive class refers to emotions such as happiness, jubilation, etc. while negative class refers to emotions such as fear, sadness, disgust, etc. In order to extract the relevant clips, movies were shown to 20 volunteers and they marked the segments of the movie where emotion was expressed either through facial expressions or through the dialogs. The clips were manually labeled with one of two classes by the volunteers. Marked segments were then extracted out of the movie into short clips of around 4-5 seconds duration. Dialogs for individual clips were extracted from the movie subtitles which contain the spoken dialogs along with the timings when they were spoken. Out of the 350 clips used for each class, 250 were used for training and 100 for testing.

4.1.5 Experiments

Speaker Identification

Identifying the speaker amongst all the characters in the scene is done by lip motion analysis. The approach has found to be successful in [68]. The charac-

ter for which the sum of magnitudes of inter-frame displacements of lip FFPs has the maximum mean value is taken to be the speaker. Lip FFPs are manually marked in the first frame of the sequence and tracked thereafter (4.1.1). Figure 4.3 demonstrates an instance of speaker identification. Note that the actual speaker has significantly higher motion profile than the rest of the characters.

Experiments Using Visual Clues Alone

Training Experiments were conducted to measure the stand alone efficacy of the FER algorithm without using the lexical clue. For training the system, positions of FFPs in the neutral is required. A frame containing the neutral face is selected from the image sequence. If the sequence does not contain any neutral face, selection is made from other parts of the movie. The pose of the neutral face should be as close as possible to the expressive face. To avoid errors due to variation in scale of the faces, residues are normalized using the Euclidean distance between the point midway between FFPs 1 and 2 (defined as FFP no. 0), and the nose tip. The normalization distance is given by $D_{norm} = d(0, 3)$ where $d(i, j)$ denotes the Euclidean distance between FFPs i and j . This distance is robust towards pose variations in yaw direction, which are found to be more prominent. Normalized residues for all the FFPs might not be relevant and thus feature selection is done on the training data using the significance ratio test.

Figure 4.1 shows 5 most relevant features selected, marked with a red color (FFPs No. 5, 10, 11, 13 and 15). An SVM classifier with an RBF kernel is trained with the selected features (Sec. 4.1.1). The parameters for the RBF kernel are $C = 4096$ and $\gamma = 0.125$ (found using grid search).

Testing A test clip is preprocessed and features are extracted in the same manner as for training sequences. SVM classification is done using selected

Table 4.4: Effectiveness of fusing visual and lexical clues. Confusion matrices for classification a) Using visual features alone, b) Using lexical analysis alone, c) By fusion using linear opinion pool (Lin) approach, and d) By fusion using logarithmic opinion pool (Log) approach. ARR: Average Recognition Rate. For each class, 250 clips: Training; 100 clips: Testing.

(a) Visual alone (ARR=76.7%)	(b) lexical alone (ARR=86.7%)	(c) Fusion (Lin) (ARR=92.7%)																											
<table border="1" style="border-collapse: collapse; width: 100%;"> <thead> <tr> <th></th> <th>Pos</th> <th>Neg</th> </tr> </thead> <tbody> <tr> <th>Pos</th> <td>73.3</td> <td>26.7</td> </tr> <tr> <th>Neg</th> <td>20.0</td> <td>80.0</td> </tr> </tbody> </table>		Pos	Neg	Pos	73.3	26.7	Neg	20.0	80.0	<table border="1" style="border-collapse: collapse; width: 100%;"> <thead> <tr> <th></th> <th>Pos</th> <th>Neg</th> </tr> </thead> <tbody> <tr> <th>Pos</th> <td>86.7</td> <td>13.3</td> </tr> <tr> <th>Neg</th> <td>13.3</td> <td>86.7</td> </tr> </tbody> </table>		Pos	Neg	Pos	86.7	13.3	Neg	13.3	86.7	<table border="1" style="border-collapse: collapse; width: 100%;"> <thead> <tr> <th></th> <th>Pos</th> <th>Neg</th> </tr> </thead> <tbody> <tr> <th>Pos</th> <td>90.7</td> <td>9.3</td> </tr> <tr> <th>Neg</th> <td>5.3</td> <td>94.7</td> </tr> </tbody> </table>		Pos	Neg	Pos	90.7	9.3	Neg	5.3	94.7
	Pos	Neg																											
Pos	73.3	26.7																											
Neg	20.0	80.0																											
	Pos	Neg																											
Pos	86.7	13.3																											
Neg	13.3	86.7																											
	Pos	Neg																											
Pos	90.7	9.3																											
Neg	5.3	94.7																											
(d) Fusion (Log) (ARR=89.7%)																													
<table border="1" style="border-collapse: collapse; width: 100%;"> <thead> <tr> <th></th> <th>Pos</th> <th>Neg</th> </tr> </thead> <tbody> <tr> <th>Pos</th> <td>89.3</td> <td>10.7</td> </tr> <tr> <th>Neg</th> <td>10.0</td> <td>90.0</td> </tr> </tbody> </table>				Pos	Neg	Pos	89.3	10.7	Neg	10.0	90.0																		
	Pos	Neg																											
Pos	89.3	10.7																											
Neg	10.0	90.0																											

features. Results of FER experiments, averaged over 10 runs are given in the form of a confusion matrix in Table 4.4 (a). In a confusion matrix, an element (i, j) (i.e. row corresponding to class i and column to class j) shows the number of test samples which were predicted to belong to class j when the sample actually belongs to class i . Each run had the training and testing sets randomly selected.

Upon experimentation using only visual clues, failures were encountered at different stages of the algorithm. In Figure 4.4a. cartoon faces were detected as faces. Consider a scene in which a person has such faces on his T-shirt. Deformation of his T-shirt might give an impression of the cartoon face moving leading to a wrong conclusion about the facial expression. Tracking the FFPs was affected by large facial motion or due to occlusion. In Figure 4.4b., the handphone confused the tracker while in Figure 4.4c., lip motion caused tracking failure.

Accuracy using just the visual clues was not very good even for the two class case. This is considering the complexity of human emotions in a natural

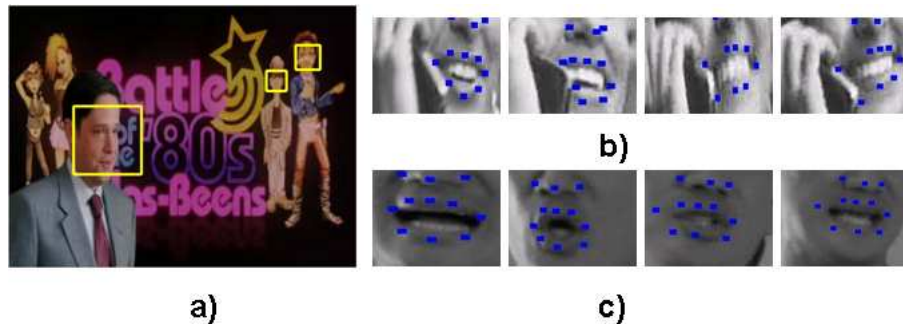


Figure 4.4: Examples of failures using visual clues. a) Failure in face detection. Recognizing cartoons as faces can be misleading if there is a motion of the surface on which cartoons are painted. b) Failure in FFP tracking due to lips being occluded by the handphone. Frame numbers 1, 10, 20 and 28 are shown b) Tracking failure due to fast motion of lips and head while speaking. Frame numbers 1, 15, 30 and 45 are shown.

environment. Results would have been worse if more classes of facial expressions would have been considered.

Experiments Using Lexical Clues

For the total 700 movie clips used in the experiment, there were 90 instances where semantic orientation could not be calculated because dialogues were very short. As a result, no word or word-pair satisfied the rules for being considered as relevant (Table 4.2). In such cases, lexical clues were given zero weight in the fusion process. From the SOs of some of the dialogues in Table 4.1, it can be seen that SO gives a good indication of emotion in a dialogue. Results for classification of dialogues into having positive or negative emotions is given in Table 4.4 (b). The result is better than the visual clues probably because of the variations of pose, scale, expression intensity, rigid motion etc. in the image sequences.

There were cases when SO could not be found out using the PMI-IR algorithm especially in the case of using word-pairs. This was because there was no any instance found on the internet where the two words forming the word pair co-occurred, e.g. the phrase “Gargantuan Beasts”.

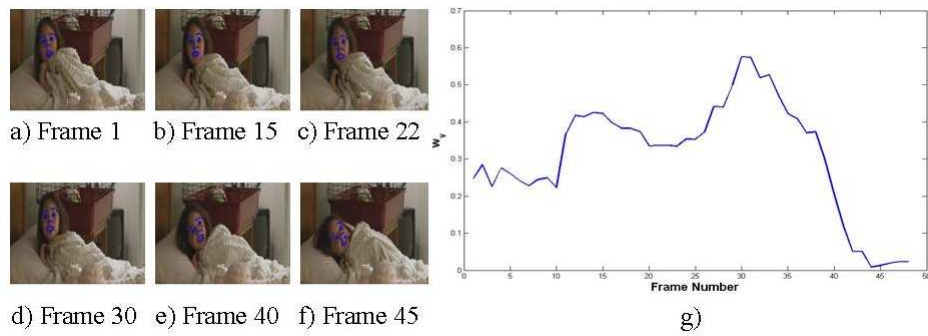


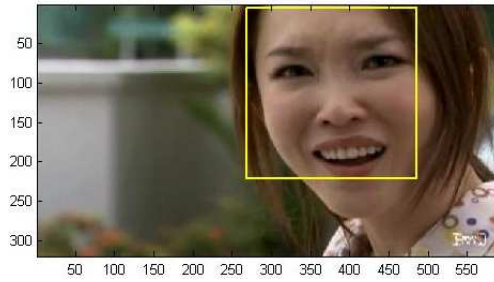
Figure 4.5: Variation of visual weight with head pose. Observe the sudden reduction of weight at around frame 40 as the face tilts to the left. Weights are computed as described in section 4.1.3.

Sometimes a particular word is used in different connotations. E.g. the word *Sissy* can be used either to address one’s sister or in a pejorative way to address a cowardly person. SO of -0.15 for ‘sissy’ indicates a negative connotation ignoring its positive use. Another example is the phrase “Business must be good!” with an SO value of +1.09. If the same message was conveyed by saying “Business might never be this good!”, the SO value changes to -0.43 even when the phrase is negatively oriented.

Experiments Combining Visual and Lexical Clues

The results of fusion are shown in Table 4.4(c) and (d) using *linear opinion pool* approach and *logarithmic opinion pool* approach, respectively. These results are average result over 10 runs and in each run training and test sets for visual experiments were randomly chosen. It is observed that *linear opinion pool* performs better than *logarithmic opinion pool*. There is an improvement in the classification accuracy when the lexical information was also used along with the visual information.

Fusion of the clues depends a lot on the weightage given to each clue. A variation of the visual weights with the progression of a test sequence is given in Figure 4.5. It can be seen that with the head tilting, visual weight suddenly



(a)

Text	<i>You are not following the rules!</i>
FER	Positive
LA	Negative (SO=-0.68)
FER+LA	Positive

(b)

Figure 4.6: An example of failure of the proposed algorithm due to contradictory information from visual and lexical clues. Actress seems to be smiling while saying a negative statement. Top row shows a sample frame from the clip and bottom row shows the text LA: Lexical Analysis.

decreased. With such choice of weight, possibility of errors due to variations in visual clues can be reduced. It was identified in some cases that either of the two modalities were insufficient for predicting the emotion however fusing the two modalities gave correct result (See Table 4.3).

Fusing the two clues also gave errors in few cases where ground truth information from visual clues was opposing that from the lexical clues (Figure 4.6). This gave an unexpected result.

Comparison with Fusing Visual and Acoustic Clues

Since most of the multimodal approaches for ER use acoustic features in conjunction with visual features, the approach using acoustic features was compared with the proposed approach. In combining visual and acoustic clues, features were chosen at clip level similar to the work in [30]. Visual features are the number of frames classified by the FER algorithm (Sec. 4.1.1) into each of the two classes. This leads to a two dimensional visual feature vec-

Table 4.5: a) and b) show confusion matrices for classification results using only acoustic (AC) features, and by fusing visual and acoustic features (FER+AC), respectively. c) Comparison between performance (Avg. Recognition Rate in percentage) of Lexical Analysis (LA) and acoustic features (AC) and between two fusion approaches. For each class, 250 clips: Training; 100 clips: Testing.

a. Acoustic alone

	Pos	Neg
Pos	73.0	27.0
Neg	19.7	80.3

b. Fusion

	Pos	Neg
Pos	73.3	26.7
Neg	13.3	86.7

c. FER+LA vs. FER+AC

Method	ARR
LA	86.7%
AC	76.7%
FER+LA	92.7%
FER+AC	80.0%

tor. Acoustic features are same as in [30]. Fusion was done using the product combining criterion described in [30] as it gave the best performance for them. Results of using acoustic features alone and fusing them with visual features are given in Table 4.5 a. and b. Lexical features (LA) were found to perform better than acoustic information (AC). Also fusion results are poorer in combining visual with acoustic clues (FER+AC) than those by fusing lexical clues with FER clues (FER+LA). This is because of movies being closer to natural environment than lab settings. Note that the approach in [30] was for ER in a controlled setting. Comparison between two multimodal fusion modalities is summarized in Table 4.5c.

4.2 Multi-actor ER vs. Single Actor ER

Considering the applications of ER in movies, we might often be more interested in emotions conveyed by the scene as a whole and not just by a single actor in the scene. In movies, there are often multiple actors in a scene. Current approaches for ER recognize emotions of a single subject (single actor approach). But emotion of just one actor may not reflect the emotion of the scene. The main contribution of the proposed work is to analyze *multi-actor*

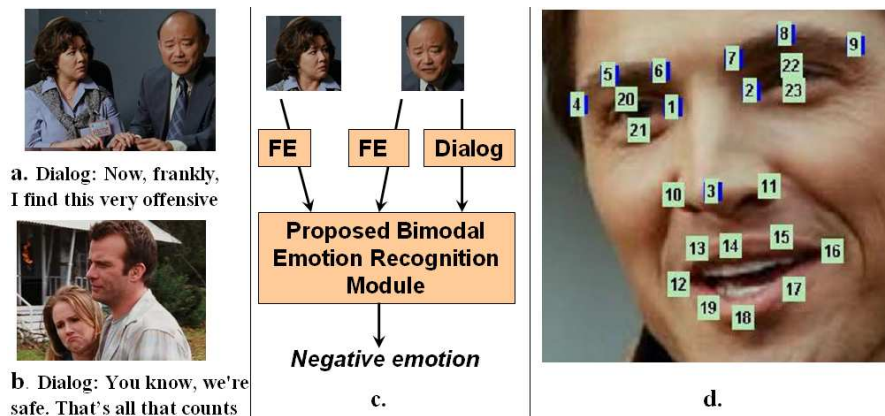


Figure 4.7: Fusing emotional clues from multiple actors. Consider two scenes with negative emotions and male actors being the speaker. **a.** Non-speaker’s neutral expression might not be helpful. **b.** Note that 1) Speaker’s positive emotions are insufficient in indicating about the tragedy in the scene, 2) Emotion of actress with near frontal face may be more reliable than those of the actor with profile view face. **c.** Proposed bimodal multi-actor emotion recognition framework. **d.** 23 Facial Feature Points (FFPs) used for FER.

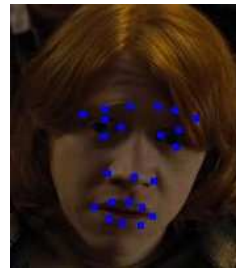
emotion recognition i.e. recognizing emotions in movie clips by fusing emotional clues from different actors.

There is a need to analyze the effect of recognizing emotions of all the relevant actors in a movie scene (multi-actor approach). Such an analysis can be interesting considering the different possibilities that may arise in a movie scenario. For example, Figure 4.7b. depicts a scene where an actor consoles the actress after a tragedy. If we consider clues from both the dialog and the facial expression, the speaker (male actor) has a positive emotion while the other actor² has a sad expression. The scene represents a sorrowful situation which is indicated by the non-speaker. In this case, combining emotional information from a non-speaker will be of help. Consider another case in Figure 4.7a. where the speaker (male) has a negative expression but the non-speaker has an almost neutral expression. If the emotion clue from the non-speaker is fused with that of the speaker, the probability of predicted emotion being negative reduces. So, it is a debatable issue whether emotion clues of actors other than

²For convenience, females are also referred to as actors; in this part of the work



a. Profile face:
 $pose1 = 0.04, pose2 = 22.8$



b. Non-profile face:
 $pose1 = 0.67, pose2 = 2.3$

Figure 4.8: Dealing with pose variations using two pose parameters

speaker are relevant for recognizing emotion of the scene, or not.

Amongst other challenges, predicted emotion for each actor might not be equally reliable as in Figure 4.7b. where the prediction of ER algorithm for the actor with a frontal face may be more reliable than that for the non-frontal face. The reliability of emotion of each actor needs to be determined. In the cases discussed above, we have emotional clues from facial expressions (visual clue) of two actors and dialog (lexical clue) of the speaker. It is not easy to devise a method to determine relevance of each of these clues and fuse them automatically to determine emotion of the scene.

4.2.1 Facial Expression Recognition (FER)

Note that facial expressions are one of the clues for internal emotions, other being the spoken words (in the context of this work). The algorithm used for recognizing facial expressions of each actor is same as that presented in section 4.1.1. In addition, variations in face pose have been dealt with in the following manner:

Dealing With Large Variations in Facial Pose

In order to deal with the large range of facial poses encountered in the data, two SVM classifiers were trained, one for profile and other for non-profile views.

In the testing phase, the test clip is classified into non-profile or profile and then the corresponding classifier is used for classifying that clip. The SVM classifier trained on profile faces is used to test profile clips and the classifier trained on non-profile faces is used to test non-profile clips. In order to classify facial pose in a particular clip into profile or non-profile, two parameters were used which are defined as follows:

1. *pose1*: If d_{mn} denotes the euclidean distance between FFPs number m and n , *pose1* is defined as:

$$pose1 = \frac{d_{49}}{(d_{34} + d_{39})} \quad (4.14)$$

For a frontal face this distance should be around 0.5 while for a profile face, due to a small value of d_{49} , value of *pose1* is close to zero. Here d_{ij} refers to euclidean distance between FFPs number i and j , as numbered in Fig.4.7d.

2. *pose2*: *pose2* measures the nature of the spread of the FFPs in the xy plane. If we perform PCA on the position of FFPs, since FFPs for a frontal face are evenly spread along the principal directions, the two eigenvalues are of similar order. However, for a profile face spread of FFPs is not the same in both the principal directions. This makes the first eigenvalue to be much larger than the second one. *pose2* is defined as $pose2 = \lambda_1/\lambda_2$, where λ_1 and λ_2 are the first and second eigenvalues respectively after applying PCA on the FFP coordinates. The value of *pose2* is much larger for profile faces as compared to frontal faces.

For classification of facial pose, the first frame of the clip is considered. If *pose1* is less than a certain threshold and *pose2* is greater the another threshold, for the first frame, the face pose in the clip is classified as profile otherwise it is

classified as non-profile. The thresholds are manually chosen based on training data. Fig.4.8 shows values of $pose1$ and $pose2$ for sample clips.

4.2.2 Lexical Analysis of Dialogs

Lexical analysis of the dialogs is performed using Semantic Orientation approach mentioned in section 4.1.2. However, SO is calculated using Whissell's Dictionary of Affect in Language (DAL)[198] since the PMI-IR algorithm was unsuccessful to find SO in many cases as mentioned in section 4.1.5.

DAL gives the emotional connotation of 8742 words along three dimensions viz. evaluation, activation and imagery. Scores for pleasantness range from 1 (unpleasant) to 3 (pleasant), for activation range from 1 (passive) to 3 (active) and for imagery range from 1 (difficult to form a mental picture of this word) to 3 (easy to form a mental picture). To calculate SO, DAL scores are mapped to a range of -1 to +1. For each word in the dialog, evaluation is taken as a direct measure of its SO. If a word is missing in the dictionary, evaluation score of its synonyms or related forms are used. Since very few words were missing in DAL, synonyms and related forms could be manually entered. SO for entire dialog is given by a mean SO value of individual words. The probability of a dialogue with SO value S being positive is given by:

$$P_{s_L}^{pos} = S/SO_+ \quad \text{if } S \geq 0 \quad (4.15)$$

$$= 1 - S/SO_- \quad \text{if } S < 0 \quad (4.16)$$

where SO_+ and SO_- are the maximum and minimum SO values for training examples and s refers to the speaker.

4.2.3 Fusing Visual and Lexical Cues

For fusing visual with lexical cues, identifying the speaker amongst all the actors in the scene is important since lexical cue from dialog can be combined with the visual cue only from the speaker's face and not those of other actors. In some cases, instead of the speaker, audience is seen in the video which makes it necessary to identify whether speaker is present or not. Speaker detection is performed by lip motion analysis (section 4.2.4).

Visual and lexical cues are fused in a weighted manner. The weights are determined using the approach in section 4.1.3 with some improvements:

Finding weights

Weights for the lexical cues are calculated in the same manner as before (in section 4.1.3).

To calculate weights for visual cues, in addition to the factors identified before, tracking failure has also been considered as one of the factors affecting the performance of emotion recognition from visual cues. In this section only the improvements to the previous weighting approach will be mentioned. Moreover, the parameter for scale has been modified as mentioned in Section 4.2.4.

The parameter related to tracking accuracy has been introduced. Any tracking failure is expected to reduce FER performance. To detect tracking failure, an assumption is made that motions of pairs of some adjacent points on the face will be almost similar, when facial expression changes. E.g. when a person lifts his eyebrow, FFPs 4 and 5 (Fig.4.7d.) usually move upwards with similar displacement magnitudes. As long as both the FFPs belonging to a chosen pair are correctly tracked, their inter-frame displacements will be similar. Any tracking failure is expected to displace the wrongly tracked point more as compared to the correctly tracked point.

Differences of inter-frame displacements of the two points gives an indication of the tracking failure. These differences are summed up for all the chosen pairs of FFPs. The chosen pairs are (4,5), (5,6), (7,8), (8,9), (10,11), (12,13) and (15,16). This summation (say D) should be under a certain threshold (say D_{min}) for the tracking to be reasonably good. In case of a tracking failure, due to wrongly predicted position of an FFP, D would cross D_{min} . D_{min} is chosen manually from training data by analyzing values of D for both correct and incorrect tracking. The parameter associated with tracking accuracy is given by $p_3^f = 0$ when $D > D_{min}$, otherwise $p_3^f = 1$.

The other parameters are same as in section 4.1.3 but notations have been changed. p_1^f and p_2^f continue to relate to face pose while parameters related to scale, intensity of expression and head motion for the f^{th} frame are represented by p_4^f , p_5^f and p_6^f , respectively.

Weight given to the visual cue for a frame f is calculated as

$$w_v^f = \frac{\frac{\alpha_v^f}{\alpha_{max}}}{\left(\frac{\alpha_v^f}{\alpha_{max}} + \alpha_s\right)}, \alpha_v^f = \frac{p_3^f p_4^f p_5^f}{(1 + a_1^f |p_1^f - 1|)(1 + a_2^f |p_2^f|)(1 + a_3^f p_6^f)} \quad (4.17)$$

and α_{max} is the maximum value of α_v^f for the training data. a_1^f , a_2^f and a_3^f are chosen manually as mentioned in section 4.1.3. Note that Eq. 4.17 additionally considers effect of tracking failure (Parameter p_3^f) which wasn't the case with Eq. 4.13.

Once weights to individual cues are determined, fusion of $Pi_V^{pos}(f)$ and P_S^{pos} can be achieved in two ways:

Scheme 1:

1. Fuse $P_S^{pos}(f)$ with P_S^{pos} using weights ws_s and ws_v^f , respectively. Let $P_S^{pos}(f)$ be the resultant probability for f^{th} frame. Here ws is used to

indicate weight for the speaker ('s').

2. Combine $P_{s^{pos}}(f)$ with $P_{i_V^{pos}}(f)$ in a weighted fashion; where i denotes all actors except the speaker. Weight for $P_{s^{pos}}(f)$ is mean of w_s and $w_{s_v^f}$, while weights for $P_{i_V^{pos}}(f)$ are $w_{i_v^f}$ for different values of i . Mean probability over all frames is the probability of the clip having positive emotion.

Scheme 2:

Combine $P_{i_V^{pos}}(f)$ for all actors including the speaker (for all i s) with weights given by $w_{i_v^f}$, then fuse the resulting visual probability with $P_{s_L^{pos}}$. Weight for $P_{s_L^{pos}}$ is given by w_s . Again, probability of the clip having positive emotion is given by mean probability over all frames.

4.2.4 Experimental Results and Discussions

Dataset

The data for the experiments consists of 388 movie clips from 17 movies classified into 5 genres-Comedy(6), Action(2), Adventure (3), Drama(1), Horror(5). 232 clips contain negative emotions while 156 contain positive. Each of the clips contains more than one persons and were manually labeled as having positive or negative emotions by 20 volunteers. Out of the clips used for each class, 75% were used for training and 25% for testing. The movie scenes contained different variations possible in real world such as low illumination, pose variations, occlusion and small face size . Some example scenes are depicted in Fig 4.1.

Speaker Detection

Presence of the speaker in the video is detected by analyzing the ratio (R_s) of the inter-frame displacements of the lip FFPs (FFPs 12 to 19, Figure 4.7 d.) to

Table 4.6: Confusion matrices for FER using a) Only speaker’s facial expression, and b) Facial expressions of multiple actors c) Only lexical cue (SO). ARR: Average Recognition Rate; Posit: Positive; Negat: Negative. For each class, 291 clips: Training; 97 clips: Testing.

(a) ARR=76.3%	(b) ARR=74.8%	(c) ARR=84.7%																											
<table border="1" style="border-collapse: collapse; width: 100%;"> <thead> <tr> <th></th> <th>Posit</th> <th>Negat</th> </tr> </thead> <tbody> <tr> <th>Posit</th> <td style="text-align: center;">78.5</td> <td style="text-align: center;">21.5</td> </tr> <tr> <th>Negat</th> <td style="text-align: center;">25.9</td> <td style="text-align: center;">74.1</td> </tr> </tbody> </table>		Posit	Negat	Posit	78.5	21.5	Negat	25.9	74.1	<table border="1" style="border-collapse: collapse; width: 100%;"> <thead> <tr> <th></th> <th>Posit</th> <th>Negat</th> </tr> </thead> <tbody> <tr> <th>Posit</th> <td style="text-align: center;">79.2</td> <td style="text-align: center;">20.8</td> </tr> <tr> <th>Negat</th> <td style="text-align: center;">29.6</td> <td style="text-align: center;">70.4</td> </tr> </tbody> </table>		Posit	Negat	Posit	79.2	20.8	Negat	29.6	70.4	<table border="1" style="border-collapse: collapse; width: 100%;"> <thead> <tr> <th></th> <th>Posit</th> <th>Negat</th> </tr> </thead> <tbody> <tr> <th>Posit</th> <td style="text-align: center;">88.2</td> <td style="text-align: center;">11.8</td> </tr> <tr> <th>Negat</th> <td style="text-align: center;">18.8</td> <td style="text-align: center;">81.2</td> </tr> </tbody> </table>		Posit	Negat	Posit	88.2	11.8	Negat	18.8	81.2
	Posit	Negat																											
Posit	78.5	21.5																											
Negat	25.9	74.1																											
	Posit	Negat																											
Posit	79.2	20.8																											
Negat	29.6	70.4																											
	Posit	Negat																											
Posit	88.2	11.8																											
Negat	18.8	81.2																											

that of other FFPs. If R_s exceeds a threshold, speaker is present and the actor for which R_s has the maximum mean value is taken to be the speaker.

Facial Expression Recognition

Training: FER is performed using the algorithm outlined in section 4.2.1. The position of FFPs in the neutral face is found by selecting a neutral frame of the actor with facial pose closest to that present in the training video. For scale normalization, the inter-eye distance cannot be used due to pose variations. Instead, the distances between all possible pairs of FFPs is calculated and the maximum of those distances is taken as the normalizing distance. After feature selection using significance ratio test [197], the top 5 most relevant residues correspond to FFPs 6, 10, 11, 12 and 16. Selected residues are used for training SVM classifier with an RBF kernel with parameters $c = 4096$ and $g = 0.125$ (found using grid search). Frames with tracking failure were automatically excluded using method described in section 4.2.3.

Testing: A test clip is preprocessed and features are extracted in the same manner as for training sequences (Section 4.2.1). FER experiments were performed for both single actor and multi-actor cases. SVM classification results for both cases, averaged over 10 runs are given in the form of confusion matrices in Table 4.6(a) and (b) respectively. It is observed that recognition accuracy



a. Dialog by actor on left: “No, you see, you have the wrong idea.”



b. Dialog by female actor: “How are you feeling? Okay?...the greatest of luck.”

Figure 4.9: Examples of failures in using the multi-actor approach. a. Contradicting facial expressions of actors led to a wrong FER result. This brings out the need to use lexical cues from dialogs, b. In some cases, even fusion of visual and lexical cues gave incorrect results.

Table 4.7: Confusion matrices for classification a) Fusion Scheme 1, single actor, b) Fusion Scheme 1, multi-actor, c) Fusion Scheme 2, single actor, and d) Fusion Scheme 2, multi-actor. ARR: Average Recognition Rate (%); Pos: Positive; Neg: Negative. For each class, 291 clips: Training; 97 clips: Testing.

(a) ARR:88.3			(b) ARR:83.9			(c) ARR:90.9			(d) ARR:89.7		
	Pos	Neg		Pos	Neg		Pos	Neg		Pos	Neg
Pos	78.5	21.5	Pos	69.2	30.8	Pos	83.8	16.2	Pos	81.5	18.5
Neg	1.8	98.2	Neg	1.5	98.5	Neg	2.1	97.9	Neg	2.1	97.9

decreased slightly, in multi-actor approach. This is possible due to the fact that speaker might be more intensely displaying facial expressions as compared to other actors (E.g. Figure 4.7a.). Another reason can be contradiction in facial expressions of different actors (Figure 4.9a. and b.).

This brings in the role of using lexical cues. Using lexical cues alone gave an Average Recognition Rate of 84.7% (Table 4.6c.).

Experiments combining visual and lexical cues

Fusion was performed using two fusion schemes (Section 4.2.3) results of which, averaged over 10 runs, are shown in Table 4.7. For each run, training and test sets were randomly chosen. Results are shown for both single actor and multi-actor cases using both fusion schemes. Even after fusing lexical cues, recognition accuracy is better in single actor case.

An example of failure of multi-actor approach is shown in Figure 4.9b. On the other hand, an example of effectiveness of using multi-actor approach is in the case of Figure 4.7b. which is already discussed in section 4.2. It is observed that in cases when the speaker's emotions are not indicative of emotion of the scene, multi-actor approach was helpful but in our database such cases are not many which has led to a poorer performance of Multi-actor approach for ER.

Chapter 5

Emotion Recognition Applied to Automated Personality Assessment

Personality tests have become ubiquitous tools for analyzing personality traits. They are designed to test candidates for suitability for a particular job, test for specific character traits, find personal preferences for facilitating personalized services [3], design personal assistants like robotic butlers etc. Research in psychology has shown significant correlation between personality traits and service preferences [145]. For example, a person seeking sensation would prefer rock, heavy metal and punk music rather than religious music [106]. An extrovert person prefers cheerful music with vocals. Consequently, it has been shown that knowledge about user's personality can significantly benefit a music recommendation system.

Existing personality assessment methods are mostly based on self-reporting where the individual being assessed (referred hereafter as *target*) fills up a questionnaire. A score indicative of a measure of the target's personality trait is computed based on his responses. However, it has been observed that personality of a person varies over time [165]. Also, conducting personality tests at regular intervals may lead to certain psychological disorders in the target [202]

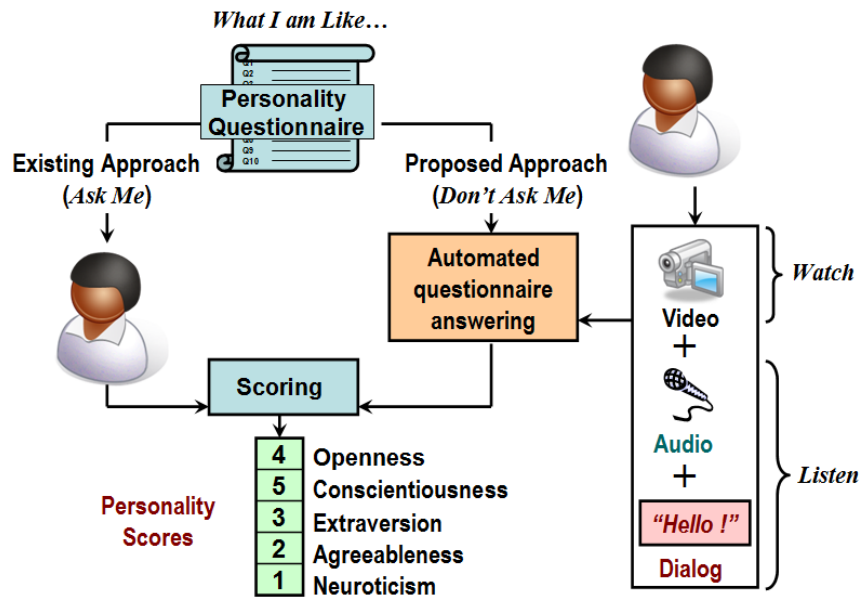


Figure 5.1: Proposed framework for automating the process of answering personality questionnaire by utilizing the audio, visual and lexical information. Additionally, answers to the personality questionnaire are utilized to obtain personality scores along the Big-Five dimensions, i.e. Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism.

and so is not recommended.

To remove intervention of the target in personality assessment, one alternative is to automate the answering of personality questionnaires. In this work, an approach (Figure 5.1) is proposed in this direction to automate answering one of the most popular personality questionnaires among professionals, the Big Five Inventory (BFI) originally containing 44 questions. A shorter version of the BFI comprising of 10 questions (BFI-10) is used as a first step towards evaluating our approach, which is found to be reliable [144] even though it cannot substitute for the full version of the BFI. BFI is based on the Five Factor Model (FFM) [116] which defines the personality along five dimensions, i.e. *Openness, Conscientiousness, Extraversion, Agreeableness* and *Neuroticism* (Section 5.1.2). Henceforth, usage of the term ‘BFI’ refers to BFI-10.

The proposed approach is evaluated by answering BFI for characters in movies (Figure 5.2). In the first step, clues from video, audio and dialog are

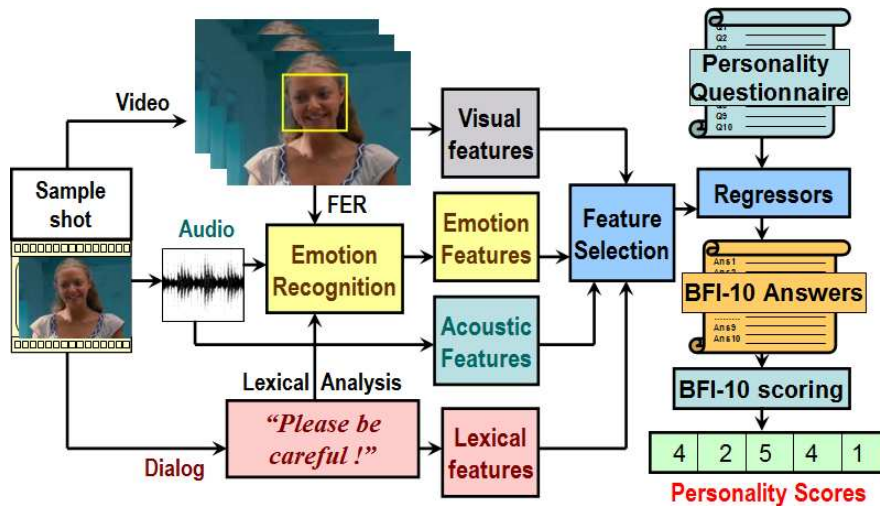


Figure 5.2: Algorithm of the framework proposed to automatically answer personality questionnaire (Big-Five Inventory-10, in this case). Evaluation is performed on movie characters. Answers to the questionnaire are mapped to the personality scores along the Big-five dimensions (Table 5.1).

combined for recognizing emotions of the character and subsequently extracting relevant ‘Emotion features’. Emotion features along with other features (Table 5.3), inspired by psychological studies, are used to learn linear regression models which relate the features to BFI answers. The linear regression models are regularized by simultaneous sparse and low rank priors, thus providing insights into features relevant for answering each question. Then the selected features are further used to learn kernel regression models which predict answers more accurately. In addition, final personality scores, in the range of 1 to 5, are computed from the predicted answers using BFI scoring scheme (Section 5.3.3), which is widely adopted by psychologists.

The main contributions of this work are summarized as follows. (1) Automating the answering of a personality questionnaire can help psychologists in cases where manually filling it up may not be feasible. (2) Inspired by psychological studies, high-level multi-modal features are presented which are extracted from audio/visual/lexical streams of individuals. These features are used to answer the personality questionnaire (specifically BFI-10). (3) Novel

regression models are learnt from the features based on sparse and low rank transformation and are used to predict the answers of the questionnaire.

5.1 Background

5.1.1 Models of Personality

Researchers in the field of psychology have proposed different models to describe human personality. Usually, personality is described along well defined dimensions. One of the earlier personality models was that of Cattell who described personality along 16 dimensions, commonly known as the 16 Personality Factors (16PF) [34]. Few of these dimensions are *Warmth*, *Liveliness*, *Emotional Stability*, etc. However, the factor analysis on which Cattell's model was based was questioned by other researchers [21]. Furthermore, the basis for choosing the 16 factors was unclear [183]. Later, Norman [125] suggested that Cattell proposed too many superfluous factors and his model could be reduced to just five factors. These five factors were initially labeled as Extraversion (talkative, assertive, energetic), Agreeableness (good-natured, cooperative, trustful), Conscientiousness (orderly, responsible, dependable), Emotional Stability (calm, not neurotic, easily upset) and Culture (intellectual, polished, independent-minded) [91]. These factors later on came to be known as the "Big-Five" [75] with Emotional Stability and Culture renamed to Neuroticism and Openness, respectively. The Big Five traits are also referred to as the Five Factors in the "Five Factor Model" (FFM) [116].

Another popular personality assessment tool is the Myers Briggs Type Indicator [123] which has also been found to correlate with the FFM [115]. It is based on a psychological model which evaluates personality based on four dimensions namely Extraversion/Introversion, Sensing/Intuition, Think-

Table 5.1: The Big-Five Dimensions and the associated traits.

Dimension	Associated traits of someone scoring high along the dimension
Openness	Artistic, Imaginative, Curious, Valuing intellectual matters
Conscientiousness	Self-disciplined, Orderly, Dutiful, Achievement striving, Responsible, Deliberation
Extraversion	Warm, Gregarious, Assertive, Active, Excitement seeking, having positive emotions
Agreeableness	Trustworthy, Straightforward, Altruistic, Modest, Tender-minded, Compliant
Neuroticism	Anxious, Depressed, Impulsive, Self-pitiful, Self-conscious, Fluctuating moods

ing/Feeling and Judgment/Perception.

The personality models mentioned above provide a complete description of personality. Apart from such models there are models which deal with specific aspects of personality. For example, an Egogram deals with the way one communicates socially [56]. Similarly Locus of Control [149] (LoC) indicates a belief whether one's actions determine the result one experiences (internal LoC) or do the results depend on some other agency (external LoC).

Out of the different personality models mentioned above, the work proposed in this work uses the Five Factor Model (FFM) as the basis, as it appears to be gaining more attention [91] in recent years. Another reason for choosing the FFM is the simplicity of a shorter version of the Big Five Inventory; which is a questionnaire based on the FFM (Section 5.1.2).

5.1.2 Five Factor Model (FFM) and the Big Five Inventory (BFI)

The FFM studies human personalities along five dimensions viz. **Openness, Conscientiousness, Extraversion, Agreeableness** and **Neuroticism**. The characteristics associated with these five dimensions are given in Table 5.1. Scores

along these dimensions are computed for a person from his response to a questionnaire. These questionnaires are also known as personality tests.

There are different personality tests available, based on the FFM, such as the Trait Descriptive Adjectives (TDA), Revised NEO Personality Inventory (NEO PI-R), Big Five Inventory (BFI), the ten item version of BFI (BFI-10) and so on. Each of these tests contain different number of multiple choice questions. As a first step towards automatic personality assessment, we have used BFI-10 test in this work due to its simplicity. BFI-10 contains 10 questions (Table 5.5) for assessing personality. Answering each question requires selecting one choice (value) from *Disagree Strongly(1)*, *Disagree a little(2)*, *Neither agree nor disagree(3)*, *Agree a little(4)*, *Agree Strongly(5)*. A pair of these questions together give a score for each FFM factor. For example, the score on Extraversion is computed by finding the average of the value for question 6 (*is outgoing, sociable*) and the reverse value for question 1 (*is reserved*) (6 - value for question 1). Note that questions 1 and 6 indicate opposing sentiments. Thus the personality score along the five factor dimensions of FFM are obtained.

5.2 Multimodal Feature Extraction

In this work, the personality questionnaire for movie characters is automatically answered based on multimodal features. Given a movie, before feature extraction, identification of characters and shot boundary is performed using the method proposed in [68]. Some manual intervention may be needed to correct the errors in recognizing the characters. After this step, each recognized face is tracked over the shot. The dialog for a shot is found by matching the begin and end times of the shot with the timings of the dialogs from the subtitles. Note that we consider only leading characters in the movie for our analysis chosen as mentioned in Section 5.5.1. Multimodal features (Table 5.3)

Table 5.2: Notations used in defining the features.

i	The i^{th} shot
c	Character c
n_c	Number of shots of character c
nd_c	Number of dialogs spoken by character c
$D_c^{(i)}$	Dialog
$p(w)$	Polarity of word w (Section 5.2.1)
E	Emotion set {An, Di, Fe, Ha, Sa, Su, Ne}
t_1	Begin time for dialog
t_2	End time for dialog
z	Acoustic pitch
en	Spectral energy
$ff1, ff2$	First and second formant frequencies
e	A certain emotion in the emotion set E

are extracted from each shot of a character (Section 5.2.1). Relevant features are selected for answering each question (Section 5.3.2). Notations used in defining features are described in Table 5.2.

5.2.1 Feature Extraction

Each of the used features is chosen based on its utility in answering BFI questionnaire, as identified by psychological studies. Note that the features selected are related to the individual personality and not the environment such as the amount of lighting, motion etc. This prevents personality prediction being affected by the genre type. Movie environment in itself conveys different emotions in different genres.

The features extracted from each shot (Table 5.3) are explained below along with the reason for their choice. The number in bracket tells the number of features associated with each item. Note that ‘ c ’ stands for the c^{th} character and ‘ i ’ stands for the i^{th} shot. Details about the methodology and experiments are presented in Sections 5.3 and 5.5, respectively.

1. $s_c^{(i)}$: Whether the character ‘ c ’ speaks in his i^{th} shot or not, provided that there are more than one character present in the shot. Speaker detection

Table 5.3: Explanation of the features extracted from each shot of the character (Notations defined in Table 5.2)

Feature (Explanation)	Computation
$s_c^{(i)}$ (Character speaker indicator)	= 0 (non-speech) = 1 (speech)
$e_c^{(i)}(e)$ (Emotion probability)	Emotion Feature Estimation (Section 5.2.2)
n_w (Number of words in a dialog)	From dialog
$t_c^{(i)}$ (Time spent per dialog)	$t_c^{(i)} = (t_2 - t_1)/n_w$
$p(D_c^{(i)})$ (Dialog polarity)	$= \frac{1}{n_w} \sum_{w \in D_c^{(i)}} p(w)$
$n^{(i)}$ (Number of persons)	Maximum no. of faces in the scene containing the shot
$h_{mean}^{(i)}, h_{min}^{(i)}, h_{max}^{(i)}, h_{re}^{(i)}$ Mean, minimum, maximum and relative entropy of h over the shot, where $h \in \{z, en, ff1, ff2\}$	Acoustic analysis (Section 5.2.1)

(Section 4.1.1) is used to identify the speaker in a shot. If the character ‘ c ’ is identified as the speaker and there are other faces detected (using a frontal face detector [190]) apart from ‘ c ’; $s_c^{(i)}$ is assigned a value 1, otherwise $s_c^{(i)} = 0$.

2. $e_c^{(i)}$ (Emotion probability for the shot, feature no. 1 to 7): It is a 7-element vector containing the probability of each of the seven emotions being present in the shot. Note that in computing emotion probability, positive and negative surprise are clubbed into one expression class namely surprise. It is calculated using the Emotion Recognition algorithm which is described later (Section 5.2.2). The probability of an emotion $e \in E$ is represented as $e_c^{(i)}(e)$. Emotions are related to personality traits. For example, reserved persons are found to be less happy as compared to those not so reserved [137]. Nervousness is related to fear [113].
3. $t_c^{(i)}$ (Feature no. 8): Time taken to speak a dialog.

4. n_w (Feature no. 9): Number of words in a dialog. Number of words combined with the time taken to speak a dialog is indicative of the speech rate which is low for a person with artistic interests [10].
5. $p(D_c^{(i)})$ (Dialog Polarity, Feature no. 10): Whether the dialog has positive ($p(D_c^{(i)}) > 0$) or negative ($p(D_c^{(i)}) < 0$) connotation as determined using Dictionary of Affect in Language (DAL) [198]. E.g. the word ‘excellent’ has positive connotation while ‘poor’ has negative connotation. Polarity of the dialog is calculated by averaging the polarity of all the words in the dialog which are present in DAL. Persons expressing more positive emotions are expected to be more trustworthy as compared to those expressing mostly negative emotions [181].
6. $n^{(i)}$ (Feature no. 11): Number of persons in the scene containing the i^{th} shot. A sociable person is expected to be with more people as compared to a reserved person [138]. All persons in a scene may be visible only in few frames of the scene. Therefore the number of persons in a scene is computed as the maximum number of faces detected over all the frames in the scene.
7. $h_{mean}^{(i)}, h_{min}^{(i)}, h_{max}^{(i)}, h_{re}^{(i)}$ (Features no. 12 to 27): Mean, minimum, maximum and relative entropy of the parameter h in a shot, where $h \in \{z, en, ff1, ff2\}$. Variable ‘ z ’ has been used to represent acoustic pitch in order to avoid confusion with probability represented by ‘ p ’. en , $ff1$ and $ff2$ represent spectral energy, first and second formant frequencies for the audio signal related to the shot. Spectral energy is a measure of energy at different frequencies and is high at dominant frequencies. First Formant frequency is the amplitude peak in the frequency spectrum of the audio signal corresponding to the lowest frequency. The amplitude peak with the next higher frequency is the second formant fre-

quency. Relative entropy of a random variable h with possible values $x_1, x_2, x_3, \dots, x_n$, is a measure of the uncertainty of a random variable and is defined as [118]:

$$h_{re} = \frac{1}{\log(|h|)} \left(- \sum_{i=1}^n P(x_i) \log(P(x_i)) \right). \quad (5.1)$$

Here $|h|$ denotes the cardinality of h .

Note that the features in item 7 above indicate voice properties such as loudness of the speech. These voice properties are useful in answering some of the questions. For example, it has been found that a fault finder is more likely to shout at others [181]. Also a person who can handle stress well is usually soft spoken [26].

5.2.2 Emotion Feature Estimation

The probability of the presence of each of the seven emotions in a shot is predicted by combining predictions from facial expression recognition, acoustic analysis and lexical analysis using Support Vector Machines (SVM) in a late fusion scheme. The details are explained below.

Facial Expression Recognition (FER) For each shot of a character, faces detected in each frame are resized to 50×50 pixels and Gabor features were extracted [18]. Principal Component Analysis (PCA) is applied on the resultant feature matrix to reduce its dimension from $50 \times 50 \times 8 \times 9 \times N (= 180,000 \times N)$ to $100 \times N$, where N is the number of frames. Reduced feature matrix is fed to SVM [36] to predict the posterior probability of each of the seven emotions.

Posterior probability for the whole shot is obtained as the mean of posterior probabilities of all the frames in that shot.

Acoustic Analysis From the audio track of a shot, prosodic features, namely pitch, spectral energy, frequency, Mel Frequency Cepstral coefficients (MFCC)

features, formant frequencies, Linear Predictive Coding coefficients, standard deviation of amplitude and mean, maximum and minimum values and standard deviation of pitch are used for emotion recognition. An SVM classifier trained using the training data predicts the probability of the audio track of a test shot belonging to each of the seven emotions.

Lexical Analysis of Dialogs Each shot is accompanied by one or more dialog. In the case of multiple speakers, we only consider the dialog(s) spoken by the character whose personality is being assessed. Emotions conveyed through the dialog are recognized based on emotions associated with ‘emotional words’ in the dialog. Examples of emotional words are ‘scary’, ‘jubilant’ etc. which convey emotions. Words such as ‘and’, ‘farther’ etc. are not counted as emotional words.

In order to predict whether a word (query word) is emotional or not, an affect (emotion) corpus is processed and it is found how often that word is associated; in the corpus; with each of the seven emotions i.e. Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness (Sa), Positive Surprise (Su+) and Negative Surprise (Su-) along with neutral (no prominent emotion) [32]. For analysis, the two emotion classes Positive Surprise and Negative Surprise are clubbed together into one class namely Surprise. Henceforth, ‘class’ will refer to one of the six emotions or the neutral. The affect corpus used for this work, is the UIUC children stories corpus [7] [6] which consists of 176 children’s stories by three authors (Grimm, H. C. Andersen and B. Potter). Each sentence in the corpus has been labeled with one of the seven emotions mentioned before or neutral (Ne). If the query word is found to occur more in sentences labeled with any one of the seven emotions as compared to neutral sentences, the word is more likely to be emotional. For more details on lexical analysis please refer to [32].

Multimodal Fusion In this work, we employ decision level multimodal

Table 5.4: Accuracies in recognizing each emotion (in %). Average Recognition Rate (ARR) = 79.21%. Note that ARR won't be equal to the mean of individual accuracies due to different number of samples for each emotion. An: Anger, Di: Disgust, Fe: Fear, Ha: Happiness, Sa: Sadness, Su: Surprise, Ne: Neutral, Acc: Accuracy

	An	Di	Fe	Ha	Sa	Su	Ne
Acc.	86.8	80.2	73.3	78.2	77.7	79.0	78.6

fusion. For decision level fusion, SVM has been found to perform much better as compared to other methods and thus we use SVM for our multimodal fusion task. To fuse the three modalities, namely, Visual, Acoustic and Lexical, the probability scores from these three modalities for a particular shot are concatenated to form a feature vector which is passed to SVM and the outputs are the probabilities of the seven emotions being present in the shot. The details of the experimental settings are given in Section 5.5.2. The accuracy of emotion recognition is given in Table 5.4. Average Recognition Rate (79.2%) is low possibly because movie data is close to reality and has difficulties such as low illumination or resolution, background noise etc.

5.3 Automating Answering of Personality Questionnaires

5.3.1 Features for Regression

Feature vectors extracted from shots for which $s_c^{(i)} = 1$ are concatenated to form a feature matrix F_c . Let $nd^{(c)}$ be the number of dialogs of the character c assuming that there is at most one dialog of a character in a shot. Then if $\mathbf{f}_c(i)$ is the 27 dimensional feature vector for the i^{th} shot (Formed as described in section 5.2.1), F_c is given by:

$$F_c = [\mathbf{f}_c(1), \mathbf{f}_c(2), \dots, \mathbf{f}_c(nd^{(c)})]. \quad (5.2)$$

Feature matrices F_c 's for all the characters included in the training set are further concatenated to form the final feature matrix F_{tr} , which is used for training the regressor. Let N_{tr} denote the number of characters used for training the regressor, N_s be total number of shots used for training and N_f be the number of features, respectively, then F_{tr} is an $N_s \times N_f$ matrix formed as follows:

$$F_{tr} = [F_1, F_2, \dots, F_{N_{tr}}]. \quad (5.3)$$

The matrix F_{tes} of test shots is formed in the similar fashion by using features extracted from the shots of test characters.

Instead of directly predicting the personality based on the above discussed features, the features are first mapped to the answers for BFI, based on a low-rank and sparse regression model. The personality attributes associated with these answers are easier to be recognized as compared to the Big-Five personality traits. Then the answers are used further to predict the personality through BFI scoring scheme. This two-stage scheme is motivated by the fact that the relationship between the features and personalities are generally difficult to describe through a simple linear model. Thus, we introduce the attributes to capture the middle-level information to narrow the gap between the features and personalities. In the regression from features to answers, the answers generally live in or near a low dimensional subspace, and each answer is believed to related with few features. Therefore, we enforce the linear regression model to be sparse and low-rank simultaneously. In this section, we will first introduce the regression model from features to answers (F2A) and then elaborate on the BFI-scoring from answers to personalities (A2P).

Table 5.5: Big Five Inventory-10 (BFI-10 [144]) used to assess personality and accuracy in predicting each answer using the proposed method (using Kernel Regression (KR1: before feature selection, KR2: after feature selection)). Results are compared with methods using linear Support Vector Machines (l-SVM), sparse SVM (s-SVM), linear regression (LR1: without regularizations, LR2: with regularizations) and SVM with Radial Basis Function kernel (k-SVM1: before feature selection, k-SVM2: after feature selection).

Instruction: How well do the following statements describe the character's personality? [144]								
You see the character as someone who ...	Accuracy (%)							
	Linear				Kernel			
	l-SVM	s-SVM	LR1	LR2	k-SVM1	k-SVM2	KR1	KR2
1. is reserved	87.0	86.6	81.1	86.2	23.2	87.4	86.6	87.4
2. is generally trusting	75.0	72.1	68.9	73.8	4.9	73.8	75.1	75.1
3. tends to be lazy	45.2	94.4	86.6	93.0	45.2	94.1	94.4	94.5
4. is relaxed and handles stress well	64.7	60.3	58.4	63.2	8.4	64.6	63.5	65.1
5. has few artistic interest	94.3	93.4	86.9	93.0	67.2	68.6	94.3	94.4
6. is outgoing, sociable	75.2	73.9	71.3	75.0	3.0	6.0	75.1	75.8
7. tends to find fault with others	71.5	71.2	67.8	71.9	3.8	72.9	71.1	73.3
8. does a thorough job	84.7	82.7	79.1	84.2	22.8	84.8	84.7	84.8
9. gets nervous easily	3.9	81.1	77.6	80.4	4.0	78.1	81.1	81.2
10. has an active imagination	5.0	88.1	81.9	87.2	5.0	7.0	88.1	88.2
Mean accuracy	60.7	80.4	76.0	80.8	18.8	63.7	81.5	82.0
Approx. training time per run (in sec)	920	150	1	1	900	1000	1000	3000

5.3.2 F2A by Sparse and Low-rank Transformation

Formulation: The pursued linear transformation from features to answers bears sparse and low-rank structure simultaneously as explained later in this section. Therefore, during learning such a transformation, we enforce the desired structures via specific matrix norms in addition to minimizing the regression error.

Let Z be a $10 \times N_{tr}$ matrix containing the ground truth of the 10 answers for all the training clips, F_{tr} be the training matrix defined in Section 5.3.1 and W be a $10 \times N_f$ matrix containing the parameters of the regression model. The objective is to estimate W with sparse and low-rank structure from the training

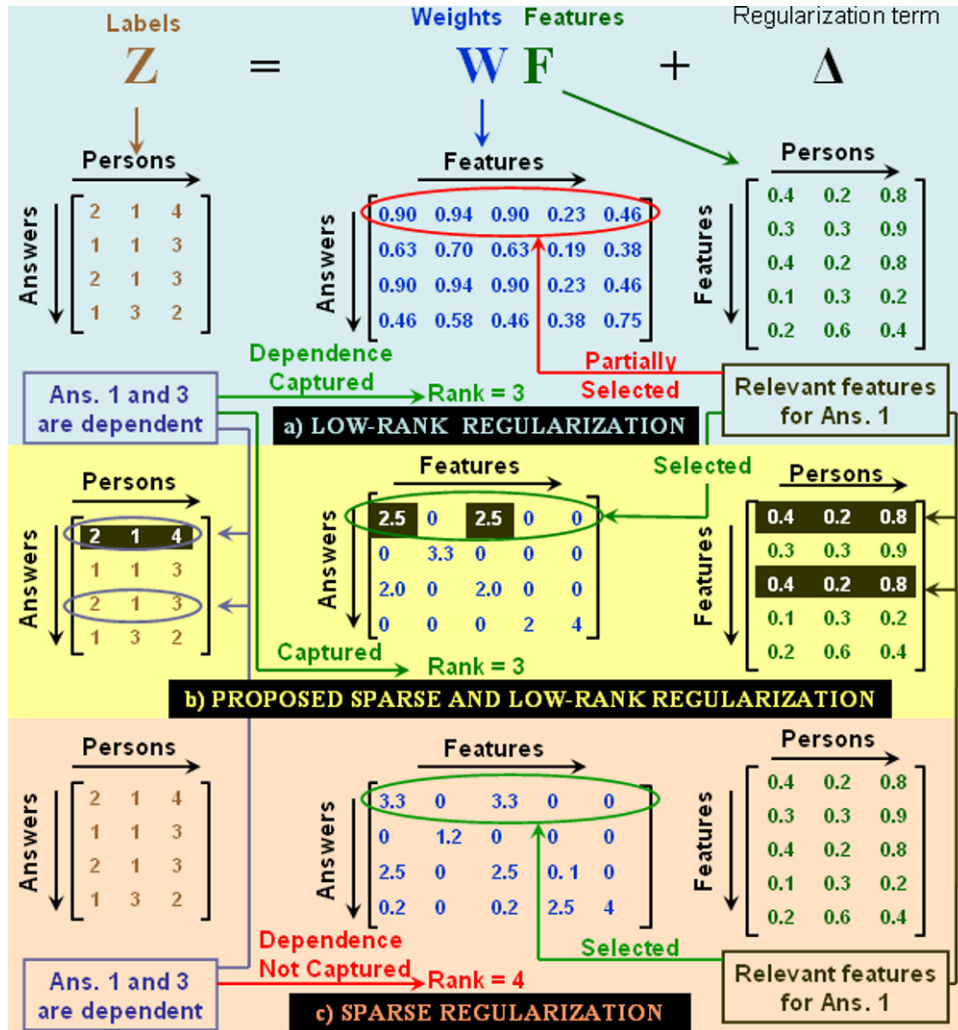


Figure 5.3: Effectiveness of using Sparse and Low-Rank regularization for our problem, as compared to Low-Rank regularization and Sparse regularization individually.

samples (F_{tr}, Z) , which can be formulated as,

$$\min_W \left\{ \frac{1}{2} \|Z - WF_{tr}\|_F^2 + \lambda_1 \|W\|_* + \lambda_2 \|W\|_1 \right\}. \quad (5.4)$$

Here $\|\cdot\|_F$ represents matrix Frobenius norm, $\|\cdot\|_*$ denotes the matrix nuclear norm which is a well-known convex surrogate for the matrix rank, and $\|\cdot\|_1$ is the matrix ℓ_1 norm which is known to be convex surrogate for the matrix ℓ_0 norm.

Reasons for Enforcing Sparse and Low-Rank Structure: In F2A, the

answers may be closely related. For example, a person who is outgoing and sociable (Question 6) would want to be thorough in his activities (Question 8) as this may attract others. These correlations would make the answers lie in a lower dimensional subspace. Considering the desired matrix W , each row in it corresponds to an answer. In case of relation between answers, the corresponding weights (and thus the corresponding rows) will be correlated. This makes W low-rank. The term $\|W\|_*$ in Equation 5.4 is a regularization to enforce W to be low rank.

Additionally each answer is believed to be related with only a subset of all the features. Consequently, many of the entries in W may be close to zero resulting in W to be sparse. The term $\|W\|_1$ in Equation 5.4 enforces W to be sparse. Therefore, we enforce the linear regression model to be sparse and low-rank simultaneously [226]. λ_1 and λ_2 are regularization parameters meant to achieve a trade off between regression error and the desired sparse and low-rank structure.

The above mentioned points are better explained using Figure 5.3 where low-rank and sparse regularization (LS) has been compared with both Low-rank regularization (LR) and Sparse regularization (SR) for F2A. For illustration, only 4 answers, 3 persons and 5 features are considered. Answers 1 and 3 are dependent being similar for the same person. Features 2 and 3 are relevant for predicting answers 1 and 2 due to a linear relationship between feature values and output labels in Z . Dependence is captured using LS since learned weight matrix is not full rank. LS is also able to select features 1 and 2 as relevant for answer 1. SR is able to select features but cannot capture dependence of answers 1 and 3. LR captures dependence since the learnt weight matrix is not full rank. But it selects feature 2 as the most significant instead of features 1 and 3.

Solving Equation 5.4 The objective function in Equation (5.4) can be ver-

ified to be convex and solved effectively. However, due to the non-smoothness of the regularization terms, the above problem can not be directly solved using traditional first-order or second-order optimization methods. In this work, we adopt the Augmented Lagrange Multiplier (ALM) method due to its efficiency and fast convergence rate [105].

Before applying ALM, we reformulate the primal problem by introducing two additional variables W_1 and W_2 as follows such that in each iteration we can have closed-form solution.

$$\begin{aligned} \min_{W, W_1, W_2} & \left\{ \frac{1}{2} \|Z - W F_{tr}\|_F^2 + \lambda_1 \|W_1\|_* + \lambda_2 \|W_2\|_1 \right\} \\ \text{subject to : } & W = W_1, W = W_2. \end{aligned} \quad (5.5)$$

It can be seen that the above problem is equivalent to the original problem in Equation (5.4). And its corresponding augmented Lagrange function is defined as:

$$\begin{aligned} \mathcal{L} = & \left\{ \frac{1}{2} \|Z - W F_{tr}\|_F^2 + \lambda_1 \|W_1\|_* + \lambda_2 \|W_2\|_1 \right. \\ & + \langle Y_1, W - W_1 \rangle + \langle Y_2, W - W_2 \rangle \\ & \left. + \frac{\mu}{2} \|W - W_1\|_F^2 + \frac{\mu}{2} \|W - W_2\|_F^2 \right\}. \end{aligned} \quad (5.6)$$

Here, μ is a positive constant to control the approximation precision and grows in the optimization iterations. The variables Y_1 and Y_2 are Lagrange multipliers for the penalty terms. By introducing the two additional variables W_1 and W_2 , the problem in Equation (5.6) can be optimized alternatively w.r.t. W , W_1 and W_2 by fixing the other two variables.

For the convenience of describing the details of the optimization process,

we first introduce the following soft-thresholding (shrinkage) operator:

$$\mathcal{S}_\varepsilon[x] = \begin{cases} x - \varepsilon, & \text{if } x > \varepsilon, \\ x + \varepsilon, & \text{if } x < -\varepsilon, \\ 0, & \text{otherwise,} \end{cases} \quad (5.7)$$

where ε is the threshold magnitude.

We optimize the problem in an alternative manner. Firstly, we fix the variables W and W_2 , then the objective function w.r.t. W_1 is

$$\min_{W_1} \left\{ \lambda_1 \|W_1\|_* + \langle Y_1, W - W_1 \rangle + \frac{\mu}{2} \|W - W_1\|_F^2 \right\}.$$

Since it is standard that [105],

$$U\mathcal{S}_\varepsilon[S]V^T = \operatorname{argmin}_X \left(\varepsilon \|X\|_* + \frac{1}{2} \|X - W\|_F^2 \right),$$

we can obtain the closed form solution for W_1 :

$$W_1 = U\mathcal{S}_{\lambda_1\mu^{-1}}[S]V^T, \quad (5.8)$$

where $(U, S, V) = \operatorname{svd}(W + \mu^{-1}Y_1)$ is obtained from the singular value decomposition.

When fixing W and W_1 , the objective function for W_2 becomes,

$$\min_{W_2} \left\{ \lambda_2 \|W_2\|_1 + \langle Y_2, W - W_2 \rangle + \frac{\mu}{2} \|W - W_2\|_F^2 \right\}.$$

Since we know [105],

$$\mathcal{S}_\varepsilon[W] = \operatorname{argmin}_X \left(\varepsilon \|X\|_1 + \frac{1}{2} \|X - W\|_F^2 \right),$$

the closed form solution to W_2 can be obtained:

$$W_2 = \mathcal{S}_{\lambda_2 \mu^{-1}}[W + \mu^{-1}Y_2]. \quad (5.9)$$

For W , the objective function is a standard quadratic function and the closed form solution of W is:

$$W = (ZF_{tr}^T + \mu(W_1 + W_2) - Y_1 - Y_2) (F_{tr}F_{tr}^T + \mu I)^{-1}. \quad (5.10)$$

The details of the optimization algorithm for estimating W are shown in Algorithm 1. The estimated model W is a $10 \times N_f$ matrix with each row corresponding to a question and each column corresponding to a feature. Predictions on test data are obtained as $Z_{tes} = WF_{tes}$ and are rounded to the nearest whole number. For Kernel Regression, the top 5 weight magnitudes for each question (row of W) are selected and the corresponding features are assumed relevant for answering that question. The selected features are further used to construct kernel for learning kernel-based regressor by slightly changing the formulations presented here.

Extension to Kernel Regression: The method presented above performs linear regression since W relates the feature matrix to the BFI answers. In order to extend it to Kernel regression, we need to Kernelize the feature matrix. As mentioned above, for predicting different answers, we select different relevant features and use the selected ones to construct the kernel matrix. Let $K^{(a)}$ be the desired kernel matrix for the a -th answer prediction whose elements $k_{ij}^{(a)}$ is calculated as:

$$k_{ij}^{(a)} = \exp\left(-\frac{1}{\sigma^2}(\|\mathbf{f}^{(a)}_i - \mathbf{f}^{(a)}_j\|^2)\right) \quad (5.11)$$

Here $\mathbf{f}^{(a)}_i$ refers to the selected features of the i^{th} sample. The problem in

Equation (5.4) is now re-formulated as:

$$\min_W \left\{ \sum_{a=1}^{10} \frac{1}{2} \|Z_a - W_a K_{tr}^{(a)}\|_F^2 + \lambda_1 \|W_a\|_* + \lambda_2 \|W_a\|_1 \right\}. \quad (5.12)$$

where Z_a and W_a denote the a -th row of the corresponding matrices. Note that in the kernelized case, sparse regularization is still imposed to encourage the regression model to select a small fraction of the training samples (serve as support samples as in SVM) for the answers prediction. Since for different answers prediction, their regression models W_a 's are independent and we can optimize them separately. And each sub-problem is similar to the optimization of Equation (5.4) and we can apply Algorithm 1 directly.

Algorithm 1 ALM algorithm for objective (5.5).

Input: $Z \in \mathbb{R}^{10 \times N_{tr}}$, $F \in \mathbb{R}^{N_f \times N_{tr}}$, λ_1 , λ_2 , iter_{\max} and ϵ .

Output: $W \in \mathbb{R}^{10 \times N_f}$.

Initialization: $W^{(0)} = W_1^{(0)} = W_2^{(0)} = Y_1^{(0)} = Y_2^{(0)} = 0$, $\mu_0 > 0$, $\rho > 1$, $k = 0$.

repeat

 Calculate $W_1^{(k+1)*} \leftarrow W_1$ according to Equation (5.8);

 Calculate $W_2^{(k+1)*} \leftarrow W_2$ according to Equation (5.9);

 Calculate $W^{(k+1)*} \leftarrow W$ according to Equation (5.10);

$Y_1^{(k+1)} = Y_1^{(k)} + \mu_k \left(W^{(k+1)*} - W_1^{(k+1)*} \right)$;

$Y_2^{(k+1)} = Y_2^{(k)} + \mu_k \left(W^{(k+1)*} - W_2^{(k+1)*} \right)$;

$\mu_{k+1} = \rho \mu_k$, $k \leftarrow k + 1$.

until $t > \text{iter}_{\max}$ or $|f(W^{(k+1)}) - f(W^{(k)})| < \epsilon$

A Note on Convergence The solution derived from the applied ALM method is guaranteed to converge to the optimal solution by following theorem [105].

Theorem 1 (Adapted from [105]). *Any accumulation point W^* of W_k^* is an optimal solution to the primal problem (5.5) and the convergence rate is at least $O(\mu_k^{-1})$ in the sense that*

$$\left| \frac{1}{2} \|Z - W^* F\|_F^2 + \lambda_1 \|W_k^*\|_* + \lambda_2 \|W_k^*\|_1 - f^* \right| = O(\mu_k^{-1}),$$

Factor	Associated Questions
Extraversion	1R, 6
Agreeableness	2, 7R
Conscientiousness	3R, 8
Neuroticism	4R, 9
Openness	5R, 10

Table 5.6: Questions in BFI-10 used for computing scores for each factor.

where f^* is the optimal value of problem (5.5).

From the above theorem, we can see that as the value of μ_k grows, the optimization algorithm will converge faster and it provides a Q-linear convergence rate [28].

5.3.3 A2P using BFI scoring scheme

Answers obtained for each of the ten questions are mapped onto personality scores along five dimensions using the standard methodology used for scoring BFI-10. The scoring scheme [144] is mentioned below.

There are questions associated with each of the five factors (Shown in table 5.6) which are used for computing score for that factor [144].

Here suffix ‘R’ to a question number indicates that the answer for this question will be reverse scored. The values of answers are in the range of 1 to 5. In order to reverse score an answer the value is subtracted from 6. For example, if questions 1 and 6 have values of 2 and 3 respectively, in order to calculate extraversion score, value for question one will be subtracted from 6 and it becomes 4 while value for question 6 remains 3. Next, for each of the Big Five factors, the values of questions associated with that factor are averaged. In our example, the final score for extraversion will be average of 4 and 3 i.e. 3.5.

5.4 Model for scoring the Questionnaire

The BFI scoring scheme presented in section 5.3.3 is very specific for the BFI-10 questionnaire. However, for another questionnaire the set of rules for scoring may be different based on complex inter-relationships between the questions. Designing an appropriate questionnaire and a set of rules that combines the answers to those questions is a non-trivial task. While more questions can help to get a better coverage, understanding interdependencies between questions to generate rules that can map to a personality score gets tougher. Thus any personality assessment method based on the questionnaire approach will be limited by these issues. In this section we present a solution to the issue raised above - the problem of *automating the process of computing a personality score for a known personality trait model from a set of questions for which the rules of combining their answers is not clear*.

Moreover an individual can have multiple shots of audio-visual feeds collected over a period of time which are also temporarily related. How to combine these shot level information to predict an overall personality assessment is another issue. For example, prediction on a noisy shot may not be very reliable. This section addresses this problem by using temporal information through a Conditional Random Field (CRF) model.

5.4.1 Conditional Random Field Model to Predict Personality Scores

In computing personality scores from questionnaires, the complex interrelationships between the questions need to be considered. Also, deciphering the relationship between questions and personality dimensions needs thorough psychological studies. Formulating a scoring scheme can be very difficult for a large questionnaire. This work proposes to use Conditional Random Fields

(CRF) for learning the scoring scheme. CRFs are discriminative undirected probabilistic graphical models proposed by Lafferty et al [101] which can be used as classifiers for temporal data. A major advantage of CRF is that its prediction is unaffected by complex interdependencies between features. According to its linear chain model, if y is a label sequence and x is an observation sequence, the probability of x having a label y is given as:

$$p(y|x) = \frac{1}{Z} \exp \sum_{t=1}^T \left(\sum_{i=1}^N \lambda_i f_i(y_t, x) + \sum_{j=1}^M \mu_j g_j(y_t, y_{t-1}, x) \right). \quad (5.13)$$

Here f_i and g_j are potential functions evaluating interaction and dependencies among features, respectively. λ_i and μ_j are weights and Z is a normalization factor.

The predicted BFI answers (Section 5.3.2) for the training data are used as features to train the CRF. Learnt CRF predicts the probabilities of each possible value of the personality score (1 to 5) for each shot of the character. Shots of a character have been used in the form of a temporal sequence with the order of the shots determined based on the order of their appearance in the movie. Using shots in this way can be justified by the observation that personality in a particular shot is closely related to the adjacent shots. Figure 5.4 shows the variation of personality scores (ground truth) across the Big-5 dimensions for the character *Jack Dawson* in the movie *Titanic*. Note that the shots with insufficient information about a personality dimension have been excluded from the plot. It is observed from the plot for Neuroticism that scores for the neighboring shots are mostly similar and their variation follows a trend of going from low to high. Similarly, a trend is observed in the variation of personality scores along other dimensions as well. These temporal relationships between adjacent shots are captured using CRF.

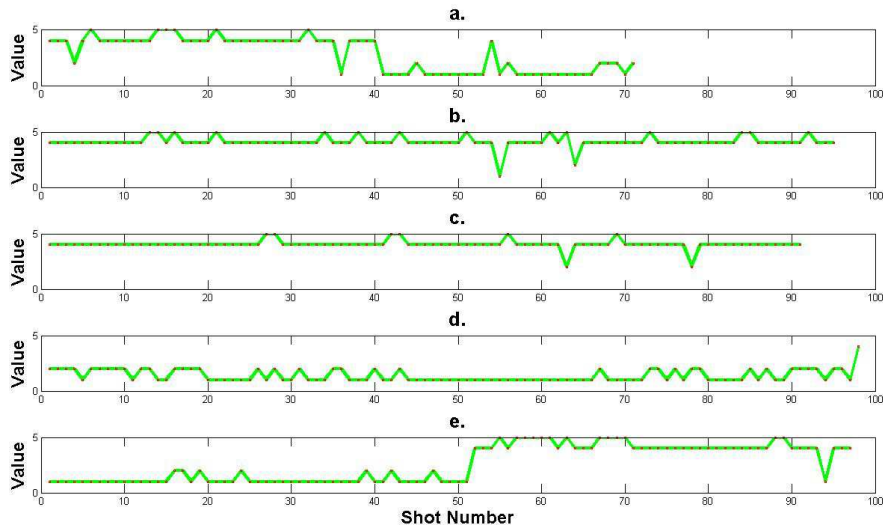


Figure 5.4: Variation of personality scores for the character Jack Dawson in *Titanic*. a. Openness, b. Conscientiousness, c. Extraversion, d. Agreeableness, e. Neuroticism. Apart from few outliers, the scores are temporally related supporting the use of CRF.

Once the personality scores are available for each shot of the character, they need to be fused together to get one value along each of the Big-5 dimensions. CRF inherently considers temporal information from all the preceding frames to predict the scores at the current frame. Cumulative personality scores are given by the prediction for the last shot since the prediction is made considering predictions in all previous shots and the dependencies between adjacent shots.

5.5 Experiments

5.5.1 Dataset

The works on personality assessment have observed human behavior in certain contextual situations such as an informal cocktail party [214], news bulletins [118] or meetings [140]. Behavior of a person in these context only reflects a particular aspect of his personality. For example, Oberlander and Nowson [127] tried to predict personality of bloggers and a possibility was expressed

that bloggers are more open than average, since distribution of Openness scores was skewed as compared to other four personality traits. It has been believed [141] that personality is a long-lived trait of a person as compared to emotions or even mood. Humans are found to get a better impression of others' personality if they interact for a longer time [33].

Considering the above points, in order to assess human personality, we must observe his behavior for a longer period of time and over different scenarios of his life. Due to non-availability of such a database, behavior of characters in movies has been analyzed. Although, movies cannot be a substitute for real life behavior, it has been found that actions of characters in movies are closer to real life as compared to acting on stage dramas [65].

Our dataset has clips from movies from different genres namely Comedy, Adventure, Drama, Fantasy and Action ¹. This makes our dataset more diverse since more scenarios of our life are covered as compared to context specific datasets.

The collected dataset consists of around 6000 clips. Selecting shots of characters with minimum 30 dialogs in a movie, reduces the final number of clips to 3907. Each clip is of around 4-7 seconds duration at 30 frames per seconds. The video resolution varies from 640×272 to 720×360 . Dialogs along with the timings are given in subtitles. To establish the ground truth, the emotion of the relevant character in each movie clip is manually labeled by 5 volunteers with one of the following: *Anger, Disgust, Fear, Happiness, Sadness, Surprise* and *Neutral*. Majority voting decides the assigned label. Besides annotating emotions, BFI questionnaire is answered by volunteers for relevant characters in each clip. Ground truth for personality scores is obtained from the ground truth for answers using the BFI scoring scheme (Section 5.3.3).

¹*Comedy: Evan Almighty, Adventure: Meet Dave, Titanic, Drama: The Prestige, Fantasy: Bedtime stories and Action: You don't mess with the Zohan.*

Table 5.7: Personality assessment accuracy comparison between Kernel Regression after feature selection (KR2), sparse SVM (s-SVM), kernel SVM (k-SVM), DirectScore (DS), Mohammadi et al. [118] and modified method of Mohammadi et al. [118]. It can be observed that KR2 performs better than the other methods for all of the Big Five personality traits. O: Openness, C: Conscientiousness, E: Extraversion, A: Agreeableness, N: Neuroticism.

Trait	Accuracies					
	KR2	s-SVM	k-SVM	DS	[118]	[118] _{mod}
O.	84.8	84.1	6.7	15.2	21.7	29.0
C.	81.1	79.4	80.7	12.5	5.3	42.3
E.	69.4	67.7	6.3	28.5	20.5	63.3
A.	60.3	57.5	58.9	12.8	12.3	17.4
N.	58.0	54.6	56.0	42.5	9.8	4.3
Avg.	70.7	68.7	41.7	22.3	13.9	31.0

The data needs to be divided into training and testing clips. Training clips are used to train (1) SVM for Facial Expression Recognition, (2) SVM for Acoustic analysis, (3) SVM for multimodal fusion (Section 5.2.2) and (4) regressors based on Sparse and Low-Rank transformations (both linear and kernel versions).

Training and Testing Data: 25% of the clips are used for training and remaining clips are used for testing. The training clips are uniformly sampled from each movie genre. 4-fold cross-validation is performed to ensure that each clip is tested once.

5.5.2 Predicting Answers using F2A Transformation

This subsection presents experimental results for F2A prediction. Averaged performance for F2A prediction is provided in Table 5.5 along with the training time². For Kernel Regression, the values of regularization parameters λ_1 and λ_2 are set as 10 and 1 respectively while they are set as 2 and 16 respectively for Linear Regression. Radial Basis Function (RBF) kernel with $\sigma = 20$ is used for Kernel Regression. These values are obtained from cross validation

²Experiments on Matlab on a PC with a 2.83 GHz, Core2 Quad, 64-bit processor and 8 GB RAM.

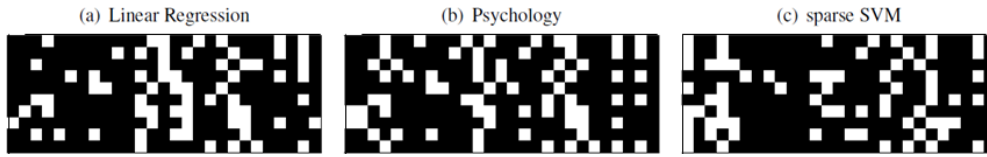


Figure 5.5: Images showing relevant features predicted using Linear Regression (LR), Sparse-SVM (s-SVM) and psychological findings. A white cell at position (i, j) means that the j^{th} feature is predicted to be relevant for predicting the i^{th} answer. The F1 scores for feature selection using LR and s-SVM are 0.81 and 0.70 respectively. A higher F1 score for LR indicates that it is better than s-SVM for feature selection.

on the training data. For ALM optimization, the parameters are selected as $\mu = 10^{-4}$, $\epsilon = 10^{-8}$, $iter_{max} = 10^3$ and $\rho = 1.1$.

Performance of our method on answer prediction is compared with alternative methods. Apart from using regression, predicting answers can also be treated as a 5-class classification problem, each class corresponding to one of the 5 values for the answers (Section 5.1.2). The classification problem is solved based on several SVM classifiers including Linear SVM (l-SVM), Sparse SVM (s-SVM) and Kernel SVM (k-SVM) using Radial Basis Kernel Function. Linear Regressions are also evaluated with (LR2) and without (LR1) the regularization term.

We also evaluate the effectiveness of selecting features relevant for answering each question by using Linear Regression (LR) (Section 5.3.2). Additionally, feature selection is also performed by using s-SVM. For predicting an answer, both LR and s-SVM output a weight to each feature. The first k highest weighted features are selected such that sum of their weights (absolute value) achieve 50% of the sum of all the weights. Each of the two feature selection techniques is compared with feature choice inspired by psychology (Fig. 5.5). The F1-score for feature selection using LR and s-SVM are found to be 0.81 and 0.70 respectively. Since LR has higher F1 score, the corresponding selected features are further used for predicting BFI answers. Out of all the selected features, only top 5 are used for further experiments.

Table 5.5 shows results of predicting answers using k-SVM before (k-SVM1) and after (k-SVM2) feature selection. The performance of Kernel regression before feature selection (KR1) is also shown. The efficacy of adding regularization constraint (sparsity & low rank) is demonstrated by better results in the case of LR2 as compared to LR1. in Table 5.5. From the results, it is observed that KR2 performs the best although it is computationally the most expensive. Prediction accuracy using KR2 (Table 5.5) exceeds 70% (well above the accuracy by chance, 20%) for all answers except answer no. 4, related to the character being relaxed and handling stress well. One possible reason for the low accuracy of k-SVM is that it overfits the training data and the generalization is not good.

5.5.3 Accuracy for Personality Prediction

After obtaining the answers to BFI, personality scores are computed based on the algorithm presented in Section 5.3.3. Ground truth for personality scores is obtained from the ground truth for answers using the BFI scoring scheme. Personality assessment accuracy comparisons are presented in Table 5.7 between the proposed approach and the following baseline approaches for personality prediction:

Using Sparse SVM (s-SVM) : Answers are predicted using s-SVM (Section 5.5.2) and then the personality scores are predicted using the BFI scoring scheme.

Using Kernel SVM (k-SVM) : Answers are predicted using k-SVM with RBF kernel (Section 5.5.2) and then the personality scores are predicted using the BFI scoring scheme.

DirectScore (DS): In the case of direct prediction from features, the input is the feature matrix F_{tr} (Section 5.3.1) and the output labels are the person-

ality scores for all characters used for training RBF kernel SVM for the five dimensions. The learned SVM then predicts the personality scores for the test characters.

Mohammadi et al. ([118]) Acoustic features including average, minimum, maximum and relative entropy of pitch, formant frequencies, energy and speaking rate are used for predicting the Big-Five personality traits using RBF kernel SVM. Scores are quantized to whole numbers from 1 to 5.

Mohammadi et al. modified ([118]_{mod}): Predicting answers based on features used by Mohammadi et al. and then predict personality traits using the BFI scoring scheme.

The results in Table 5.7 show that the proposed KR method performs the best in predicting all of the Big-Five personality traits. It is observed that predicting personality scores directly from the features (DS) is not very effective. [118]_{mod} also performs better than [118]. This opens up the avenue of answering questionnaire as an intermediate step in predicting personality traits. Performance of [118]_{mod} is still poor possibly because of not using visual and lexical features which are also found to be significant in predicting answers. Note that in [118], experimental data was captured in controlled conditions and the performance was reported by quantizing the personality scores into 2 classes: High and Low.

There exists correlation between the accuracies of predicted answers and the personality traits. Neuroticism has the worst prediction accuracy which can be attributed to the low prediction rate of answer 4, which Neuroticism is associated with. Similarly, Agreeableness is associated with answers 2 and 7 which also have lower prediction accuracy. Consequently, prediction of Agreeableness is affected.

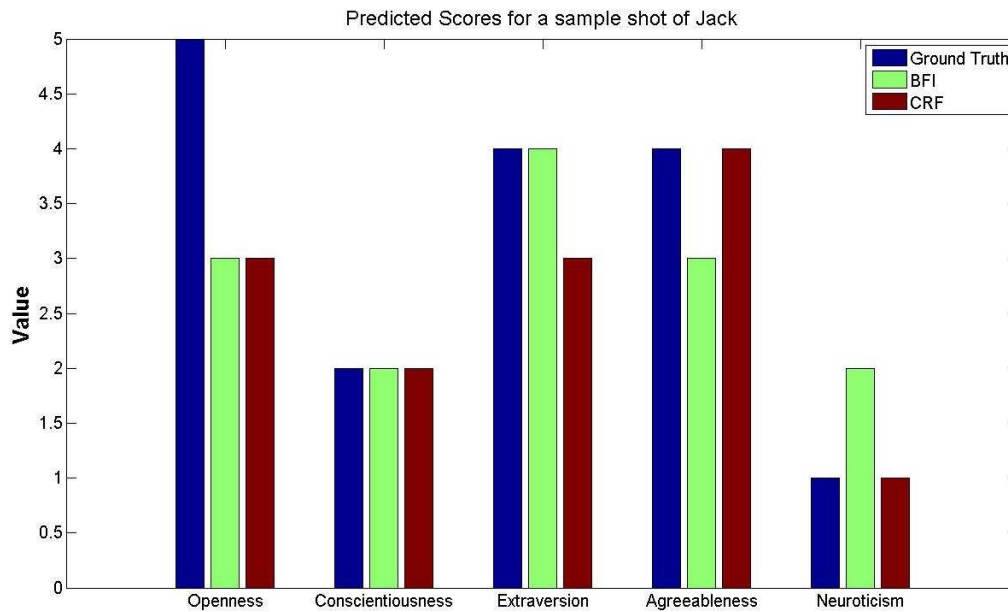


Figure 5.6: Predicted personality scores for a sample shot of Jack. Predictions using the proposed method (CRF) are comparable to the traditional method (BFI).

5.5.4 Personality Prediction Using Learnt CRF model

For predicting personality scores, a CRF model is trained using the scaled conjugate gradient algorithm implemented using Kevin Murphy’s CRF toolbox [4]. The parameters selected for training CRFs are: regularizer weight $\alpha = 1$ and maximum iterations = 200. Note that the shots with personality values labeled as 3 are excluded from training data since they lack sufficient information about the personality trait.

The output of the CRF is the personality scores for each shot of the characters. The prediction is in the form of probabilities corresponding to the 4 values for the score (Value of 3 being excluded). The value with the highest probability is taken as the predicted score if it exceeds a threshold θ . If none of the probabilities exceed θ , prediction is given a value 3. A comparison is also made with standard BFI scoring scheme (Section 5.3.3) to predict shot-level personality scores from answers.

Figure 5.6 shows the predicted personality scores for a sample shot of Jack

Table 5.8: Accuracy in scoring the BFI: Comparing CRF and BFI. O: Openness, C: Conscientiousness, E: Extraversion, A: Agreeableness, N: Neuroticism. BFI: BFI-scoring scheme, CRF: Proposed approach, CRF_{fused} : Predicting cumulative personality scores

Trait	Accuracies		
	Shot-level	Cumulative	
	<i>BFI</i>	<i>CRF</i>	CRF_{fused}
O.	86.2	88.4	84.3
C.	82.8	85.9	81.6
E.	76.6	85.1	82.0
A.	73.4	76.1	75.5
N.	69.1	76.4	77.5
Avg.	77.6	82.4	80.2

using both BFI and CRF. Predictions using CRF are close to that using BFI scoring since difference between the two predictions is never more than one unit for all the dimensions. As compared to the ground truth, CRF predicts 3 dimensions accurately as compared to 2 dimensions of BFI resulting in an accuracy of 60% for CRF and 40% for BFI *for this shot*. The prediction performances of CRF and BFI averaged over all the shots in the test dataset over four runs is reported in Table 5.8 (Shot-level). CRF is found to perform better than BFI for all the Big-5 dimensions.

Predicted scores at each shots are utilized to predict one set of personality scores for each character (Cumulative scores). Predicted values are compared with the ground truth annotations at the movie level. Prediction accuracies averaged over all the characters over 4 runs are presented in Table 5.8 last column. Even for the cumulative scores, Openness is predicted with the best accuracy. The average accuracy of 80.2% supports the use of CRF for fusing the personality scores of individual shots. Concluding issues about the proposed personality assessment are presented in the next chapter.

Chapter 6

Conclusion and Future Works

The current approaches to Emotion Recognition mostly work well on lab recorded data with sufficient lighting, face size, frontal face pose and so on. Moreover, most of the datasets have exaggerated emotions which are not close enough to the real world emotions. This thesis presented approaches to deal with the difficulties in the data namely that of face pose, illumination, face size, facial expression intensity and head motion.

In comparison with the 2D data, 3D data has an extra depth information. An approach has been proposed using 3D flows called as residues. As compared to the traditional 2D optical flow methods, 3D flow proved to perform better when evaluated on the BU-3DFE database which is one of the most widely used 3D facial expression databases. With technological advancement, capturing 3D data can be performed easily in real time making the proposed approach feasible for real life applications (See section 3.1 for details on availability of 3D data in real life).

For such applications an important issue of concern is the low intensity of facial expressions as observed in real life. Experimental analysis was conducted to assess the effect of lower intensity expression on the recognition performance. As expected, it was observed that flow for higher gradations of

experiments carry more information about the expression than the lower gradations. Consequently, the proposed approach needs to be improved to handle lower intensity of expressions.

Using 3D residues has limitations as well. Optical flow is very effective in recognizing facial expressions because of the universal nature of facial expressions which causes the motion of facial feature points to be similar across different subjects. However, it is also possible that when large interpersonal variations occur, accuracy in recognizing facial expression may decrease. We need to improve the proposed algorithm to handle more subtle and spontaneous expressions which may involve more interpersonal variations. Moreover, this approach is subject dependent approach which means that it requires a 3D model for the neutral face of the subjects.

Another subject independent approach has been proposed which is relevant to the real life situations where it might not be always possible to have the neutral facial model corresponding to a test expression. One vs. all scheme of classification was implemented using Support Vector Machines (SVM). For each individual classifier, a feature selection was performed using the significance ratio test. Promising classification results support the presented feature extraction technique.

Both the approaches proposed for 3D FER are geometry based approaches involving Facial Feature Points (FFPs). A major drawback of geometry based approaches is that the selection of these FFPs usually involves manual intervention. Choice of appropriate FFPs is also a debatable issue. A method needs to be devised to extract the feature points from the 3D mesh model. Combining appearance and geometry can be a good way to achieve this.

Considering 3D FER, there is a wide possibility of future research in the following areas:

- **Formation of a representative database:** For development of algo-

rithms for 3D FER, availability of a 3D facial expression database is a must. Databases presently available have been captured under controlled conditions mainly because available 3D scanners cannot operate satisfactorily in outdoor real life situations. However, for applying the FER algorithms in daily life applications it becomes necessary to evaluate them under natural conditions instead of a laboratory setup. Moreover such data should have naturally expressed emotions instead of exaggerated ones.

- **Dealing with computational and storage complexities:** Yin et al. [167] report facing problems in processing and storage of their 3D dynamic models. Because of these limitations, presently it is possible to record only short duration 3D videos where the person begins from neutral and deliberately shows expression and then comes back to neutral. Natural expressions are prolonged and so longer 3D videos need to be recorded. Also, the BU-4DFE database in its present form takes around 500GB of storage space. Techniques need to be devised for compact storage of the 3D models especially the dynamic models.

The difficulties faced in a real-life data can be experienced more in a movie data as compared to the currently available 3D data. Considering this, a dataset was constructed to facilitate ER in movies. A dynamic weighting based multi-modal approach has been proposed to fuse lexical and visual clues in order to recognize positive and negative emotions of actors in movies. Result for facial expression was considered to be of high confidence if face was less affected by variations in pose, scale, expression intensity etc. On the other hand, result of lexical analysis was relied upon more if its Semantic Orientation was larger in magnitude.

Several issues will be considered for future research. Detecting and local-

izing FFPs for non-frontal faces is still difficult. The face tracking algorithm has to be made robust to variations in facial pose and head motion by using a generic model for tracking in which facial shape is better defined. In extracting lexical clues, we observe problems due to different meanings of a word and due to different ways of saying the same message. Using the context can be helpful for determining Semantic Orientation under word meaning disambiguation.

The effect of considering emotions of all the relevant actors in a movie scene was analyzed for recognizing emotion of the scene. Multi-actor approach has been found to be helpful in cases when speaker displays an emotion contradicting to the emotion of the scene.

ER has been applied to automate the answering of personality assessment questionnaire. The proposed approach is evaluated by answering Big-Five Inventory (BFI) for movie characters using the proposed multimodal features which include predicted emotion, acoustic features from audio and lexical clues from dialog. The performance of the proposed algorithm is expected to be further improved by extending to the full version of BFI. Other questionnaires such as NEO-PIR [116] can also be explored. Such extensions require proposing new multimodal features.

The accuracy of the method depends on the accuracy of character identification and emotion recognition. An improvement is needed in the accuracy at these processing stages. Lexical clues from the dialog are easily available for movies. Even in real life, dialogs can be extracted using Automatic Speech Recognition (ASR) technology which is developing. Especially for a real life application using dialogs, it is possible to ensure good quality of sensors to capture speech. This ensures availability of lexical and even the acoustic cues.

A method is also presented to learn a scoring scheme using a Conditional Random Field (CRF) model which enables the approach to be generalized for scoring a wide range of personality questionnaires. Otherwise, to formu-

late a scoring scheme, complex interdependencies between questions needs to be thoroughly studied based on psychology. The proposed approach shows promising results as compared to the standard BFI scoring scheme.

Relating to the work on automated personality assessment, there are ethical and privacy issues associated with monitoring individuals. However, we conducted our work for research purposes and do not consider these issues. Even considering this issue, it is also possible that users may be willing to share their information considering the benefits they get in return. Still movies are still in somewhat controlled environment and a real-life data will be more useful for evaluating the proposed approach. This and the other limitations mentioned above open directions for future research.

Bibliography

- [1] Dimensional Imaging. Retrieved March 3 2012, from <http://www.di3d.com/index.php>.
- [2] Kinect. Retrieved March 3 2012, from <http://www.xbox.com/en-US/kinect>.
- [3] <http://hunch.com>.
- [4] Conditional Random Field (CRF) toolbox for Matlab. <http://www.cs.ubc.ca/~murphyk/Software/CRF/crf.html>.
- [5] The SEMAINE database. <http://semaine-db.eu/>.
- [6] C.O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586. Association for Computational Linguistics, 2005.
- [7] E.C.O. Alm. *Affect in text and speech*. PhD thesis, University of Illinois at Urbana-Champaign, 2008.
- [8] Z. Ambadar, JW Schooler, and JF Cohn. Deciphering the enigmatic face. *Psychological Science*, 16:403–410, 2005.
- [9] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.

- [10] W. Apple, L.A. Streeter, and R.M. Krauss. Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, 37(5):715–727, 1979.
- [11] S. Argamon, S. Dhawle, M. Koppel, and J.W. Pennebaker. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*. Citeseer, 2005.
- [12] Wendy S. Ark, D. Christopher Dryer, and Davia J. Lu. The emotion mouse. In *Proceedings of the 8th International Conference on Human-Computer Interaction: Ergonomics and User Interfaces - Volume I*, pages 818–823, Hillsdale, NJ, USA, 1999. L. Erlbaum Associates Inc.
- [13] A.B. Ashraf, S. Lucey, J.F. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon, and B.J. Theobald. The painful face: Pain expression recognition using active appearance models. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, pages 9–14. New York, USA, 2007.
- [14] A. Austermann, N. Esau, B. Kleinjohann, and L. Kleinjohann. Prosody based emotion recognition for mexi. In *Proceedings of the IEEE/RSJ Int. Conference on Intelligent Robots and Systems (IROS 2005), Edmonton, Canada, 2005*.
- [15] R. Banse and K.R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614, 1996.
- [16] 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 *CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, 2003.
- [17] MS Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 2005.

- [18] M.S. Bartlett, G. Littlewort, MG Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Proc. Conf. Face & Gesture Recognition*, pages 223–230, 2006.
- [19] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech communication*, 40(1-2):117–143, 2003.
- [20] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell, and M. Wong. You stupid tin box-children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proc. LREC*, pages 171–174, 2004.
- [21] W.C. Becker. The matching of behavior rating and questionnaire personality factors. *Psychological Bulletin*, 57(3):201, 1960.
- [22] L. Benedikt, D. Cosker, P.L. Rosin, and D. Marshall. 3D Facial Gestures in Biometrics: from Feasibility Study to Application. In *2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, BTAS*, pages 1–6, 2008.
- [23] L. Benedikt, V. Kajic, D. Cosker, PL Rosin, and D. Marshall. Facial Dynamics in Biometric Identification. In *British Machine Vision Conference*, 2008.
- [24] S. Berretti, A.D. Bimbo, P. Pala, B.B. Amor, and M. Daoudi. A set of selected sift features for 3d facial expression recognition. In *20th International Conference on Pattern Recognition (ICPR)*, pages 4125–4128. IEEE, 2010.
- [25] M.J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1997.
- [26] M.R. Bonner. Changes in the speech pattern under emotional tension. *The American Journal of Psychology*, 56(2):262–273, 1943.
- [27] S.E. Bou-Ghazale and J.H.L. Hansen. A comparative study of traditional and newly proposed features for recognition of speech under stress. *Speech and Audio Processing, IEEE Transactions on*, 8(4):429–442, 2000.

- [28] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge Univ Pr, 2004.
- [29] S. Burger, V. MacLaren, and H. Yu. The ISL meeting corpus: The impact of meeting type on speech style. In *Seventh International Conference on Spoken Language Processing*, 2002.
- [30] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, pages 205–211. ACM, 2004.
- [31] C. Busso, S. Lee, and S. Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(4):582–596, 2009.
- [32] R.A. Calix, S.A. Mallepudi, B. Chen, and G.M. Knapp. Emotion Recognition in Text for 3-D Facial Expression Rendering. *Multimedia, IEEE Transactions on*, 12(6):544–551, 2010.
- [33] D.R. Carney, C.R. Colvin, and J.A. Hall. A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41(5):1054–1072, 2007.
- [34] R.B. Cattell, H.W. Eber, and M.M. Tatsuoka. Handbook for the sixteen personality factor questionnaire (16 pf). 1970.
- [35] Chih Chung Chang and Chih Jen Lin. *LIBSVM: a library for support vector machines*, 2001.
- [36] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

- [37] Y. Chang, M. Vieira, M. Turk, and L. Velho. Automatic 3D Facial Expression Analysis in Videos. In *Analysis and Modelling of Faces and Gestures: Second International Workshop, AMFG 2005*, page 293. Springer, 2005.
- [38] L.S. Chen. *Joint Processing of Audio-Visual Information for the Recognition of Emotional Expressions in Human Computer Interaction*. PhD thesis, University of Illinois at Urbana-Champaign, 2000.
- [39] LS Chen, TS Huang, T. Miyasato, and R. Nakatsu. Multimodal human emotion/expression recognition. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 366–371, 1998.
- [40] 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 and Image Understanding*, 91(1-2):160–187, 2003.
- [41] J.F. Cohn. Foundations of human computing: facial expression and emotion. In *Proceedings of the 8th International Conference on Multimodal Interfaces*. ACM, 2006.
- [42] JF Cohn, LI Reed, Z. Ambadar, J. Xiao, and T. Moriyama. Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *IEEE International Conference on Systems, Man and Cybernetics*, 2004.
- [43] JF Cohn, AJ Zlochow, JJ Lien, and T. Kanade. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 396–401, 1998.
- [44] T.F. Cootes, G.J. Edwards, C.J. Taylor, et al. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, 23(6):681–685, 2001.
- [45] D.W. Cunningham, M. Kleiner, H.H. Bühlhoff, and C. Wallraven. The components of conversational facial expressions. In *Proceedings of the 1st Symposium*

- on Applied perception in graphics and visualization*, pages 143–150. ACM New York, USA, 2004.
- [46] M.K. De Mooij. *Consumer behavior and culture: Consequences for global marketing and advertising*. Sage, 2004.
- [47] F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Fourth International Conference on Spoken Language, ICSLP*, volume 3, pages 1970–1973. IEEE, 1996.
- [48] L. Devillers and I. Vasilescu. Reliability of lexical and prosodic cues in two real-life spoken dialog corpora. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, 2004.
- [49] L. Devillers and L. Vidrascu. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Ninth International Conference on Spoken Language Processing*, 2006.
- [50] L. Devillers and L. Vidrascu. Real-life emotion recognition in speech. *Speaker Classification II*, pages 34–42, 2007.
- [51] G. Donato, MS Bartlett, JC Hager, P. Ekman, and TJ Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.
- [52] F. Dornaika and F. Davoine. Simultaneous facial action tracking and expression recognition using a particle filter. In *Tenth IEEE International Conference on Computer Vision, ICCV*, volume 2, 2005.
- [53] F. Dornaika and F. Davoine. Simultaneous facial action tracking and expression recognition in the presence of head motion. *International Journal of Computer Vision*, 76(3):257–281, 2008.
- [54] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60, 2003.

- [55] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.C. Martin, L. Devillers, S. Abrilian, A. Batliner, et al. The humane database: addressing the collection and annotation of naturalistic and induced emotional data. *Affective computing and intelligent interaction*, pages 488–500, 2007.
- [56] J.M. Dusay. *Egograms: How I see you and you see me*. Harper & Row, 1977.
- [57] P Ekman and W V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.
- [58] P. Ekman, W.F. Friesen, and J.C Hager. *Facs manual, investigator’s guide*, 2002.
- [59] P. Ekman, W.V. Friesen, and J.C. Hager. *Facial action coding system*. Consulting Psychologists Press Palo Alto, CA, 1978.
- [60] Paul Ekman and Erika L. Rosenberg, editors. *What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system(FACS)*. Oxford University Press, 1997.
- [61] M. El Ayadi, M.S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.
- [62] M.M.H. El Ayadi, M.S. Kamel, and F. Karray. Speech emotion recognition using gaussian mixture vector autoregressive models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007.*, volume 4. IEEE, 2007.
- [63] R. El Kaliouby and P. Robinson. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *International Conference on Computer Vision and Pattern Recognition Workshop.*, pages 154–154, 2004.
- [64] I.S. Engberg, A.V. Hansen, O. Andersen, and P. Dalsgaard. Design, recording and verification of a danish emotional speech database. In *Fifth European Conference on Speech Communication and Technology*, 1997.

- [65] K. ESCH. Movie acting: The film reader. *Film Quarterly*, 59(1):62–63, 2005.
- [66] IA Essa and AP Pentland. Facial expression recognition using a dynamic model and motion energy. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 360–367, 1995.
- [67] IA Essa and AP Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, 19(7):757–763, 1997.
- [68] M. Everingham, J. Sivic, and A. Zisserman. “Hello! My name is... Buffy” – automatic naming of characters in TV video. In *Proceedings of the British Machine Vision Conference*, 2006.
- [69] B. Fasel, F. Monay, and D. Gatica-Perez. Latent semantic analysis of facial action codes for automatic facial expression recognition. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 181–188. ACM New York, USA, 2004.
- [70] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *Biomedical Engineering, IEEE Transactions on*, 47(7):829–837, 2000.
- [71] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio. Detecting group interest-level in meetings. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 489–492, 2005.
- [72] C. Ghys, N. Paragios, and B. Bascle. Understanding 3D Emotions Through Compact Anthropometric Autoregressive Models. *Lecture Notes in Computer Science*, 4291:793–802, 2006.
- [73] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis. A dimensional approach to emotion recognition of speech from movies. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 65–68. IEEE Computer Society, 2009.

- [74] S.B. Gokturk, J.Y. Bouquet, C. Tomasi, and B. Girod. Model-based face tracking for view-independent facial expression recognition. In *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, page 287. IEEE Computer Society Washington, DC, USA, 2002.
- [75] L.R. Goldberg. Language and individual differences: The search for universals in personality lexicons. In *Wheeler (Ed.), Review of Personality and social psychology*, 1:141–165, 1981.
- [76] B. Gong, Y. Wang, J. Liu, and X. Tang. Automatic facial expression recognition on a single 3d face by exploring shape deformation. In *Proceedings of the seventeen ACM international conference on Multimedia*, pages 569–572. ACM, 2009.
- [77] James J. Gross, Steven K. Sutton, and Timothy Ketelaar. Relations between affect and personality: Support for the affect-level and affective-reactivity views. *Personality and Social Psychology Bulletin*, 24(3):279–288, 1998.
- [78] H. Gunes and M. Pantic. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *Intelligent Virtual Agents*, pages 371–377. Springer, 2010.
- [79] H. Gunes and M. Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 4, pages 3437–3443. IEEE, 2005.
- [80] M. Hahnel, A. Wiratanaya, and K. Kraiss. Facial Expression Modelling from Still Images Using a Single Generic 3D Head Model. *Lecture Notes in Computer Science*, 4174:324–333, 2006.
- [81] A. Hanjalic and L.Q. Xu. Affective video content representation and modeling. *Multimedia, IEEE Transactions on*, 7(1):143–154, 2005.

- [82] J.H.L. Hansen and S.E. Bou-Ghazale. Getting started with susas: A speech under simulated and actual stress database. In *Fifth European Conference on Speech Communication and Technology*, 1997.
- [83] J. Hirschberg, S. Benus, J.M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girard, M. Graciarena, A. Kathol, L. Michaelis, et al. Distinguishing deceptive from non-deceptive speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.
- [84] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll. Bimodal fusion of emotional data in an automotive environment. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2005.
- [85] H. Hong, H. Neven, and C. Von der Malsburg. Online facial expression recognition based on personalized galleries. In *Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. Proceedings*, pages 354–359, 1998.
- [86] R. Horlings, D. Datcu, and L.J.M. Rothkrantz. Emotion recognition using brain activity. In *Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing*, page 6. ACM, 2008.
- [87] C.W. Hsu, C.C. Chang, C.J. Lin, et al. A practical guide to support vector classification, 2003.
- [88] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, and T.S. Huang. A study of non-frontal-view facial expressions recognition. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4, 2008.
- [89] C.L. Huang and Y.M. Huang. Facial expression recognition using model-based feature extraction and action parameters classification. *Journal of Visual Communication and Image Representation*, 8(3):278–290, 1997.

- [90] Q. Ji, P. Lan, and C. Looney. A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 36(5):862–875, 2006.
- [91] O.P. John, L.P. Naumann, and C.J. Soto. Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In *O.P. John, R.W. Robins and L.A. Pervin (Eds.) Handbook of Personality: theory and research*, pages 114–158, 2008.
- [92] M. Jones and P. Viola. Fast multi-view face detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Citeseer, 2003.
- [93] A. Kapoor, W. Bursleson, and R.W. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736, 2007.
- [94] J. Kim and E. Ande. Emotion recognition based on physiological changes in music listening. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2067–2083, 2008.
- [95] J.O. Kim, K.S. Seo, C.H. Chung, J. Hwang, and W. Lee. On Facial Expression Recognition Using the Virtual Image Masking for a Security System. *Lecture Notes in Computer Science*, pages 655–662, 2004.
- [96] S. Kimura and M. Yachida. Facial expression recognition and its degree estimation. In *1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997. Proceedings.*, pages 295–300, 1997.
- [97] H. Kobayashi and F. Hara. Facial interaction between animated 3D face robot and human beings. In *IEEE International Conference on Systems, Man, and Cybernetics. 'Computational Cybernetics and Simulation'.*, volume 4, 1997.
- [98] I. Kotsia and I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*, 16(1):172–187, 2007.

- [99] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato. Pose-invariant facial expression recognition using variable-intensity templates. *International Journal of Computer Vision*, 83(2):178–194, 2009.
- [100] O.W. Kwon, K. Chan, J. Hao, and T.W. Lee. Emotion recognition by speech signals. In *Eighth European Conference on Speech Communication and Technology*, 2003.
- [101] J. Lafferty. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, 2001.
- [102] C.C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 2011.
- [103] C.M. Lee and S.S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, 2005.
- [104] D. Lin, S. Yan, and X. Tang. Comparative study: face recognition on unspecific persons using linear subspace methods. In *IEEE International Conference on Image Processing, ICIP*, volume 3. IEEE, 2005.
- [105] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *Arxiv preprint arXiv:1009.5055*, 2010.
- [106] P. Litle and M. Zuckerman. Sensation seeking and music preferences. *Personality and Individual Differences*, 7(4):575–578, 1986.
- [107] G.C. Littlewort, M.S. Bartlett, and K. Lee. Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 15–21. ACM New York, USA, 2007.

- [108] S. Lucey, A.B. Ashraf, and J.F. Cohn. Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face. *Face Recognition, Delac*, pages 275–286, 2007.
- [109] MJ Lyons, J. Budynek, and S. Akamatsu. Automatic Classification of Single Facial Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI*, 21(12):1357–1362, 1999.
- [110] A. Maalej, B. Ben Amor, M. Daoudi, A. Srivastava, and S. Berretti. Local 3d shape analysis for facial expression recognition. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4129–4132. IEEE, 2010.
- [111] F. Mairesse and M. Walker. Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 543–548. Citeseer, 2006.
- [112] F. Mairesse, M.A. Walker, M.R. Mehl, and R.K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500, 2007.
- [113] R.P. Mattick and J.C. Clarke. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behaviour Research and Therapy*, 36(4):455–470, 1998.
- [114] P. Maurel, ENS Odyssee, A. McGonigal, F. Marseille, R. Keriven, F. Paris-Est, and P. Chauvel. 3D model fitting for facial expression analysis under uncontrolled imaging conditions. In *19th International Conference on Pattern Recognition, ICPR, Tampa, US*, 2008.
- [115] R.R. McCrae and P.T. Costa. Reinterpreting the Myers-Briggs Type Indicator from the perspective of the five-factor model of personality. *Journal of personality*, 57(1):17–40, 1989.
- [116] R.R. McCrae and O.P. John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

- [117] P. Melville, W. Gryc, and R.D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284. ACM New York, USA, 2009.
- [118] G. Mohammadi, A. Vinciarelli, and M. Mortillaro. The voice of personality: mapping nonverbal vocal behavior into trait attributions. In *Proceedings of the 2nd international workshop on Social signal processing*, pages 17–20. ACM, 2010.
- [119] S. Moore and R. Bowden. The effects of Pose on Facial Expression Recognition. In *British Machine Vision Conference*, pages 1–11, 2009.
- [120] Y. Moses, D. Reynard, and A. Blake. Determining facial expressions in real time. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 296–301, 1995.
- [121] I. Mpiperis, S. Malassiotis, V. Petridis, and M.G. Strintzis. 3D facial expression recognition using swarm intelligence. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2008*.
- [122] I. Mpiperis, S. Malassiotis, and MG Strintzis. Bilinear Models for 3-D Face and Facial Expression Recognition. *IEEE Transactions on Information Forensics and Security*, 3(3):498–511, 2008.
- [123] I.B. Myers and P.B. Myers. *Gifts differing: Understanding personality type*. Davies-Black Publishing, 1995.
- [124] T.D. Nguyen and S. Ranganath. Tracking facial features under occlusions and recognizing facial expressions in sign language. In *Proc. Conf. Face and Gesture Recognition, FG08*, pages 1–7, 2008.
- [125] W.T. Norman. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, 66(6):574, 1963.

- [126] T.L. Nwe, S.W. Foo, and L.C. De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.
- [127] J. Oberlander and S. Nowson. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 627–634. Association for Computational Linguistics, 2006.
- [128] A. Ortony, G.L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge Univ Pr, 1990.
- [129] T. Otsuka and J. Ohya. Spotting segments displaying facial expression from image sequences using HMM. In *Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, page 442. IEEE Computer Society Washington, DC, USA, 1998.
- [130] C. Padgett and G.W. Cottrell. Representing face images for emotion classification. *Advances in neural information processing systems*, pages 894–900, 1997.
- [131] G. Pan, S. Han, Z. Wu, and Y. Zhang. Removal of 3d facial expressions: A learning-based approach. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*. IEEE, 2010.
- [132] M. Pantic and I. Patras. Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments From Face Profile Image Sequences. *IEEE Transactions on Systems, Man, and CyberneticsPart B: Cybernetics*, 36(2):433–449, 2006.
- [133] M. Pantic and L.J.M. Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18(11):881–905, 2000.
- [134] M. Pantic and LJM Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 34(3):1449–1461, 2004.

- [135] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, ICME. IEEE International Conference on*. IEEE, 2005.
- [136] S. Park and D. Kim. Spontaneous facial expression classification with facial motion vectors. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.
- [137] W. Pavot, E. Diener, and F. Fujita. Extraversion and happiness. *Personality and Individual Differences*, 11(12):1299–1306, 1990.
- [138] D.M. Pedersen. Personality correlates of privacy. *The Journal of Psychology*, 112(1):11–14, 1982.
- [139] V.A. Petrushin. How well can people and computers recognize emotions in speech? In *Proceedings of the AAAI Fall Symposium*, pages 141–145, 1998.
- [140] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60. ACM, 2008.
- [141] R.W. Picard. *Affective computing*. The MIT press, 2000.
- [142] R. Plutchik. *The psychology and biology of emotion*. Harper Collins, 1994.
- [143] L. Rabiner and B.H. Juang. *Fundamentals of speech recognition*. 1993.
- [144] B. Rammstedt and O.P. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203–212, 2007.
- [145] P.J. Rentfrow and S.D. Gosling. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of personality and social psychology*, 84(6):1236–1256, 2003.

- [146] G.I. Roisman, J.L. Tsai, and K.H.S. Chiang. The emotional integration of childhood experience: physiological, facial expressive, and self-reported emotional response during the adult attachment interview. *Developmental Psychology*, 40(5):776, 2004.
- [147] M. Rosato, X. Chen, and L. Yin. Automatic Registration of Vertex Correspondences for 3D Facial Expression Analysis. In *2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, 2008. BTAS 2008*, pages 1–7, 2008.
- [148] M. Rosenblum, Y. Yacoob, and LS Davis. Human expression recognition from motion using a radial basisfunction network architecture. *IEEE Transactions on Neural Networks*, 7(5):1121–1138, 1996.
- [149] J.B. Rotter. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General & Applied*, 1966.
- [150] P. Rozin and A.B. Cohen. High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion*, 3(1):68–75, 2003.
- [151] K.R. Scherer. Adding the affective dimension: A new look in speech analysis and synthesis. In *Proc. International Conf. on Spoken Language Processing*, pages 1808–1811, 1996.
- [152] K.R. Scherer and G. Ceschi. Lost luggage: A field study of emotion–antecedent appraisal. *Motivation and Emotion*, 21(3):211–235, 1997.
- [153] B. Schuller, A. Batliner, S. Steidl, and D. Seppi. Emotion recognition from speech: putting ASR in the loop. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, 2009*.
- [154] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu. Being bored? Recognising natural interest by

extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27(12):1760–1774, 2009.

- [155] N. Sebe, I. Cohen, T. Gevers, and T.S. Huang. Emotion recognition based on joint visual and audio cues. In *Proceedings of the 18th International Conference on Pattern Recognition*, pages 1136–1139, 2006.
- [156] T. Sha, M. Song, J. Bu, C. Chen, and D. Tao. Feature level analysis for 3d facial expression recognition. *Neurocomputing*, 2011.
- [157] L. Shang and KP Chan. Nonparametric Discriminant HMM and Application to Facial Expression Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Citeseer, 2009.
- [158] T. Simon, M.H. Nguyen, F. De La Torre, and J.F. Cohn. Action unit detection with segment-based SVMs. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2737–2744. IEEE, 2010.
- [159] J. Sivic, M. Everingham, and A. Zisserman. Who are you?—Learning person specific classifiers from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1145–1152, 2009.
- [160] M. Slaney and G. McRoberts. Baby ears: a recognition system for affective vocalizations. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 2, pages 985–988. IEEE, 1998.
- [161] T. Sobol-Shikler and P. Robinson. Classification of complex information: Inference of co-occurring affective states from their expressions in speech. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7):1284–1297, 2010.
- [162] M. Song, J. Bu, C. Chen, and N. Li. Audio-visual based emotion recognition—a new approach. In *Computer Vision and Pattern Recognition, CVPR. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2. IEEE, 2004.

- [163] H. Soyel and H. Demirel. Facial expression recognition using 3d facial feature distances. In *ICIAR 2007, Lecture Notes in Computer Science*, volume 4633, pages 831–838. Springer Berlin / Heidelberg, 2007.
- [164] H. Soyel and H. Demirel. Optimal feature selection for 3d facial expression recognition using coarse-to-fine classification. *Turkish Journal of Electrical Engineering and Computer Sciences*, 18(6):1031–1040, 2010.
- [165] S. Srivastava, O.P. John, S.D. Gosling, and J. Potter. Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, 84(5):1041–1053, 2003.
- [166] G. Stylianou and A. Lanitis. Image based 3d face reconstruction: a survey. *International Journal of Image and Graphics*, 9(2):217–250, 2009.
- [167] Y. Sun and L. Yin. Facial Expression Recognition Based on 3D Dynamic Range Model Sequences. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, pages 58–71. Springer-Verlag Berlin, Heidelberg, 2008.
- [168] J. Sung and D. Kim. Pose-Robust Facial Expression Recognition Using View-Based 2D+3D AAM. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 38(4):852–866, 2008.
- [169] J. Sung, S. Lee, and D. Kim. A real-time facial expression recognition using the STAAM. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, 2006.
- [170] M. Suwa, N. Sugie, and K. Fujimora. A preliminary note on pattern recognition of human emotional expression. In *International Joint Conference on Pattern Recognition*, 1978.
- [171] Jeffrey Cohn Takeo Kanade and Ying-Li Tian. Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, pages 46 – 53, March 2000.

- [172] H. Tang and T.S. Huang. 3D facial expression recognition based on automatically selected features. In *CVPR Workshops 2008.*, pages 1–8, 2008.
- [173] H. Tang and T.S. Huang. 3D facial expression recognition based on automatically selected features. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPR Workshops 2008. IEEE Computer Society Conference on*, pages 1–8, 2008.
- [174] H. Tang and T.S. Huang. 3d facial expression recognition based on properties of line segments connecting facial feature points. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.
- [175] A. Tawari and M.M. Trivedi. Speech emotion analysis in noisy real-world environment. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4605–4608. IEEE, 2010.
- [176] CE Thomaz, DF Gillies, and RQ Feitosa. A new covariance estimate for Bayesian classifiers in biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(2):214–223, 2004.
- [177] Y.I. Tian, T. Kanade, and JF Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.
- [178] Y. Tong, W. Liao, Z. Xue, and Q. Ji. A unified probabilistic framework for facial activity modeling and understanding. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.
- [179] K. Toutanova, D. Klein, C.D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

- [180] K. Toutanova and C.D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *EMNLP/VLC 2000*, pages 63–70, 2000.
- [181] W. Tov and E. Diener. The well-being of nations: Linking together trust, cooperation, and democracy. *The Science of Well-Being*, pages 155–173, 2009.
- [182] L.J. Trainor, C.M. Austin, and R.N. Desjardins. Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science*, 11(3):188, 2000.
- [183] E.C. Tupes and R.E. Christal. Recurrent personality factors based on trait ratings. *Journal of Personality*, 60(2):225–251, 1992.
- [184] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.
- [185] P.D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 417–424, 2002.
- [186] M.F. Valstar, M. Pantic, Z. Ambadar, and J.F. Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 162–170. ACM New York, USA, 2006.
- [187] M.F. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection from face video. In *2004 IEEE International Conference on Systems, Man and Cybernetics*, volume 1, 2004.
- [188] YV Venkatesh, A.A. Kassim, and OV Ramana Murthy. A novel approach to classification of facial expressions from 3d-mesh datasets using modified pca. *Pattern Recognition Letters*, 30(12):1128–1137, 2009.
- [189] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

- [190] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple. In *Proc. IEEE CVPR 2001*.
- [191] N. Vretos, N. Nikolaidis, and I. Pitas. A model-based facial expression recognition algorithm using Principal Components Analysis. In *Image Processing (ICIP), 2009 16th International Conference on*, pages 3301–3304. IEEE, 2010.
- [192] J. Wagner, E. Andre, F. Lingenfelter, J. Kim, and T. Vogt. Exploring fusion methods for multimodal emotion recognition with missing data. *Affective Computing, IEEE Transactions on*, (99):1–1, 2011.
- [193] J. Wang, L. Yin, X. Wei, and Y. Sun. 3D facial expression recognition based on primitive surface feature distribution. In *Proc. Conf. Computer Vision and Pattern Recognition, CVPR*, volume 2, pages 1399–1406, 2006.
- [194] M. Wang, Y. Iwai, and M. Yachida. Expression recognition from time-sequential facial images by use of expression change model. In *Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. Proceedings*, pages 324–329, 1998.
- [195] P. Wang, C. Kohler, F. Barrett, R. Gur, and R. Verma. Quantifying Facial Expression Abnormality in Schizophrenia by Combining 2D and 3D Features. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–8, 2007.
- [196] Y. Wang and L. Guan. Recognizing human emotion from audiovisual information. In *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, volume 2. IEEE, 2005.
- [197] Sholom M. Weiss and N. Indurkha. *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers, 1998.
- [198] C. Whissell. The dictionary of affect in language. *Emotion: Theory, research, and experience*, 4:113–131, 1989.

- [199] J. Whitehill, M. Bartlett, and J. Movellan. Automatic facial expression recognition for intelligent tutoring systems. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPR Workshops 2008*, pages 1–6, 2008.
- [200] J. Whitehill and C.W. Omlin. Haar features for FACS AU recognition. In *Proc. IEEE Intl Conf. Face and Gesture Recognition*, 2006.
- [201] Jacob Whitehill, Marian Bartlett, and Javier Movellan. Measuring the perceived difficulty of a lecture using automatic facial expression recognition. In *ITS '08: Proceedings of the 9th international conference on Intelligent Tutoring Systems*, pages 668–670, 2008.
- [202] C. Windle. Further studies of test-retest effect on personality questionnaires. *Educational and Psychological Measurement*, 15(3):246, 1955.
- [203] C.H. Wu, Z.J. Chuang, and Y.C. Lin. Emotion recognition from text using semantic labels and separable mixture models. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(2):183, 2006.
- [204] Chung-Hsien Wu and Wei-Bin Liang. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *Affective Computing, IEEE Transactions on*, 2(1):10–21, 2011.
- [205] T.F. Wu, C.J. Lin, and R.C. Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:1005, 2004.
- [206] T. Yabui, Y. Kenmochi, and K. Kotani. Facial expression analysis from 3D range images; comparison with the analysis from 2D images and their integration. In *International Conference on Image Processing*, pages 879–882, 2003.
- [207] Y. Yacoob and L.S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642, 1996.

- [208] P. Yang, Q. Liu, and D.N. Metaxas. Exploring facial expressions with compositional features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2638–2644. IEEE, 2010.
- [209] M. Yeasin, B. Bullot, and R. Sharma. Recognition of facial expressions and measurement of levels of interest from video. *IEEE Transactions on Multimedia*, 8(3):500–508, 2006.
- [210] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.
- [211] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3d facial expression database for facial behavior research. In *FGR '06: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 211–216. IEEE Computer Society, 2006.
- [212] M. Yoneyama, A. Ohtake, Y. Iwano, and K. Shirai. Facial expressions recognition using discrete Hopfield neuralnetworks. In *Image Processing, 1997. Proceedings., International Conference on*, volume 1, 1997.
- [213] J. Schooler Zara Ambadar and Jeffrey Cohn. *Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions*, 2005.
- [214] G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space speaks: towards socially and personality aware visual surveillance. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, pages 37–42. ACM, 2010.
- [215] Z. Zeng, Y. Hu, M. Liu, Y. Fu, and T.S. Huang. Training combination strategy of multi-stream fused hidden markov model for audio-visual affect recognition. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 65–68. ACM, 2006.

- [216] Z. Zeng, Y. Hu, G.I. Roisman, Z. Wen, Y. Fu, and T.S. Huang. Audio-visual spontaneous emotion recognition. *Lecture Notes in Computer Science*, 4451:72, 2007.
- [217] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T.S. Huang, D. Roth, and S. Levinson. Bimodal HCI-related affect recognition. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 137–143. ACM New York, USA, 2004.
- [218] Z.H. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. 31(1):39–58, January 2009.
- [219] Y. Zhan and D. Shen. Design efficient support vector machine for fast classification. *Pattern Recognition*, 38(1):157–161, 2005.
- [220] L. Zhang, M. Song, N. Li, J. Bu, and C. Chen. Feature selection for fast speech emotion recognition. In *Proceedings of the seventeen ACM international conference on Multimedia*, pages 753–756. ACM, 2009.
- [221] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, 27(5):699–714, 2005.
- [222] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and Gabor-wavelets-based facialexpression recognition using multi-layer perceptron. In *Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. Proceedings*, pages 454–459, 1998.
- [223] X. Zhao, D. Huang, E. Dellandréa, and L. Chen. Automatic 3d facial expression recognition based on a bayesian belief net and a statistical facial feature model. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3724–3727. IEEE, 2010.

- [224] W. Zheng, H. Tang, Z Lin, and T. S. Huang. A Novel Approach to Expression Recognition from Non-Frontal Face Images. In *IEEE International Conference on Computer Vision*, 2009.
- [225] W. Zheng, H. Tang, Z. Lin, and T.S. Huang. Emotion recognition from arbitrary view facial images. In *Proceedings of the 11th European conference on Computer vision: Part VI*, pages 490–503. Springer-Verlag, 2010.
- [226] Guangyu Zhu, Shuicheng Yan, and Yi Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ACM Multimedia*, 2010.