

University of Memphis

University of Memphis Digital Commons

Electronic Theses and Dissertations

7-20-2011

A Spatio-Temporal Probabilistic Framework for Dividing and Predicting Facial Action Units

A K M Mahbubur Rahman

Follow this and additional works at: <https://digitalcommons.memphis.edu/etd>

Recommended Citation

Rahman, A K M Mahbubur, "A Spatio-Temporal Probabilistic Framework for Dividing and Predicting Facial Action Units" (2011). *Electronic Theses and Dissertations*. 267.
<https://digitalcommons.memphis.edu/etd/267>

This Thesis is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.

To the University Council:

The Thesis Committee for A K M Mahbubur Rahman certifies that this is the final approved version of the following electronic thesis: "A Spatio -Temporal Probabilistic Framework For Dividing And Predicting Facial Action Units."

Mohammed Yeasin, Ph.D.
Major Professor

We have read this thesis and recommend
its acceptance:

Khan M. Iftekharuddin, Ph.D.

Xianguan Hu, Ph.D.

Accepted for the Graduate Council:

Karen D. Weddle-West, Ph.D.
Vice Provost for Graduate Programs

A SPATIO-TEMPORAL PROBABILISTIC FRAMEWORK
FOR DIVIDING AND PREDICTING FACIAL ACTION UNITS

by

A K M Mahbubur Rahman

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Electrical and Computer Engineering

The University of Memphis

August 2011

Copyright ©2011 A K M Mahbubur Rahman
All rights reserved

ACKNOWLEDGMENTS

I would like to thank Dr. Mohammed Yeasin for his constant support and guidance throughout the period of my graduate studies in United States. His assistance made my graduate studies possible. His valuable help has been extremely appreciated. I would also like to thank Dr. Khan M. Iftekharuddin and Dr. Xiangen Hu for being in my thesis committee.

This research was partially supported by grant NSF-IIS-0746790. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and not necessarily reflect the views of the funding institution. (Portions of) the research in this paper uses the MMI-Facial Expression Database collected by Valstar and Pantic.

I would like to thank my parents for always inspiring and supporting me. Without their support I would not be able to come along this far in my life. This thesis is dedicated to them.

ABSTRACT

Rahman, A K M Mahbubur. M.S. Electrical and Computer Engineering. The University of Memphis. August 2011. A Spatio-Temporal Probabilistic Framework for Dividing and Predicting Facial Action Units. Major Professor: Mohammed Yeasin, Ph.D.

This thesis proposed a probabilistic approach to divide the Facial Action Units (AUs) based on the physiological relations and their strengths among the facial muscle groups. The physiological relations and their strengths were captured using a Static Bayesian Network (SBN) from given databases. A data driven spatio-temporal probabilistic scoring function was introduced to divide the AUs into : (i) frequently occurred and strongly connected AUs (FSAUs) and (ii) infrequently occurred and weakly connected AUs (IWAUs). In addition, a Dynamic Bayesian Network (DBN) based predictive mechanism was implemented to predict the IWAUs from FSAUs. The combined spatio-temporal modeling enabled a framework to predict a full set of AUs in real-time. Empirical analyses were performed to illustrate the efficacy and utility of the proposed approach. Four different datasets of varying degrees of complexity and diversity were used for performance validation and perturbation analysis. Empirical results suggest that the IWAUs can be robustly predicted from the FSAUs in real-time and was found to be robust against noise.

TABLE OF CONTENTS

| Chapter | Page |
|---|------|
| 1 Introduction | |
| Overview | 1 |
| Research Challenges | 2 |
| Proposed Method | 3 |
| Technical Definition of the Proposed Method | 3 |
| Implementation Framework for the Proposed Method | 4 |
| Definitions | 5 |
| Major Contributions | 6 |
| Outline | 7 |
| 2 Literature Review | |
| Background and Related Works | 8 |
| Limitations | 9 |
| 3 FACS and AU Relations | |
| Facial Action Coding System(FACS) | 11 |
| AU Relations | 11 |
| 4 Modeling AU Relations | |
| Modeling AU Relations with Static Bayesian Network(SBN) | 16 |
| Structure Learning of SBN | 16 |
| Score of a SBN | 17 |
| Notations | 18 |
| 5 Identify Strong AU Relations | |
| Proposition | 20 |
| Explanation of the Proposition | 20 |
| 6 Finding Significant Subset | |
| Strong Relations vs. Weak Relations | 25 |
| FSAUs | 26 |
| Parameter Learning for Final SBN | 26 |
| IWAU Prediction | 27 |
| 7 Modeling Temporal Relationship between AUs using Dynamic Bayesian Network | |
| Temporal Evolution with AUs | 29 |
| Modeling Temporal Evolution of AUs with DBN | 30 |
| Parameter Learning | 32 |
| Prediction of IWAUs using DBN | 32 |
| 8 Data Annotation | |
| Data Collection | 34 |
| Data Annotation (FACS Annotation) | 34 |
| Scoring Procedure | 36 |
| Data Treatment and Analysis | 36 |
| 9 Experimental Results | |
| Databases | 39 |
| Utility of Scoring Algorithm in Logical Division of AUs | 40 |
| IWAUs Predictions | 40 |

| | |
|---|----|
| Empirical Analysis using Posed Expressions | 41 |
| Generalization using Mixed Expressions | 42 |
| Analysis with Natural Expressions | 43 |
| Performance Analysis with Perturbation | 44 |
| Noise Analysis in Predicting IWAUs | 44 |
| Noise Analysis in Facial Expression Recognition | 45 |
| 10 Conclusion | 50 |
| Bibliography | |

LIST OF FIGURES

| | | |
|------|--|----|
| 1.1 | Proposed Framework | 4 |
| 2.1 | Research trends in AU Recognition | 9 |
| 3.1 | Example of Action Units [1] | 12 |
| 3.2 | Example of Action Units [1] | 13 |
| 3.3 | Example of Action Units [1] | 14 |
| 3.4 | Muscles in Face | 15 |
| 4.1 | Example BN | 18 |
| 5.1 | Possible Events for Add operation | 23 |
| 5.2 | Possible Events for Reverse operation | 23 |
| 5.3 | Possible Events for Delete operation | 24 |
| 6.1 | Rate of Change of scores | 25 |
| 6.2 | Strong relations between AUs | 26 |
| 7.1 | Five Slice DBN Example | 29 |
| 7.2 | Two Slice DBN | 31 |
| 9.1 | Experiment Setups | 38 |
| 9.2 | 2nd derivative of scores | 39 |
| 9.3 | ROCA for 3-fold cross validation with CK+ | 42 |
| 9.4 | Recognition Performance of IWAUs for MMI dataset | 43 |
| 9.5 | Performance evaluation for IWAUs in spontaneous datasets | 44 |
| 9.6 | Noise Tolerance Score | 45 |
| 9.7 | Performance of IWAU Prediction against Noise | 46 |
| 9.8 | Expression Bayesian Net | 47 |
| 9.9 | Performance of Facial Expression Recognition against Noise | 48 |
| 9.10 | Average Performance of Facial Expression Recognition | 49 |

Chapter 1

Introduction

1.1 Overview

In person-to-person nonverbal interaction, human psychology enables people to take appropriate actions through perceived emotional experiences. Emotion perception depends on human behavior, which is composed of facial expression, voice, action, and movement of different body parts. In human to human communication, faces are one of the most important input attributes to the cognitive procedure which governs our emotional and affective states. Facial behavior can provide information about: (i) affective state, (ii) cognitive activity, (iii) temperament and personality, (iv) truthfulness and (v) psychopathology just to name a few. Actually facial behavior is characterized through facial expressions. Although there are six basic facial expressions described by Ekman et al. [2], spontaneous behavior of a human face can range up to thousands of intensities over those six basic expressions. Moreover, psychologists are interested in complex emotional states of the human mind rather than six basic emotions. Therefore, one of the grand challenges in emotion research is to make artificial agents and machines capable of understanding the mechanisms of how human beings interact with the world and each other. In fact, human-like communication is desirable between man and computer agents. For example, in e-learning, autonomous agents could provide formative feedback based on assessment and identification of users' emotion and affective states using state of the art facial image analysis. In addition, functional relevance of learning with emotion, affective state, and their interplay can be used to enhance learners' experience as well as the utility of existing MetaTutors such as Auto-Tutor [3], Meta-cognitive Tutors [4].

Autonomous analysis and synthesis of facial expressions and emotions are emerging issues in affective computing and agent-human communication. Capturing facial features and measuring their appearances are the core discipline of facial expression recognition research. Though a human observer can easily perceive the changes in facial features quite easily, objective definition of each facial feature is necessary for a machine to perceive automatically. The mostly used definition of facial features is Facial Action Coding System(FACS). In FACS, Ekman et al.[5] described facial behavior using 32 Facial Action Units(AUs). AUs are the lexicon of the facial behavior. Each AU is defined by movement of a certain set of muscle movements. Quantitative measurements are also incorporated for each AU. That is, various combinations of AUs are responsible for different facial expressions.

1.2 Research Challenges

Facial action units (AUs) defined in Facial Action Coding System (FACS) [2] have been widely used in recognizing facial expressions (i.e., [6]), emotions (i.e., [7]), and affective states [8] to compute description of facial behavior. However, real-time tracking and spotting of continuous emotions from video require robust identification of all or a majority of the AUs. Despite the recent surge of methods, robust and real-time recognition of AUs remains challenging due to inaccuracies in measurements of subtle facial deformation, pose, and out of plane head movements. State-of-the-art methods in recognition of AUs are limited to subsets of *posed expressions*. They are inadequate for recognition of spontaneous facial expressions, modeling blended emotions and are also unsuitable for real-time applications. For example, subsets of AUs with size 14, 17, 18, and 20 were recognized by Tong *et al.* [6], Lucey *et al.* [9], Zhang *et al.* [10], and Bartlett *et al.* [11], respectively. The choices of subsets of AUs in the reported literatures were done mostly by *ad hoc*-principles for a variety of reasons (that include but are not limited to): (i) skewed and non-uniform distribution of “representative examples” for AUs in existing emotion databases [9, 12] and databases containing natural emotions, (ii) trade-offs between higher resolution image and uncertainty in measuring spontaneous facial movements from visual data, (iii) lack of systematic approaches to determine significant subset of AUs even in the context of a niche application to characterize facial behavior, and (iv) lack of framework for real-time processing of full set of AUs. In addition, methods based on *heuristics* and *ad hoc* rules may preclude important relationships between AUs available as evidences in the data.

Furthermore, continuous-emotion analysis necessitates real time AU recognition from a video of a particular subject. AU recognition consists of two successive processes. Perception of the facial features is usually done by computer vision techniques. Classifications are executed to determine the presence of each AU based on perceived facial features. However, modeling the spontaneous variation of facial features is the key challenge for real time AU recognition due to the very high dimension of feature space. Different AUs require capturing different sets of facial features and different decision algorithms. Since computer vision techniques are responsible for capturing facial features from real time video, large sets of AUs increase the computer-vision complexities. Additionally, a large sets of AUs requires complex classification algorithms that cause enormous mathematical computations. For example, if one binary classifier is associated to recognize one AU, the number of required binary classifiers for the AU recognition system is equal to the size of AU set. Consequently, training each binary classifier with appropriate training examples creates overhead of the process.

Both complexities described above are the main sources of limitation for processing speed. Reducing the number of AUs will help to improve both complexities. However, reducing the number of AUs without any impact on continuous emotion recognition creates a vast window of the recent research trends.

1.3 Proposed Method

Some of the above mentioned problems can be addressed by dividing the AUs intuitively into two subsets: frequently occurred and strongly connected AUs (FSAUs) and infrequently occurred and weakly connected AUs (IWAUs). By exploiting the physiological constraints, the IWAUs can be predicted from FSAUs in real time using spatio-temporal relations of the AUs. To understand the AU relations and how to use them for reliable prediction, let us consider a few illustrative examples. Muscle group of *Corrugator supercilii*, *Levator labii superioris alaeque nasi*, *Buccinator*, and *Depressor Anguli Oris* produce the constraints for activation of AU 4 (Brow Lowerer), AU 9 (Nose Wrinkler), AU 14 (Dimpler), and AU 15 (Lip Corner Depressor) respectively. Among them, recognition of AU 14 is more error prone due to higher uncertainties in measuring muscle movements from the visual cue [11][13] while AU 4 and AU 15 are relatively easy to recognize. The physiological constraints between AUs 4, 9, 14, and 15 can be used to reliably predict AU 14 without directly recognizing it. In addition to physiological constraints, AUs are evolved over time when facial expressions continue from neutral to apex and then relax. By modeling both the spatial and temporal relations, it is possible to rely on a smaller significant subset of AUs to infer the occurrences of other AUs. Moreover, spatio-temporal model of the physical constraints can help to reduce the effect of inaccuracies in FSAUs. Since, none of the AU recognition systems are fully robust against noise, error in the FSAUs are expected to be minimized when predicting IWAUs.

1.3.1 Technical Definition of the Proposed Method

This thesis introduces a scoring process to determine the FSAUs and IWAUs and exploits the spatio-temporal evolution of AUs to predict the IWAUs from the FSAUs in real-time. The goal is to divide the full set of AUs (U) into two subsets: (i) subset $P = \{p_1, p_2, \dots, p_n\}$ containing FSAUs and (ii) subset $S = \{s_1, s_2, \dots, s_{m-n}\}$ containing IWAUs. The key objective is to keep the size of P as small as possible while maintaining robustness in inferring the set S . The spatial relations and physiological constraints among the AUs were modeled and synthesized with Static Bayesian Network (SBN) while their temporal evolution is modeled using Dynamic Bayesian Network (DBN). The scoring mechanism

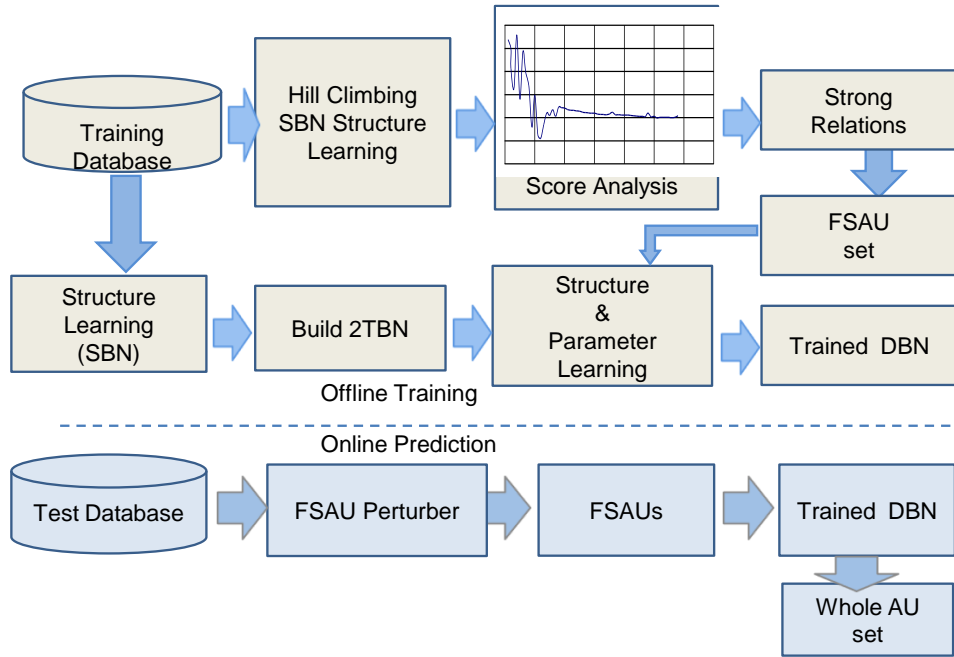


Figure 1.1: Proposed Framework

has been defined using an optimized SBN that captures the strength of AU relations learned from the database. Figure 1.1 illustrates the proposed concept in graphical form. Based on the scores obtained using optimal SBN structure, AUs were logically divided into FSAUs and IWAUs. In addition, a framework was developed for real-time recognition and prediction of a large number of AUs. IWAUs act as auxiliary AUs for modeling the facial expressions.

1.3.2 Implementation Framework for the Proposed Method

Description of the framework shown in figure 1.1 has been illustrated here. Upper portion illustrates the training phase while lower phase describes testing procedure.

- **Training dataset** has been used for finding the FSAUs and for finding the correct structure of the AU relations.
- **“Hill Climbing SBN Structure Learning”** generates the intermediate scores during modified hill climbing learning technique.
- These scores are analyzed in the **“Score Analysis”** to calculate how the scores change over time.

- After analyzing the scores, strong relations are identified through the proposition described in the chapter 5. AUs involved in the strong relations are defined as **FSAUs**.
- The most compatible AU relation structure has been identified in “**Structure Learning**” with couple of random starts.
- After finding the SBN structure, Dynamic Bayesian Network has been built through most popular **2TBN** (Two Slice Bayesian Network).
- In **Structure & Parameter Learning**, the interrelationships between SBN slices are learned using temporal training data. The parameter distributions are also learned here. At this block, the FSAUs are marked with observed nodes since these are the available nodes during online prediction.

In the online prediction stage, the **FSAU Perturber** takes the ground truth FSAUs from the testing datasets and adds noise up to some specific amount. Noisy FSAUs are transmit to the **Trained DBN** as inputs. The FSAUs are working as observed nodes in the **Trained DBN** and IWAUs are defined as hidden nodes. Inference algorithms are used to infer the probability distributions of the IWAUs. Based on the probability distributions, IWAUs are predicted and transferred to the output of the system.

1.3.3 Definitions

We define FSAUs as the significant Action units that play more important role in facial expressions. Thats why they are defined as Frequently occurred and Strongly Connected. Strong connections between significant AUs are mostly responsible in modeling facial behavior. Using a linguistic analogy, one can define the AU tree and the root based on frequently occurred strong relations that are present in the expressions database(s). The FSAUs are frequently occurred and strongly connected for the most of the data points in the dataset. Thus, recognizing FSAUs using computer vision is sufficient enough to infer the other set IWAUs. The reason is that the strongly connected AUs are the most important ones in the face and the others are insignificant. So, other can be easily predicted from the significant ones. IWAUs are Infrequently occurred and weakly connected AU. These AUs are weakly connected with each other and not much frequent in the facial expressions. However, to predict the wide range of emotional states, these are also necessary to disambiguate the emotion prediction.

1.4 Major Contributions

Adhoc selection of AUs [6], [13],[10] bypass the problem of finding of significant AUs. A number of closely related works (i.e., [10, 6, 14]) use Bayesian analysis with frequent AUs. The proposed approach is significantly different from these techniques in a number of ways. The key difference is the use of AU relations obtained using the “scoring mechanism” as opposed to “frequency” that clearly separates this work from reported related works. Since the combination of AUs and their temporal evolutions are mostly responsible for spontaneous facial behavior, the proposed solution incorporates a spatio-temporal statistical approach to capture the *AU relations* from the evidences. The scoring process defined using the SBN is novel (to the best of our knowledge) though it uses existing tools and was found to be very robust. Moreover, the prediction performance of IWAUs outperformed the contemporary related works.

Additionally, perturbation (noise) analysis was performed at AU level as well as at the level of categorical emotion recognition. At first, robustness in predicting IWAUs against perturbation in the FSAUs were analyzed. It was observed that the proposed approach is stable in the presence of varying degrees of perturbation in the FSAUs while maintaining reasonable IWAU recognition performance. However, different IWAUs have different level of noise tolerance. Empirical analyses were performed to determine the range of allowable noise in FSAUs. None of the recent research works have identified the noise tolerance capabilities of the AUs. Also, empirical analysis was performed to characterize the effect of perturbation at FSAUs on the robust prediction of facial expressions. The key contributions of this thesis are summarized as follows:

1. Identifying the needs for logical division of AUs based on physiological relationships among AUs computed from evidences and their impact on realtime recognition of AUs and emotion.
2. Modeling spatio-temporal relationship between AUs.
3. Development of a spatio-temporal probabilistic scoring algorithm to divide the AUs into FSAUs and IWAUs based on the relations among the AUs.
4. Designing and developing a DBN based framework to predict IWAUs from the FSAUs in real time.
5. Analyzing the noise tolerance capability of IWAUs when FSAUs are noisy.
6. Building bayesian network based framework for recognition of six basic expressions.
7. Studying the noise tolerance capability in facial expression recognition.

1.5 Outline

The rest of the thesis is organized as follows:

Chapter 2 explores the available literatures for automatic AU recognition and modeling to provide the research context. Chapter 3 gives an overview of the FACS and brief description about AU relationships through muscle connections. Chapter 4 describes the mathematical description to model AU relations and the necessity to find the strong relations among AUs. Chapter 5 describes the mathematical derivation of the proposition to find strong relations between AUs. Chapter 6 illustrates the procedure to find out the significant AUs (FSAUs). Chapter 7 introduces the technical details of temporal modeling of the AU relations. Chapter 8 describes data annotation process of Emotion Elicitation dataset. Chapter 9 presents the experiment setups and the results with detail discussion. Chapter 10 concludes the thesis with future plans.

Chapter 2

Literature Review

Research works of AU recognition can be divided into two categories. First category includes the literatures that involve computer vision techniques to extract facial features that are supplied to binary classifiers while the other category of the research works describes AU recognition using spatial and temporal relations between AUs. Figure 2.1 shows the research trends in automatic AU recognition.

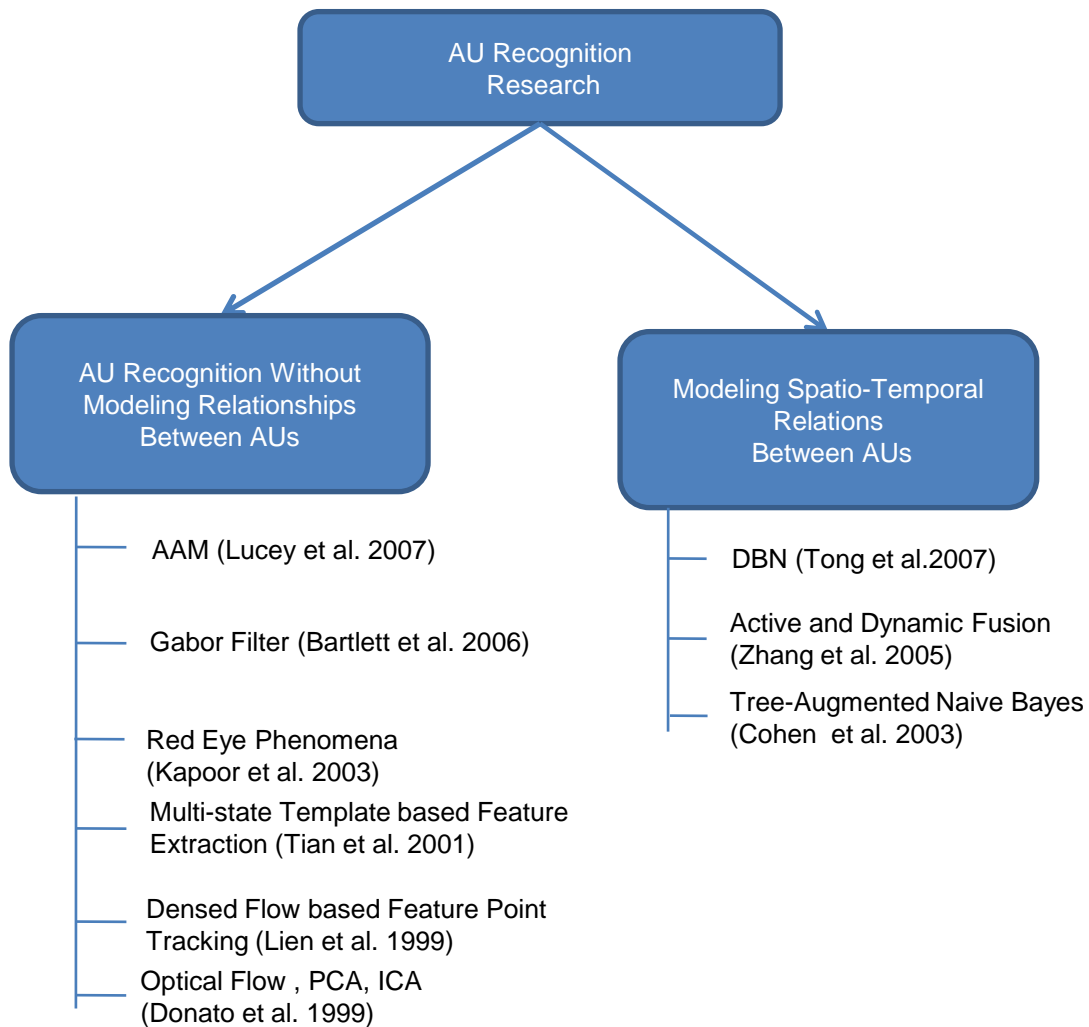


Figure 2.1: Research trends in AU Recognition

2.1 Background and Related Works

Tian et al. [15] utilized a multi-state template based feature extraction while using a neural network to recognize 16 AUs. Lien used dense flow based feature point tracking and edge extraction to detect and track facial action units in [16]. Donato et al. [17] compared AU detection systems based on several feature extraction techniques like optical flow, Principal Component Analysis (PCA), Independent Component Analysis (ICA), Gabor Wavelet based local features etc. However, their technique required manual alignment of images while the images have to be free from head movement. Bartlett et al. [18] used optical flow based local feature analysis. In later development [19], they provided the improved recognition system for 20 AUs. Gabor wavelets are used to extract features from frames using holistic approach while 20 separate adaboost binary classifiers are used. Kapoor et al. [20] extracted features utilizing a robust algorithm to detect red eye phenomena in addition to the shape of eyes and eyebrows. Lucey et al. [13] used different computer vision techniques wherein Active Appearance Model (AAM) has been used to extract features that consists of shape and appearance information. 15 AUs are subject to be recognized in this work.

Recent research works have analyzed the spatio-temporal relations between AUs to improve the AU recognition. In the work in [6], relations between 14 AUs are modeled with Static Bayesian Net (SBN) to formulate static relations. Dynamic Bayesian Network (DBN) has been exploited to model temporal evolution of targeted AUs. The computer vision technique involves multi-resolution based gabor-wavelets for feature extractions. Then resulting features are fed to 14 individual adaboost classifiers. The research group of Cohen et al. [21] uses Gaussian Tree-Augmented Naive Bayes (TAN) to learn the dependencies among different facial motion features in order to classify facial expressions. However, the TAN is not suitable to handle complex relationships such as hierarchical dependencies between facial features, or temporal changes. Other researchers also tried various hybrid and heuristic methods for action unit and facial expression recognition. Zhang and Ji[10] utilized a hybrid method of 26 landmark points, appearance information of crow feet wrinkles, nasal labial furrows as well as probabilistic and dynamic framework for spontaneous facial expression representation and recognition.

2.2 Limitations

Recent works are struggling to model spontaneous and natural facial action unit recognition.

- Some of them focus on frequent AUs. There is no systematic way to find out significant AU relations which are mostly responsible for spontaneous facial behavior. For example, Tong et al.

[6] selected only 14 AUs which are most frequent in the facial expressions. Selection of frequent AUs are done with Cohn-Kanade dataset by heuristics. The simple heuristic of the strategy is to select AUs which are most frequent in video sequences. This kind of simple heuristic can exclude very important relationships between AUs. Additionally, some AUs with high frequencies can have significant relations with other AUs with low frequencies. Rather than individual occurrences of AUs, relationships between them show the most significance in spontaneous facial behavior. Not all facial expressions can be modeled with 14 AUs. Similarly, [22] and [19] have selected 15 and 20 AUs respectively based on the frequencies for recognition.

- In holistic approaches, the same extracted features are fed to each classifier. The process suffers from a large scale of complexities for feature extraction. Since same features are used for each AU, redundant information is incorporated within the training data. In these situations, redundant information makes classification models worse.
- Employing a large number of individual binary classifiers increases the number of computations and memory consumption. Though recognition accuracies are reasonable, global feature extractions from the face suffer enormous complexities and delays. Spontaneous facial behavior includes more AUs which can increase the complexities further.
- [21] is not quite effective due to TANs structure limitations. TAN cannot handle complex relationships such as hierarchical dependencies between facial features, as well as temporal changes. Moreover, strength of the relationships does not represent probabilistic dependencies.
- Additionally, computer vision techniques are not robust in the natural environment and thus can produce erroneous recognition of AUs for varying lighting conditions, head movements, varying facial features across the ethnicities etc.
- Robust recognition must ensure correctness of computer vision techniques and classification algorithms for each and every AU which is practically impossible.

It is more desirable to keep concentration on the most significant AUs and their recognition processes. If we can ensure the reasonable accuracies, then real time estimation of IWAUs will do the rest of the recognition without any computer vision techniques and classifications.

Chapter 3

FACS and AU Relations

This chapter presents an overview of the Facial Action Coding System(FACS) and the relationships existing between different AUs .

3.1 Facial Action Coding System(FACS)

Facial Action Coding System (FACS) is the most widely used and versatile method for measuring and describing facial behaviors [2]. Paul Ekman and W.V. Friesen developed the original FACS in the 1970s by determining how the contraction of each facial muscle changes the appearance of the face. The Facial Action Coding System breaks facial muscles into particular 'action units' that are responsible for unique human facial expressions [23]. They are groups of muscles instead of specific muscles. This framework allows an researcher/psychologist to document and categorize a person's facial expressions. By understanding which muscle action units activate with which expression, we can better discern if someone is happy or sad. The concept is robust because these facial expressions and body movements appear to have universal characteristics amongst humans. The muscle actions are called Facial Action Units(AUs). As AUs are independent of any interpretation, they can be used for computing affective state, recognition of facial expressions et.

Each AU is identified by a number and name (for example, one AU explained is 'AU 4 - Brow Lowerer'). Names like 'Brow Lowerer' are provided as a more meaningful handle than the more arbitrary numbers, and might make it easier for us to relate to the AUs. Figures 3.1, 3.2, 3.3 show common AUs along with the related muscle groups [1].

3.2 AU Relations

In spontaneous facial behavior, combinations of AUs create meaningful facial expressions [2]. The relations between AUs are functions of physiological constraints as well as facial expressions. Figure 3.4 illustrates the muscle relations. Physiological constraints that are resulted from the anatomy of the human face are critical for analyzing relations between AUs. A number of examples are briefly described. AU 1 and AU 2 are connected with each other through the *Frontalis* muscle while *Frontalis* muscle causes the relation between AU 1 and AU 2 to be stronger. Similarly, AU 5 has been activated by *Levator palpebrae superioris* which is connected to *Orbicularis oculi*. Again, the *frontalis* muscle and its fascia are connected with the *orbicularis oculi* muscle at the level of the eyebrow. Thus AU 5 may











| AU | Description | Facial muscle | Example image |
|-----------|---------------------|--|--|
| <u>1</u> | Inner Brow Raiser | <i>Frontalis, pars medialis</i> |  |
| <u>2</u> | Outer Brow Raiser | <i>Frontalis, pars lateralis</i> |  |
| <u>4</u> | Brow Lowerer | <i>Corrugator supercilii, Depressor supercilii</i> |  |
| <u>5</u> | Upper Lid Raiser | <i>Levator palpebrae superioris</i> |  |
| <u>6</u> | Cheek Raiser | <i>Orbicularis oculi, pars orbitalis</i> |  |
| <u>7</u> | Lid Tightener | <i>Orbicularis oculi, pars palpebralis</i> |  |
| <u>9</u> | Nose Wrinkler | <i>Levator labii superioris alaquae nasi</i> |  |
| <u>10</u> | Upper Lip Raiser | <i>Levator labii superioris</i> |  |
| 11 | Nasolabial Deepener | <i>Zygomaticus minor</i> |  |
| <u>12</u> | Lip Corner Puller | <i>Zygomaticus major</i> |  |

Figure 3.1: Example of Action Units [1]

affect AU 2 in a different scale of dominance. Similarly, AU 15 pulls the corners of the lips down that have been affected by AU 17 (Chin Raiser). Both AUs are connected through *Orbicularis oculi* and *Mentalis* resulting in a stronger relation.

Again, AU 23 (Lip tightener) and AU 24 (Lip presser) are controlled through the muscle








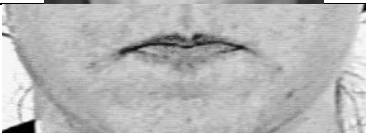
| AU | Description | Facial muscle | Example image |
|-----------|----------------------|--|--|
| 13 | Cheek Puffer | <i>Levator anguli oris (a.k.a. Caninus)</i> |  |
| 14 | Dimpler | <i>Buccinator</i> |  |
| <u>15</u> | Lip Corner Depressor | <i>Depressor anguli oris (a.k.a. Triangularis)</i> |  |
| 16 | Lower Lip Depressor | <i>Depressor labii inferioris</i> |  |
| <u>17</u> | Chin Raiser | <i>Mentalis</i> |  |
| <u>20</u> | Lip stretcher | <i>Risorius w/ platysma</i> |  |
| <u>23</u> | Lip Tightener | <i>Orbicularis oris</i> |  |
| <u>24</u> | Lip Pressor | <i>Orbicularis oris</i> |  |

Figure 3.2: Example of Action Units [1]

Orbicularis oris. AU 24 might not be scored separately for the top or bottom lip, even though the pressing action may be greater in one lip than the other. Therefore, AU 23 and AU 24 are strongly connected. The strength of the relationship between AU 23 and AU 24 may not be the same as that between AU 1 and AU 2.




| AU | Description | Facial muscle | Example image |
|------------------|---------------|--|--|
| <u>25</u> | Lips part** | <i>Depressor labii inferioris or relaxation of Mentalis, or Orbicularis oris</i> |  |
| <u>26</u> | Jaw Drop | <i>Masseter, relaxed Temporalis and internal Pterygoid</i> |  |
| <u>27</u> | Mouth Stretch | <i>Pterygoids, Digastric</i> |  |

Figure 3.3: Example of Action Units [1]

Conversely, AU 25 (Lips parted) is associated with *Orbicularis oris* that exists in the lower part of the face. *Levator palpebrae superioris* and *Orbicularis Oris* are indirectly connected through *Levator Labii Superioris*, *Zygomaticus minor and major*, and *Orbicularis oculi*. Indirect connections of muscles result in a weak relationship between AU 5 and AU 25. Additionally, *Zygomaticus Minor* and *Risoricus* are not connected to each other but the muscles *Zygomaticus*, *Orbicularis Oculi*, and *Masseter* are responsible for the weaker relation between AU 11 and AU 20. AU 9 is related to the muscle *Levator Labii Superioris*. The muscle *Frontalis* is not directly connected but is indirectly through the muscle *Orbicularis Oculi* representing another weak relation with AU 1 and AU 2.

Relations between AUs are also functions of facial expressions. Though AU 6 and AU 12 are related to the different facial regions, they are involved in happy expressions frequently resulting strong relation among themselves. Relations between AU 2 (Outer brow raiser) and AU 27 (mouth stretch) possess a mixture of the both kinds of relations. Though AU 2 and AU 27 are situated in upper face and lower face respectively, large mouth stretch causes significant movement of *Frontalis* through *Labii Superioris* and *Orbicularis Oculi*. Additionally, the relation between AU 2 and AU 27 is an example of an expression oriented relationship in terms of a surprise expression. Therefore, expression oriented relations are also the contributed factors in the relation analysis in addition to physical connections.

Using a linguistic analogy, one can define the AU tree and the root based on frequently occurred strong relations that are present in the expressions database(s). The main objective here is to

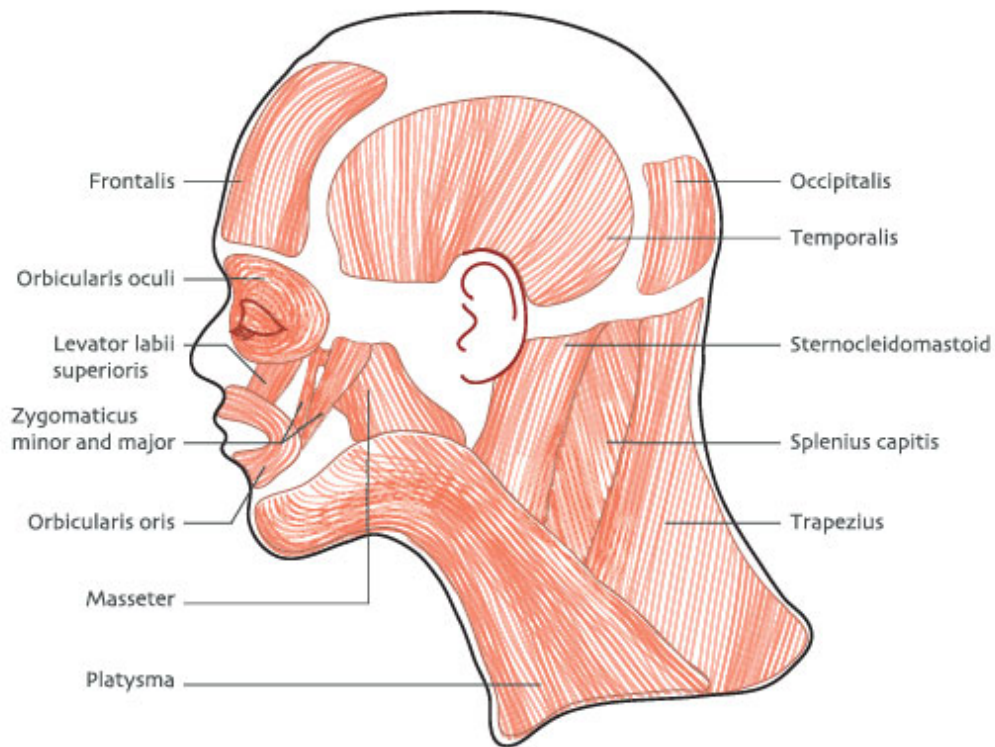


Figure 3.4: Muscles in Face

separate the strong relations between AUs from weaker ones. The next sections discuss in detail the scoring techniques and a framework to distinguish the FSAUs and IWAUs using the proposed approach. Figure 3.4 depicts the muscles on a face.

Chapter 4 Modeling AU Relations

4.1 Modeling AU Relations with Static Bayesian Network(SBN)

AU relationships are modeled with SBN where AUs are represented as nodes [6]. Specifically, each node i corresponds to particular AU. Relationships between AUs are represented as directed edges. Directed edges represent the dependencies of a child node on its parents. Each node acts as a random variable. Nodes are binary nodes whose values are assigned by the occurrences of corresponding AUs. Therefore, possible value of each node can be true(numerical value 1) or false(numerical value 0). AU intensity is not included. If n AUs are modeled, one child can have at most $n-1$ parents.

A SBN has two properties. The structure of SBN depicts the dependencies/relations between the nodes. Conditional Probability Table (CPT) will represent the nature of the dependencies/relations. Learning the structure and the CPTs from given data is more desirable rather than expert guess. Learning from data ensures that the learned structure and CPTs will reflect the data mostly.

Structure learning is performed on an initial SBN. Initial SBN may be a random network or empty network. Structure learning step provides the final BN structure which represents the relationships between nodes. Relationship between nodes reflect the AU relationships in specific dataset. Learned structure has been used to learn CPTs through parameter learning process.

4.2 Structure Learning of SBN

Structure learning is an iteration process in which the network is learned according to the given dataset. In score-based process, network learning depends on the predefined scoring function. In this work, scoring function defined by Cooper et al. [24] has been used.

Having defined a score, we need to identify a network structure with the highest score by a searching algorithm. Thus, the structure learning becomes an optimization problem: find the structure that maximizes the score. Among score based processes, Hill climbing is most popular. A modified hill climbing approach has been used find the final structure.

1. **Initialize the starting network structure:** First, a random network structure or empty structure is used as the starting network structure . In this work, we start with the empty network (B_0) which is nothing but network with all nodes without any edges.

2. Find a final structure:

- Starting from B_0 , compute the score of each nearest neighbor B_S of B_0 , which is generated from B_0 by adding, deleting, or reversing a single arc, subject to the acyclicity constraint.
- Update B_0 with the BN (B_S) that has the maximum score among all of the nearest neighbors and go back to the previous step until no neighbors have higher score than the current structure.

4.2.1 Score of a SBN

Let in a SBN, there are n discrete nodes, where a node i can have r_i possible value assignments:

$(v_{i1}, \dots, v_{ir_i})$.

Let D be a database of m cases, where each case contains a value assignment for each node in SBN. Let B_S denote a belief-network structure. Let x_i is variable in data set D corresponds to node i . Score of B_s is defined by function score.

$$\text{score}(B_S, D) = \log(P(B_S|D)) \quad (4.1)$$

$$= \log P(B_S, D) - \log P(D) \quad (4.2)$$

Here $\log P(D)$ is constant for given data. $P(B_S, D)$ denotes

$$P(B_S, D) = \int_{\theta} P(D|B_S, \theta) f(\theta|B_S) P(B_S) d\theta \quad (4.3)$$

Where θ is a vector of parameters associated with belief-network structure B_S , and f is the conditional-probability density function over θ given B_S . For simplicity, it is assumed that the density function $f(\theta|B_S)$ is uniform rather than multinomial prior distribution.

Since, D is a complete data set, $P(B_s, D)$ can be calculated using Cooper's work et al.,[24]. Before discussing Cooper's method, let us introduce some useful notations.

4.2.2 Notations

We represent the parents of node i as a list (vector) of variables, which we denote as π_i . We shall use w_{ij} to designate the j^{th} unique instantiation of the values of the variables in π_i , relative to the ordering of the cases in D . We say that w_{ij} is a value or an instantiation of π_i . Suppose there are q_i such unique instantiations of π_i . Define N_{ijk} to be the number of cases in D in which variable x_i has the value v_{ik} and π_i is instantiated as w_{ij} .

Define,

$$N_{ij} = \sum_{k=1}^{r_i} N_{ijk} \quad (4.4)$$

For special case, if an node i has no parents, j has no value associated with node i . So, j has null value. In this situation, define $N_{i\emptyset k}$ instead of N_{ijk} for particular node i where $N_{i\emptyset k}$ defines, how many times k^{th} possible value of node i has been occurred. And

$$N_{i\emptyset} = \sum_{k=1}^{r_i} N_{i\emptyset k} \quad (4.5)$$

Example:

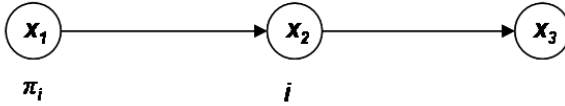


Figure 4.1: Example BN

To clarify the notations, let us discuss about the simple Static Bayesian Net BN in 4.1 for data set shown in table 4.1. Dataset D has three variables x_1, x_2, x_3 . Variable x_i corresponds to the node i .

In dataset, each variable can have any of two values: false or true. So for each node i , $r_i = 2$. And $(v_{i1} = false, v_{i2} = true)$. π_i denotes the parent of node i . if node i has more than one parent, π_i will be just like a list to store the parents such as $\pi_i(1), \pi_i(2), \dots$. For node $i = 2$ as in the figure 4.1, following discussion has been carried out to describe the notations. In the example BN, $\pi_i = 1$. Now, π_i has two unique instantiation in the table because 1^{st} value of the variable x_1 is false. 2^{nd} value of the

Table 4.1: Data set D

| x_1 | x_2 | x_3 |
|-------|-------|-------|
| false | true | false |
| false | false | true |
| true | true | false |
| false | false | false |
| true | true | false |

variable x_1 is also false which is not a unique one. But 3^{rd} value is true which is the unique instantiation for node π_i . So for node i , $q_i = 2$. And, $w_{i1} = false$ and $w_{i2} = true$. Now, for first instantiation of π_i , $j = 1$ and $w_{i1} = false$. Calculate N_{ijk} for $k = 1$ and $k = 2$. For $k = 1$, N_{ijk} is simply N_{i11} . Following the definition of N_{ijk} , N_{i11} stands for, how many times $x_1 = false$ and $x_2 = false$ occurred in Dataset D. From the example, $N_{i11} = 2$.

Similarly, for $k = 2$, N_{ijk} is simply N_{i12} and it means, how many times, $x_1 = false$ and $x_2 = true$. Table 1 shows that $N_{i11} = 1$. Same calculation follows to calculate N_{ijk} for second instantiation of π_i . Here $j = 2$ and $w_{i2} = true$. From the example, $N_{i21} = 1$ and $N_{i22} = 2$. Please note that, if π_i is a list of parents, we need to calculate q_i and w_{ij} for every parent separately to calculate N_{ijk} .

Let us introduce the calculation for special case, when node i has no parents. node $i = 1$ has no parents in the example. No calculation for j needed here. For $k = 1$, $N_{i\emptyset 1}$ refers to how many times, $x_1 = false$ occurs in the data set. So, from the given dataset, $N_{i\emptyset 1} = 3$. For $k = 2$, $N_{i\emptyset 2}$ refers to how many times, $x_1 = true$ occurs in the data set. $N_{i\emptyset 1} = 2$.

Chapter 5

Identify Strong AU Relations

We developed a proposition to identify strong relations between AUs based on a scoring mechanism while empty SBN is used as the initial network in the hill climbing algorithm. Hill-climbing search procedures examine all possible local changes in each step and apply the one that leads to the biggest improvement in score. The usual choice for local changes are edge addition, edge deletion, and edge reversal. Defined bayesian scoring function [24] has been customized in hill climbing process without constraints.

5.1 Proposition

Strong relations are modeled in the earlier iterations while weak relations are modeled at the later iterations of the SBN structure learning process using hill climbing algorithm.

Before describing the proposition, some notations are introduced here. We represent the parents of node i as a list π_i . w_{ij} designates the j^{th} unique instantiation of the values of the variables in π_i . Suppose there are q_i such unique instantiations of π_i . Define N_{ijk} to be the number of cases in D in which variable x_i has the value v_{ik} and π_i is instantiated as w_{ij} . Define N_{ij} as, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

For special case, if a node i has no parents, define $N_{i\emptyset k}$ instead of N_{ijk} for a node i where $N_{i\emptyset k}$ defines, how many times k^{th} possible value of node i has been occurred. And consequently, $N_{i\emptyset} = \sum_{k=1}^{r_i} N_{i\emptyset k}$.

5.2 Explanation of the Proposition

Let in a SBN, there are n discrete nodes, where a node i can have r_i possible value assignments: $(v_{i1} \dots v_{ir_i})$. D is a database of m cases. B_S denotes a belief-network structure. Let x_i is a variable in data set D corresponds to node i . Score of B_s is defined by function $\Psi(B_S, D)$ as follows.

$$\Psi(B_S, D) = \ln P(B_S, D) - \ln P(D) \quad (5.1)$$

Here $\ln P(D)$ is constant for given data. Where, $P(B_S, D)$ denotes

$$P(B_S, D) = \int_{\theta} P(D|B_S, \theta) f(\theta|B_S) P(B_S) d\theta \quad (5.2)$$

where θ is a vector of parameters associated with B_S , and f is the conditional-probability density function over θ given B_S . $P(B_S, D)$ can be calculated by Cooper's[24] method. For simplicity, it is assumed that the density function $f(\theta|B_S)$ is uniform rather than Dirichlet distribution.

$$P(B_S, D) = P(B_S) \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \quad (5.3)$$

Substituting the value of $P(B_S, D)$ using [24] follows,

$$\Psi(B_S, D) = \ln P(B_S) + \underbrace{\ln \left(\prod_{i=1}^n \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \right)}_{\gamma} - \ln P(D) \quad (5.4)$$

Since, no prior network has been used, $\ln P(B_S)$ is uniform over all possible structures. Only γ is responsible to raise score in any iteration. Specifically, at iteration t ,

$$\gamma(t) = \ln \left(\prod_{i=1}^n \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \right) \quad (5.5)$$

Some definitions are used to simplify $\gamma(t)$.

Definition 1: An orphan node has no parents. Score of an orphan node i is defined by,

$$\Psi_{orp}(i) = \frac{(r_i - 1)!}{(N_{i\emptyset} + r_i - 1)!} \prod_{k=1}^{r_i} N_{i\emptyset k}! \quad (5.6)$$

Definition 2: If there is directed edge from a parent node π_i to node i , the score for the arc is defined by,

$$\Psi_{arc}(\pi_i) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \quad (5.7)$$

If node i has more than 1 parent, π_i works like a list of parent nodes. $\Psi_{arc}(\pi_i)$ will be calculated for each parent in the list. Basically, $\Psi_{arc}(\pi_i)$ defines the strength of the corresponding arc from node π_i to node i [24]. The larger the value of $\Psi_{arc}(\pi_i)$ is, the stronger the edge will be. So, for a best network at

iteration t , above definitions can be used to simplify $\gamma(t)$ as,

$$\gamma(t) = \ln\left(\prod_{i \in \text{orphan}} \Psi_{orp}(i) \prod_{u \in \text{child}} \Psi_{arc}(\pi_u)\right) \quad (5.8)$$

where $i \neq u$ and π_u is the list of parents of child node u . Following procedure describes the change of $\gamma(t)$ at each iteration based on the possible edge addition, reversal, or deletion.

Initialization: For empty network, each node is an orphan node. So, initially,
 $\gamma(0) = \ln \prod_{i=1}^n \Psi_{orp}(i)$.

Updates at each iteration: In hill climbing approach, one arc can be selected to be added, reversed, or deleted to the current network in each iteration that maximizes the score.

Addition: Let, one arc from π_i to node i has been selected to be added at iteration t . The change of $\gamma(t)$ would follow any of the two possible events.

Event 1: If any node π_i is being connected as parent to an orphan node i , node i will be no longer an orphan node.

$$\begin{aligned} \gamma(t) &= \gamma(t-1) - \ln \Psi_{orp}(i) + \ln \Psi_{arc}(\pi_i) \\ &= \gamma(t-1) - \ln\left(\frac{(r_i-1)!}{(N_{i\emptyset}+1)!} \prod_{k=1}^{r_i} N_{i\emptyset k}!\right) + \ln\left(\prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}+r_i-1)!} \prod_{k=1}^{r_i} N_{ijk}!\right) \end{aligned} \quad (5.9)$$

Event 2: If any node π_i is being connected to a non-orphan node i , then,

$$\begin{aligned} \gamma(t) &= \gamma(t-1) + \ln \Psi_{arc}(\pi_i) \\ &= \gamma(t-1) + \ln\left(\prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}+r_i-1)!} \prod_{k=1}^{r_i} N_{ijk}!\right) \end{aligned} \quad (5.10)$$

Both events are illustrated in figure 5.1 where edge marked with \checkmark indicates the addition.

Reversal: Similarly, the change of $\gamma(t)$ follows from any of the four different events. Equations in 5.11 show the changes of $\gamma(t)$ for corresponding events where figure 5.2 shows the possible events. Dashed edge is subject to be reversed and (\checkmark) indicates the new edge after reversal.

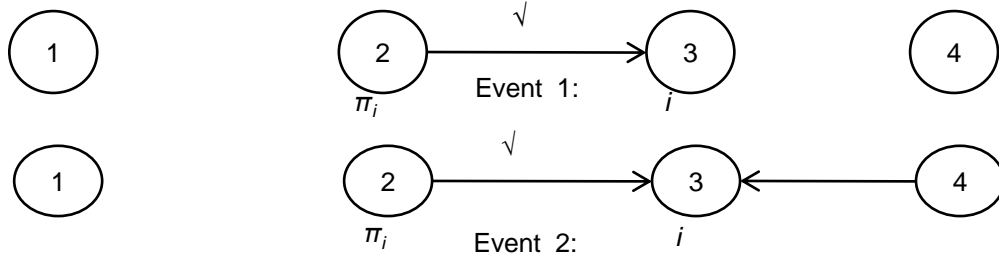


Figure 5.1: Possible Events for Add operation

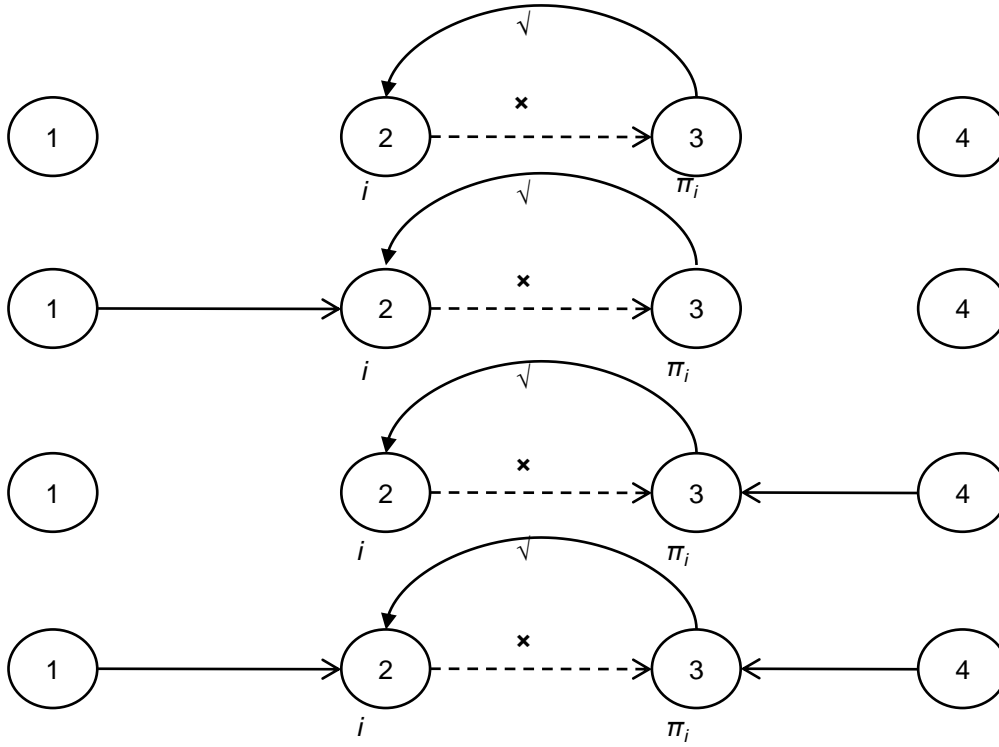


Figure 5.2: Possible Events for Reverse operation

$$\gamma(t) = \gamma(t-1) + \ln \Psi_{orp}(\pi_i) - \ln \Psi_{orp}(i) + \ln \Psi_{arc}(\pi_i) - \ln \Psi_{arc}(i)$$

$$\gamma(t) = \gamma(t-1) + \ln \Psi_{orp}(\pi_i) + \ln \Psi_{arc}(\pi_i) - \ln \Psi_{arc}(i)$$

$$\gamma(t) = \gamma(t-1) - \ln \Psi_{orp}(i) + \ln \Psi_{arc}(\pi_i) - \ln \Psi_{arc}(i)$$

$$\gamma(t) = \gamma(t-1) + \ln \Psi_{arc}(\pi_i) - \ln \Psi_{arc}(i)$$

(5.11)

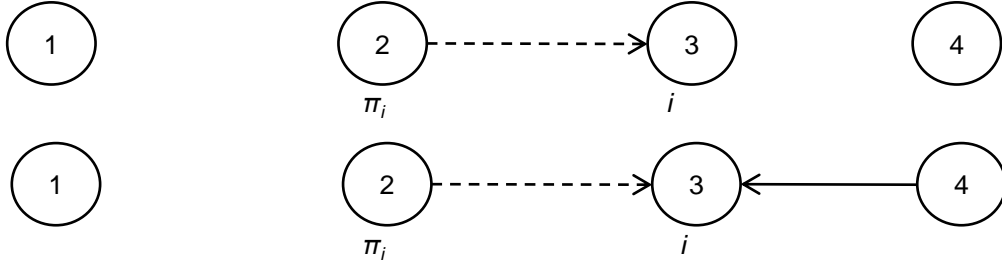


Figure 5.3: Possible Events for Delete operation

Deletion: Delete operation might have any of the two events in figure 5.3 where dashed edge indicates the deletion. 5.12 corresponds the events for deletion.

$$\begin{aligned}\gamma(t) &= \gamma(t-1) + \ln \Psi_{orp}(i) - \ln \Psi_{arc}(\pi_i) \\ \gamma(t) &= \gamma(t-1) - \ln \Psi_{arc}(\pi_i)\end{aligned}\tag{5.12}$$

It can be easily shown that the upper bound of $|\ln \Psi_{orp}(i)|$ is given by $|\ln \frac{1!}{(R+1)!} (\frac{R}{2})! (\frac{R}{2})!|$ when $r_i = 2$ for a complete dataset with R records. The contribution of $|\ln \Psi_{orp}(u)|$ is insignificant to the score since $|\ln \frac{1!}{(R+1)!} (\frac{R}{2})! (\frac{R}{2})!| \ll \ln \Psi_{arc}(v)$ for any u, v where $u \neq v$ verified by a number of simulation with different FACS datasets. Consequently, delete operations are not performed in the entire learning process since it decreases $\gamma(t)$. Actually, the $\ln \Psi_{arc}(\pi_i)$ of corresponding arc from node π_i to node i is responsible for increasing the $\gamma(t)$ in any event due to add or reverse operation.

In the hill climbing process, arcs which increase the score significantly will be selected during earlier iterations. And increase of score is defined by $\ln \Psi_{arc}(\pi_i)$ in each of the $\gamma(t)$ function. Since $\ln \Psi_{arc}(\pi_i)$ shows the strength of relationship, we would say that strong relations (frequent relations) are modeled at earlier iterations. At later stages, weak relations are modeled. When the parameter distribution $f(\theta|B_S)$ comes from Dirichlet Distribution, the proposition can be proved in the same way.

Chapter 6

Finding Significant Subset

6.1 Strong Relations vs. Weak Relations

Following the proposition, the entire structure learning process is divided into two parts, where the first part (“Buildup Area”) contains iterations which are involved to model stronger edges and the later part “Tuning Area” is responsible for modeling weak relations. In the “Buildup Area”, the stronger AU relation increases the score at a very high rate of change albeit one at a time. Therefore, the main portion of score buildup has occurred in “Buildup Area”. In “Tuning Area”, the structure is being tuned with weaker edges.

Initially, Extended Cohn-Kanade dataset (CK+) [9] has been used to build the SBN structure to identify the strong relations. 23 AUs are used since they are appeared more than 9 video sequences. In figure 6.1, before the 8th iteration, rates of change of network score are high. Alternately, after the 8th iteration, rates of change of score are very low. A sharp transition is observed at iteration 8. Using the proposition, we would say that up to iteration 8, all frequent relations are modeled in SBN structure. Weaker relations are added in later iterations. Therefore, the SBN structure shown in figure 6.2 obtained at iteration 8 gives us all strong or frequent edges.

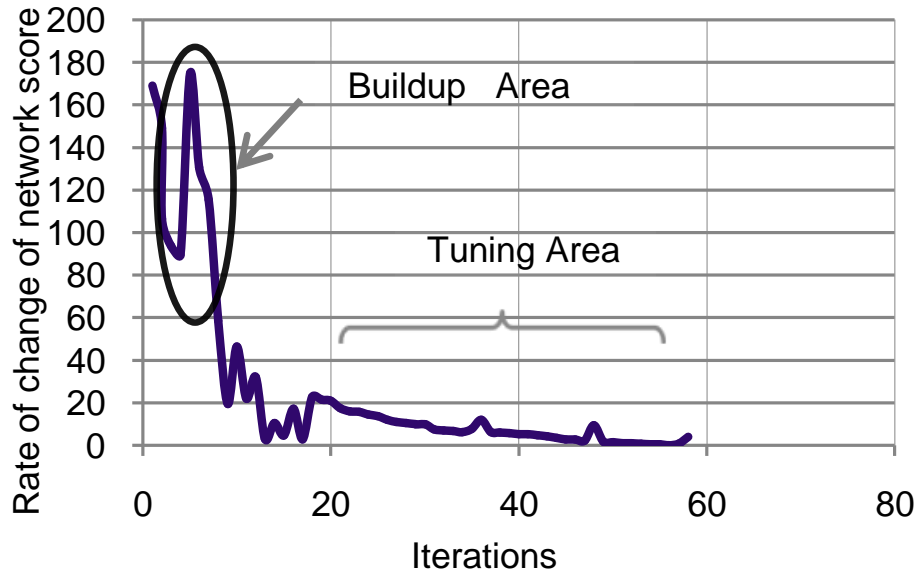


Figure 6.1: Rate of Change of scores

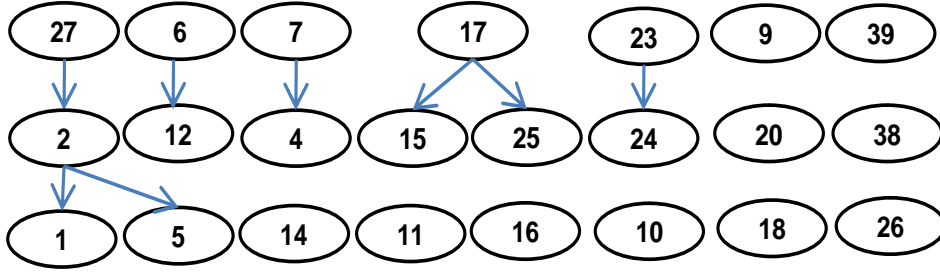


Figure 6.2: Strong relations between AUs

6.2 FSAUs

Since every facial expression can be derived from some specific strong relations, the strong relations between AUs represent the root relations between AUs. Figure 6.2 shows that particular AUs are involved in building the strong relations. The responsible AUs to build strong relations are defined as FSAUs as they are the building blocks of root relations. In other words, FSAUs are the most important participants to perform a wide range of facial behaviors. They are AU (1, 2, 4, 5, 6, 7, 12, 15, 17, 23, 24, 25, and 27).

Now, the right side in figure 7.2 shows the final structure of SBN where boxed nodes represent FSAUs. The final structure has been obtained later with greedy hill climbing approach with a number of random restarts to make the model optimal. Note here that node 1 and node 5 are FSAUs but they are leaf nodes. The status of node 1 and node 5 do not affect any other nodes. AU 1 and AU 5 are removed from FSAU set and are considered as member of IWAUs. So, now we have 11 FSAUs.

6.3 Parameter Learning for Final SBN

The structure learning process gives us the final structure of the BN. Figure 5 shows the Final structure a B_S after the whole structure learning process. The box nodes indicate the FSAUs in the structure. The structure B_S is used to learn the parameters.

Let θ denote the vector of parameters for a given BN structure B_S . $\theta = \{\theta_1, \theta_2, \dots, \theta_n\}$ Let θ_{ijk} denote the conditional probability $P(x_i = v_{ik} | \pi_i = w_{ij}, B_S)$. Parameter learning step will give us θ which will maximize the posterior distribution $p(\theta | D, B_S)$. Assuming θ_{ij} are mutually independent, $p(\theta | D, B_S)$ is defined by,

$$p(\theta|D, B_S) = \left(\prod_{i=1}^n \prod_{j=1}^{q_i} p(\theta_{ij}|D, B_S) \right) \quad (6.1)$$

where $\theta_i = ((\theta_{ijk})_{k=1}^{r_i})_{j=1}^{q_i}$ and $\theta_{ij} = \{\theta_{ij1}, \theta_{ij2}, \dots, \theta_{ijr_i}\}$

For the parameter distribution, the conjugate prior come from Dirichlet family. So the posterior distribution will be given by

$$p(\theta_{ij}|D, B_S) = Dir(N_{ij1} + \alpha_{ij1}, \dots, N_{ijr_i} + \alpha_{ijr_i}) \quad (6.2)$$

where α_{ijk} is about prior information. α_{ijk} can be calculated just as N_{ijk} . Instead of calculating MAP of θ_{ijk} , approximated ML of θ_{ijk} can be easily calculated from a given dataset D.

$$\theta_{ijk} = \frac{N_{ijk} + \alpha_{ijk}}{N_{ij} + \alpha_{ij}} \quad (6.3)$$

where, $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$.

6.4 IWAU Prediction

Significant AUs are represented as boxed nodes in the final SBN. For the inference procedure, nodes for corresponding significant AUs act as observed nodes. That is, boxed nodes are observed nodes. IWAU nodes are hidden nodes(unobserved). Presence/absence information from a particular data point of data set for corresponding boxed nodes are applied to the appropriate observed nodes as evidence. Prediction step includes inference procedure which will predict the status of hidden nodes. Status of hidden nodes represent whether they are present or not. Standard SBN inference algorithm(Junction tree) has been used to infer the probabilities.

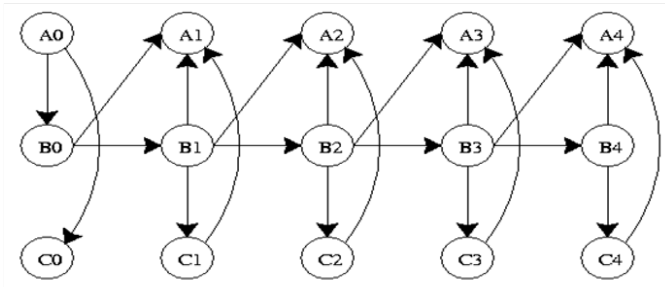
The final SBN shows that all hidden nodes are not connected directly to the observed nodes. But, to infer a node correctly from a bayesian net, all parents of that node should be observed. Hidden nodes are in different depth relative to observed nodes. For example, node 16 is descendent of node 23 and node 4 with depth. But it is also descendent of node 17. So, we need to infer the status of node 26 and node 18 before the inference of node 16. This procedure should repeat for larger depth. For

node 14, we need to infer node 9 at first. But node 9 also depends on an unboxed node 20. Node 20 depends on another un-boxed node 26. To overcome the problem, a Depth First Search(DFS) algorithm has been applied to predict the nodes in a sequence so that any node can be inferred from its parents' status. With DFS, inferred nodes will be the visited nodes. The final BN has been fed to DFS algorithm. Initialization step marks all observed nodes(boxed nodes)as visited nodes. In the searching step, when any node is being marked with visited node, status of that particular node will be inferred from its parents' nodes. Let us describe an example. At the initial situation, SBN is the graph where boxed nodes are marked visited nodes. Searching step starts from node 1. Since node 1 is already visited, search step jumped to node 2. It is also visited. Search step continues to follow in increasing nodes. When it will jump to node 9, it finds that parent nodes except node 20 are visited. So search step will discovered node 20. Node 20 has one unvisited node 26. So search step will discover node 26 which has no unvisited node. So, the search step will mark the node 26 as visited and the status of node 26 will be inferred from node 4. Then search process will back track to node 20. Now node 20 has no parents unvisited. In this point, node 20 will be inferred from all of his parents' status. After the inference of node 20, search process will back track to node 9 and infer the status of node 9. Then search process will jump to node 10. The search process will continue until all hidden nodes are not inferred.

Chapter 7

Modeling Temporal Relationship between AUs using Dynamic Bayesian Network

Time series data evolves over time. Static Bayesian Network (SBN) is not capable to represent the evolution of time series data. Since time series data are represented as sequential data along time, it is natural to use directed graphical model that can hold the sequential nature of the data. So SBNs connected by directed edges can be adapted to model the time series data. And the architecture is called known as Dynamic Bayesian Network (DBN). Actually, DBN generalizes hidden Markov models (HMMs) and linear dynamical systems (LDSs). Figure 7.1 illustrates an example of DBN. It is an unrolled version of a DBN. That means, five SBNs are unrolled with the same connection between them. Same Conditional Probability Distribution (CPD) is associated for each inter-slice connection. Sequential data with length five can be intuitively modeled with this DBN. For the sake of simplicity, first-order markov assumption is taken for the experiments. That is, the parents of a node can only be in the same time slice or the previous time slice, i.e., arcs do not across slices. Inter-slice arcs are all from left to right, reflecting the flow of time. The relationships between two neighboring time slices are modeled by an HMM such that random variables at time t are influenced by other variables at time t , as well as by the corresponding random variables at time $t-1$ only. Each time slice is used to represent the snapshot of an evolving temporal process at a time instant.



simultaneous with or closely followed by one or more associated action units, such as AU 6, AU 15 or AU 17. Generally, subset of these AUs follow AU 12 depending on different kinds of smile [25]. In most of the smiles indicating enjoyment with multiple action units, AU 6 is the first one to follow AU 12. The work also shows that AU 6 appeared an average of 11 frames after the beginning of the smile. For the disgust expression, AU 9 (Nose Wrinkler) is preceded by the activation of AU 10 (Upper Lip Raiser). At the beginning of the disgust expression, upper lip raises and then the nose wrinkler begins to deepen. Their relaxations also maintain the sequence. Moreover, the transition between AUs varies in spontaneous expressions. It depends on subjects also. So, in the context of temporal evolution of AUs in terms of DBN, IWAUs are dependent on the AUs at earlier time slices.

If we can model the temporal evolution of AUs, IWAUs can be predicted more accurately using the evolution model. Since we use FSAUs as evidences at each time frame, we use the evolution model to predict the IWAUs only.

7.1.1 Modeling Temporal Evolution of AUs with DBN

Dynamic Bayesian network consists of time slices of static bayesian net(SBN). SBN denotes our final BN. In a time slice, corresponding SBN represents the snap shot of relationships between AUs in one video frame. So, the temporal evolution of AUs in a video sequence can be represented with interconnected SBNs. The directed edges between SBNs represent the temporal relationship. Therefore in DBN, a AU node in time slice t depends on AUs at time $t - 1$ as well as other AUs at time t . Since we assume that FSAUs are always available from the recognition machine, and we use eleven IWAUs as evidence at each frame, construction of DBN is not straight forward. We impose some constraints to build the DBN. FSAUs at time t can not be connected as children of any AU node at $t - 1$. Moreover, IWAUs are observed and their status are available at each time t .

We can model the DBN using 2 time slice Bayesian net (2TBN). That is, two SBN are connected with temporal edges. First SBN represents the time slice $t - 1$ and the second SBN represents the time slice t . Temporal edges describe the structure of DBN. From the structure learning of SBN, we know the intra connections of each SBN. But, inter-slice relationships are undefined. As of SBN, we need to search out the temporal edges which are appropriate for certain data set. Learning a DBN needs the data-set divided into time slices. For this purpose, frames for each video sequence of Cohn kanade dataset have been labeled with AU codes. ISL dataset provides the codes for 14 AUs for each frame. We have labeled the other 9 AUs for each frame of the video sequences. As a result, the modified data set has video sequences each of which have multiple frames coded with AU codes. Now

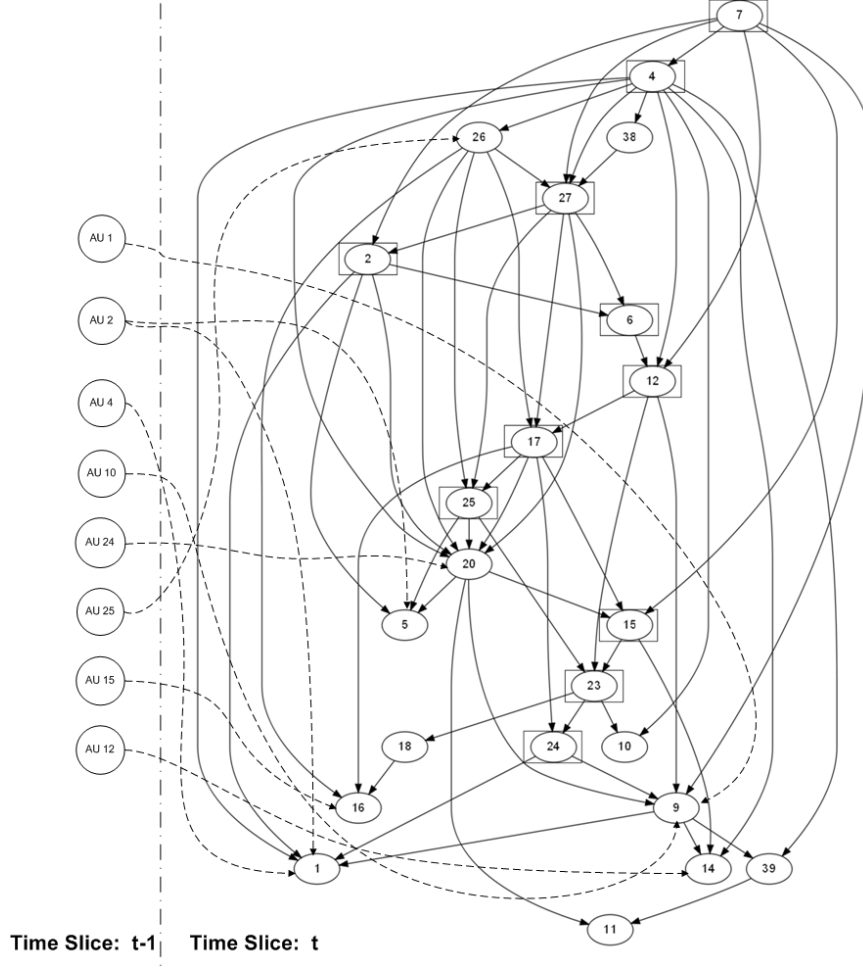


Figure 7.2: Two Slice DBN

one data point $D(t, l)$ denotes the vector of the AUs of t_{th} time slice from l^{th} video sequence. Hill climbing method has been used to search out the temporal edges. Structure learning for 2TBN does not allow to modify edges within the same SBN. It searches for the temporal edges between 2TBN. The hill climbing method also maintained the above constraints about significant AUs. After the structure learning, resultant 2TBN in figure 7.2 shows the temporal edges from one slice to another. Here, the AUs of time slice t depends on the AUs at time slice $t - 1$ as well as AUs at time slice t . Edges from node i of time slice $t - 1$ to node j of time slice t represents the AU evolution over time.

7.1.2 Parameter Learning

Parameter learning of 2TBN is similar to SBN. Only the difference is, in each time slice, each AU(except FSAUs) has parent nodes in the same time slice as well as earlier time slice. Parameter learning provides CPTs for both intra-slice and inter-slice dependencies. To shrink the space, we escape the structure learning and parameter learning step.

7.2 Prediction of IWAUs using DBN

As it is too difficult to guess all temporal evolution, training by a good database is essential. CK+ dataset [9] has been applied to learn the temporal edges between two slices as well as to learn the CPTs. Figure 7.2 shows the final 2TBN where temporal edges (dotted lines) from time slice $t - 1$ to time slice t represent the temporal evolution. Temporal edges between same IWAUs are not shown here to keep the figure simple. Let denote the FSAUs at time slice t as a vector P_t where $p_t \in P_t$. And the IWAUs are defined by S_t where $s_t \in S_t$. U_t denote the whole set of AUs at time t . Updating process continues as a 2TBN such that SBN at t is connected with previous SBN at $t - 1$ using the proposed approach in [6] with essential modifications. For the inference procedure, nodes for corresponding FSAUs (boxed nodes) act as observed nodes while IWAUs are hidden nodes.

Initialization: As the first step, SBN for $t = 0$ is empty. In this situation, SBN for $t = 1$ acts as one slice SBN and S_1 are inferred from the P_1 .

Prediction: Given that, estimated probability distribution $P(s_{t-1}|P_{t-1}, U_{1:t-2})$ is known at $t - 1$. Use estimated probability to calculate the predicted probability $P(s_t|U_{1:t-1})$ using Junction Tree inference engine.

Rolling: Delete the SBN for $t - 1$ and use $P(s_t|U_{1:t-1})$ as a new prior for SBN at t

Estimation: Observe P_t and calculate estimated probability distribution $P(s_t|P_t, U_{1:t-1})$. Add SBN for time at $t + 1$.

Chapter 8

Data Annotation

Continuous Emotion Recognition is to be robust enough to capture real life scenarios. Real life scenarios contain spontaneous facial behavior. Modeling spontaneous facial behavior is the key challenges in analyzing emotions. Lack of spontaneous facial data is one of the most important reason for slow progress in this area. Spontaneous facial behavior includes various combinations of AUs that can be different from combinations in posed expressions. In deliberate expression data-set such as Cohn-Kanade dataset [26], subjects are asked to display certain facial expressions. Moreover, they are instructed to display single or combination of AUs. In the situation where subjects are asked to display facial behavior in these ways, spontaneous facial behavior is rare. Activations and relaxing of AUs are experiencing much more artificial variations rather than naturalistic ways. Posed expressions are different in terms of activation time and appearances from spontaneous expressions [25]. Moreover, while some people can perform some AUs voluntarily, many other can do that in spontaneous fashion [26]. In addition, the temporal evolution AUs are different from those in prototypic and posed expressions. The reason is that the transitions between AUs do not include neutral all the time. Ideally, in most of the existing well published facial expression data-set, it is assumed that the input expressions are isolated or pre-segmented, showing a single temporal activation pattern (*onset* – *>* *apex* – *>* *offset*) of either a single AU or an AU combination that begins and ends with a neutral expression. In spontaneous behavior, such segmentation is not available; facial expressions are more complex and transitions from an action or combination of actions to another does not have to involve intermediate neutral state. AU 12 (Lip Corner Puller) can be followed by AU 20 (Lip Stretcher) immediately without full relaxation of AU 12. AU 9 showing disgust can be preceded by the activation of AU 1 and AU 2.

To cope with the real situation, recognition systems must be able to model spontaneous facial behavior. However, spontaneous facial expression data is rare compared to posed expressions. Unavailability of facial behavior data in natural environment derives us to collect and annotate data in a natural situation. Additionally, since AUs are the scope of the thesis, spontaneous facial behaviors should be coded with FACS.

8.1 Data Collection

Generally psychologists collect emotion data using "Emotion Elicitation" method. There are a number of emotion elicitation procedures followed by psychologists in laboratory setting. The most used methods include a) interaction with trained personnel such as interview, group discussion; b) hypothesis ; c) facial muscle movements; d) imaginary; e) music; and f) music. Among them, emotion elicitation with films has couple of advantages in the context of spontaneous facial behavior [27]. Emotion Elicitation from films facilitates to evoke emotions in a dynamic way rather than static. Visual and auditory stimuli from film help the subjects to show actual emotions without deception. Since watching film is a continuous process, the subjects undergo through a series of emotional processes that are also continuous with time. The emotional reactions are expressed mainly through facial behavior and the body postures since most of the time subjects are not allowed to talk.

In the data collection procedure, 21 subjects are observed when they watch movie clips. Six movie clips are chosen carefully to provide sufficient visual and auditory stimuli in order to arouse spontaneous emotional reactions. Emotional reactions include the changes of facial behavior. Actually, each of the subject has been asked to watch one movie clip in a large screen. One frontal video camera has been used to capture the face video for the entire watching period. The video camera has been placed just under the large screen to maintain frontal view of the whole face.

The whole procedure has been described by Yeasin et al. [7]. To retain the natural quality of the data, subjects are kept unaware of the real purpose of the data collection. To ensure their concentration in the experiment, they are informed that their eye movements would be recorded. Each subject has watched particular movie clip for 24 minutes. Therefore, each face video is 24 minute long associated with each subject. In the data annotation section, we use the term "Face Video" constantly to indicate the recorded face video for particular subject.

8.2 Data Annotation (FACS Annotation)

Since, the work [7] has been done with facial expression recognition, the video sequences of the subjects are annotated with six basic expressions. Though, the annotated video sequences are used for expression recognition, they are rich in terms of the AUs. Throughout the face videos, different AUs are activated to build up the expressions. Since, subjects show their emotional reactions over the whole period, annotation with FACS will provide a very rich diverse AU-coded data set with different AU combinations. Consequently, spontaneous facial behavior of the face can be annotated with FACS to provide the test beds of our emotion research. Six face videos have been chosen for the annotation. To

retain the cultural and racial variations, we take six different subjects from different races, cultures, and geographic regions shown in table 8.1.

Table 8.1: Subjects

| <i>Subject</i> | <i>Ethnecity</i> |
|----------------|------------------|
| Subject 1 | Indian |
| Subject 2 | Chinese |
| Subject 3 | American |
| Subject 4 | Russian |
| Subject 5 | African-American |
| Subject 6 | European |

The face videos are segmented in smaller clips to make the annotation process easy and simple. Segmentation is also a time consuming task. And, the quality of the segmentation will affect the scoring procedure in later phase. High quality segments ensure the reduction of complexities and coding time. The quality of the segmentation is defined in terms of the position of the segment and the length of the segmented clip. In order to reduce complexities and coding time, segmentation has been done by using following guidelines.

1. Use Video Cutter and Splitter In Depth 1.2 to segment the videos. It s a free software product. It has several advantages to get information in frame level.
2. Segmented video clips should be as small as possible. If possible, only one facial expression should be in one clip.
3. If possible, every clip will start with neutral face then it will contain expression and again neutral at end.
4. Cropping a clip with in expression is not allowed.
5. If possible, a clip must have *neutral* – > *onset* – > *apex* – > *offset* – > *neutral* states of the targeted facial expression
6. Start and end position of each clip must be recorded for further use..

The face videos were segmented into 3-20 seconds clips. After the segmentation, we have got 500 video clips. Some of them are discarded because no AU activations are taking place in them (empty clips). After throwing out empty clips, we have 415 video clips. Most of the clips are 3-10 second

long. They are responsible for one expression. Five percent of the clips are 15-20 second long. Long clips contain two or more successive emotional events that couldn't be separated. Careful observation reveals that 10-15 second long clips contain long expressions with subtle changes in the facial features. Though subtle changes do not cause substantial changes in the expressions, they are responsible for diverse facial behavior. So, the length of the clips hold a great attention while AU coding continues.

Trained judge has been exploited to annotate the frames in the clips according to FACS. The judge has the expertise on the FACS manual. Along the coding procedure, trained judge is not aware of the emotion in the clips. The judge concentrates only to the AU states. That is, clips are not associated with emotional terms. Also, the judge is not trained with emotion annotation.

8.3 Scoring Procedure

Trained judge watched each clip and recorded the AUs when they occur. The judge was flexible to pause the video clip anywhere and to search for the AUs. He could go back and forth to identify the AU activation effectively. Since the intensities are not targeted for the experiments, judge recorded the AUs when they were in the apex. The judge recorded the frame number to use the information in the further study. Since spontaneous behavior is being coded through FACS, onset and offset frames for targeted AUs have not been recorded. The judge coded 64 AUs according to the FACS manual.

We know that about 20 AUs listed in [19] are responsible for different kinds of facial expressions. However, in the above procedure, more complex emotional states such as amusement, contentment, boredom, confusion etc are elicited. Therefore, only 20 AUs are not sufficient to analyze the spontaneous facial behavior during watching movies. This is the strong motivation to code the head movements, eye positions and eye movement, and other miscellaneous AUs.

8.4 Data Treatment and Analysis

Reliability of the annotation is a major concern in establishing ground truth data. The mostly used measure of reliability for annotation is having observers independently code a portion of the same data. Then agreement between the observers are calculated in terms of kappa statistics. The coefficient kappa is preferable to raw percentage of agreement between observers. Generally Kappa quantified inter-observer agreement after correcting for level of agreement expected by chance. In the area of FACS annotation, manual AU coding is time consuming and labor intensive job. Therefore, generally AU coding for the whole data-set is done by one trained judge. Then 15-20 percent of data will be coded by another independent trained judge. After that, Kappa statistics will be calculated to measure

the reliability. For our annotation task, randomly selected 50 percent of the data has been chosen to be coded by second trained judge. The second judge will be completely unaware of the annotation by first judge.

Table 8.2: Proportion Of Facial Action Units

| Category | Facial Action Unit | Description | Percentage |
|--------------------------|--------------------|-------------------|------------|
| Upper Face | AU 1 | Inner Brow Raiser | 13.2 |
| | AU 2 | Outer Brow Raiser | 9.5 |
| | AU 4 | Brow Lowerer | 14.19 |
| | AU 6 | Cheek Raiser | 23.76 |
| | AU 7 | Lid Tightener | 6.9 |
| | AU 43 | Eye Closure | 23.43 |
| Lower Face | AU 9 | Nose Wrinkler | 11.22 |
| | AU 14 | Dimpler | 6.60 |
| Lower Face-Mouth and Lip | AU 10 | Upper Lip Raiser | 25.08 |
| | AU 12 | Lip Corner Puller | 43.89 |
| | AU 14 | Dimpler | 6.60 |
| | AU 20 | Lip Stretcher | 7.59 |
| | AU 24 | Lip Pressor | 8.25 |
| | AU 25 | Lips Part | 37.75 |
| Head Position | AU 55 | Head Tilt Left | 2.97 |
| | AU 56 | Head Tilt Right | 12.87 |

Since, annotation by second judge has not been performed yet, the following section describes the statistical analysis of the first annotation. Only 10 percentage of the data frames contain single AU. 39 percent of the data frames include the combination of AUs that are quite responsible for various facial expressions. Twenty four percent data frames include AUs that are associated with head-eye positions as well as movements.

Table shows the raw percentage of frequent AUs. Frequent AUs come in more data frames rather than the infrequent AUs. The AUs that are associated not only with prototypic facial expressions but also complex emotional states. Thus, the coded data-set is rich for both facial expressions and complex emotional states. Complex emotional states such as contentment, frustration, delight, and confusion are associated with AU 1 , AU 2 , AU 4, AU 7, AU 12, AU 25, AU 43, AU 55, and AU 58. So, the data-set can be used as test bed for expression analysis as well as mental affect judgements.

Chapter 9

Experimental Results

Empirical analyses using a number of different datasets consisting of varying degrees of complexities and variabilities were used to illustrate the utility of the proposed approach. In particular, four different datasets, **CK+** dataset (posed expressions [9]), Bosphorus dataset (posed 3D expressions [28]), M and M Initiative (**MMI**) dataset (mixture of posed and natural expressions [12]), and Emotion Elicitation (**EE**) dataset (natural expressions [7]) were used for empirical analyses. The key objectives of the experiments are as follows:

1. To show the utility of the scoring function in logically dividing the AUs into FSAUs and IWAUs using diverse datasets.
2. To show the robustness of predicting the IWAUs from the FSAUs and to compare and contrast the performance with the recently reported literatures.
3. To study the effect of perturbation at the level of FSAUs and their impacts on the robustness in predicting IWAUs as well as categorical emotions.

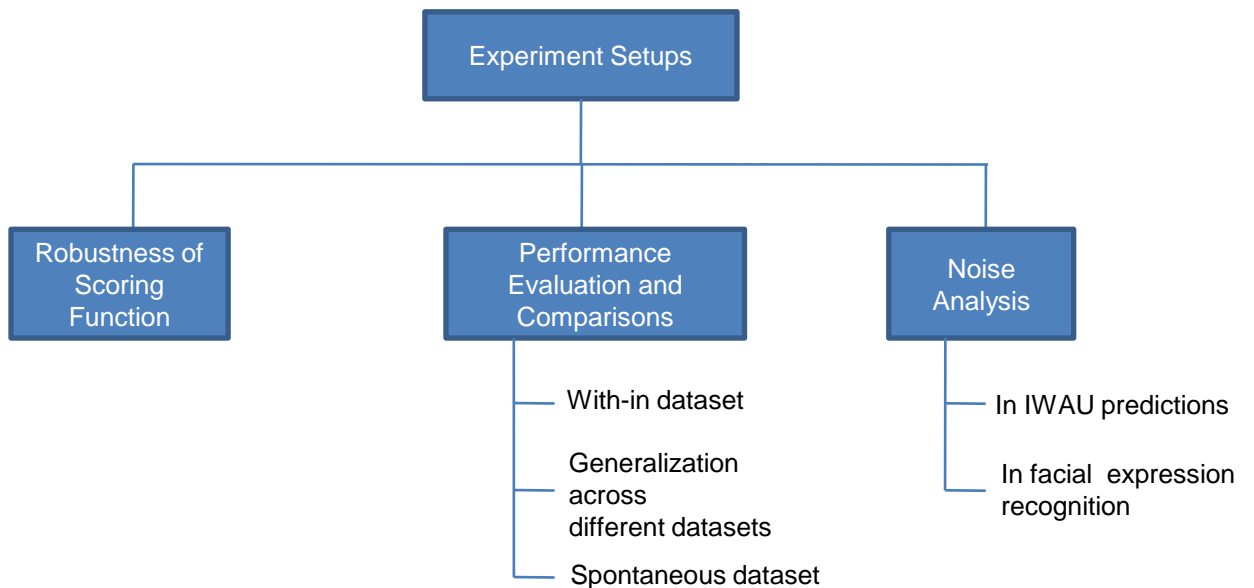


Figure 9.1: Experiment Setups

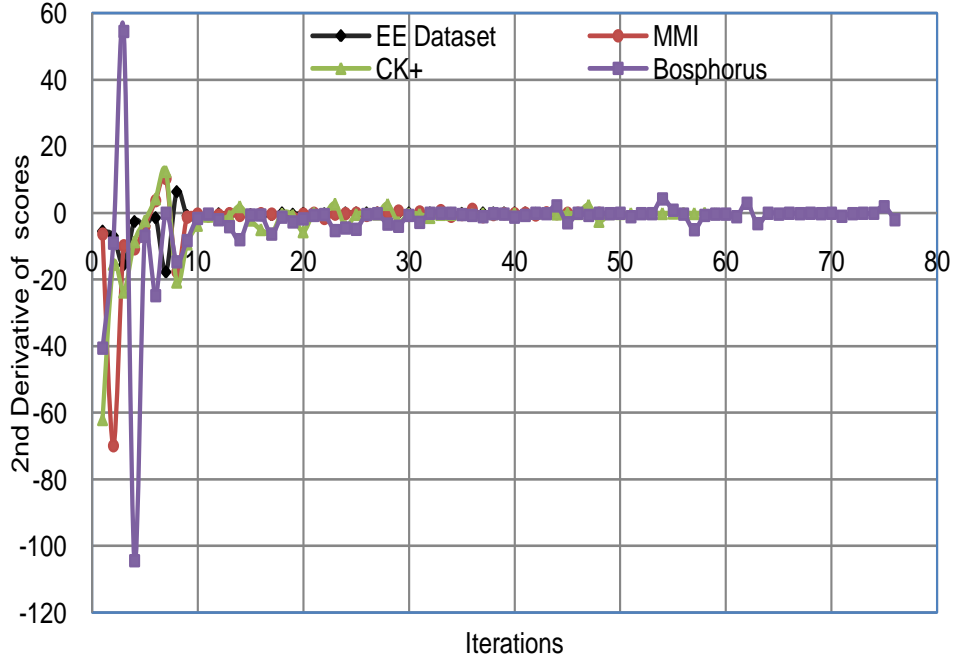


Figure 9.2: 2^{nd} derivative of scores

9.1 Databases

CK+[9] includes 593 video sequences from 123 subjects. Enhanced Cohn-Kanade AU-coded dataset [6] provides codes for 14 AUs frame by frame. Accordingly, the remaining 9 AUs were coded manually frame by frame basis by the authors. Second database used is **Bosphorus dataset** [28] that is intended for research on 3D and 2D facial image analysis. It contains 105 subjects and a maximum 35 different expressions are captured. The dataset was acquired by varying lighting condition, occlusion, and poses. **MMI Facial Expressions database** by Pantic *et al.* [12] is employed to test the generalization across people and facial expressions. The fourth dataset is known as the **Emotion Elicitation (EE)** dataset and was captured using “*Emotion Elicitation by Films*” [7]. There is 21 subjects and the videos are annotated according to the FACS manual. This dataset provides an opportunity to study the robustness of the proposed method for natural facial behavior. This dataset also includes frequent out of plane head movements, occlusions, and varying lighting condition.

9.2 Utility of Scoring Algorithm in Logical Division of AUs

It is widely acknowledged that the FACS provides a mechanism for analysis and synthesis of facial behavior that is consistent across culture, ethnicity, gender, and age groups. Hence, relations among the AUs are expected to be similar across datasets. The scoring proposition described in the Chapter 5 was used to compute the relations among AUs. The trends in the rate of change of scores over iterations were used as an indicator to divide the AUs into FSAUs and IWAUs. Figure 9.2 shows the 2^{nd} derivative of the scores over iterations computed during the SBN structure learning process. From figure 9.2 it is easy to note that the “Buildup Area” and “Tuning Area” were divided after 8^{th} iteration. All four datasets were used in the experimentation to check consistency in dividing the AUs into FSAUs and IWAUs. It was observed that the strong relations were found after 8^{th} iteration - these were **identical** across the datasets. It was also observed that few weak relations were different across the datasets due to a number of variabilities. However, generalization of the strong relations and FSAUs make the probabilistic model robust in IWAU prediction.

9.3 IWAUs Predictions

Subsequent paragraphs illustrate the versatility and the robustness of predicting the IWAUs from the FSAUs. Since the most of the state-of-the-art AU recognition works focus on *ad hoc* process to select subset of AUs, it is difficult to compare the reported results with the proposed work. However, a number of reported techniques that have recognized most of the IWAUs were used for the comparison. In these experiments, FSAUs from ground truth are used. To illustrate the efficacy of the proposed approach, noise have been added to the ground truth data of FSAUs in a way that input to the DBN incorporates the same errors as the reported literature presents. For example, to compare with CK+ dataset, ground truth of FSAUs are corrupted to produce following recognition performance. Since error in the FSAUs will propagate through the DBN, incorporating noise accordingly make more sense in comparing the prediction performance of IWAUs. Noisy FSAUs with following recognition performance are generated by randomly flipping the FSAUs with particular probabilities.

Three different validation experiments were performed. Firstly, performance evaluation and comparison with recently reported literature were performed using the CK+ dataset. Secondly, generalization of the proposed approach was tested using MMI dataset while the DBN had been trained with the CK+. Furthermore, EE dataset [7] was used to illustrate the suitability of the proposed approach in dealing with spontaneous emotion in an uncontrolled environment.

Table 9.1: Recognition Performance

| AUs | ROCA |
|-----|------|
| 2 | .975 |
| 4 | .925 |
| 6 | .975 |
| 7 | .925 |
| 12 | .975 |
| 15 | .925 |
| 17 | .95 |
| 23 | .85 |
| 24 | .92 |
| 25 | .96 |
| 27 | .99 |

9.3.1 Empirical Analysis using Posed Expressions

The CK+ was used to quantify the performance of the proposed system and also to perform comparative analysis with [29]. The area under the Receiver Operating Characteristic (ROC) Curve was used as a metric and was abbreviated as ROCA. FSAUs from CK+ were perturbed to add noise in such a way that their recognition performance followed the Table 9.1 that is reported in [29]. Then, noisy FSAUs were fed to the DBN and subsequently IWAUs were predicted.

Figure 9.3 showed the comparative analysis of prediction accuracies of IWAUs. Numbers within the blue bars indicated the ratios between positive and negative samples. Figure 9.3 suggests that ROCAs for AU 1, 5, 10, 11, 16, 20, and 26 were increased compared to [29]. Particularly, performance of AU 10 was increased by 19%, AU 11 by 9%, AU 16 by 8.64%, and AU 26 by 12.5%. It indicates that the performance of the proposed approach is significantly better compared to concurrent methods even though IWAUs were predicted without any computer vision techniques. The key observation is that the particular AUs that are difficult AUs (AU 10, 11, 14, 16, 20, and 26) were predicted with significant ROCA. Moreover, AU 11, 16, 18, and 38 have ROCA more than 0.80 that are neglected by most of the AU detection techniques. Another comparative analysis was performed with the work reported by Tong *et al.* [6] to provide additional perspectives in the light of dynamic modeling of AUs where the True Positive rate (TPR) of AU 1, 5 and 9 are 0.8, 0.75 and 0.9, respectively. The corresponding number for the proposed approach are 0.93, 0.97, and 0.98, respectively. Clearly, the proposed approach outperformed [6] in predicting AU 1, 5, and 9.

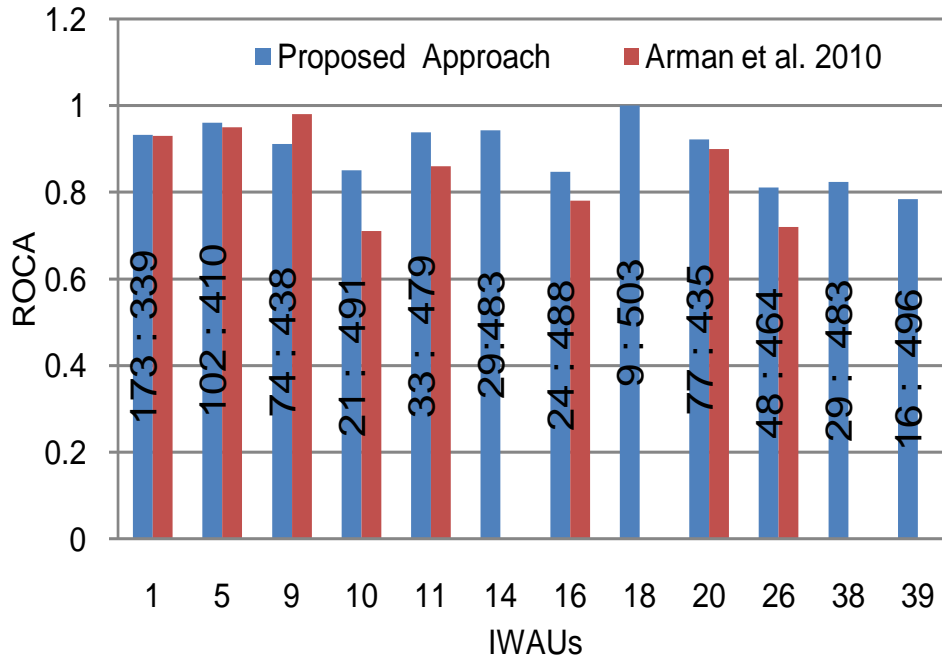


Figure 9.3: ROCA for 3-fold cross validation with CK+

9.3.2 Generalization using Mixed Expressions

MMI dataset was used to test the generalization of the proposed approach. This experiment used CK+ dataset for training the DBN and the MMI dataset for testing while testing set included the samples that have more than one AU activated simultaneously. In this experiment, F-measure was used as performance metric. The input to the DBN were corrupted to produce FSAUs with F-measures (table 9.2) reported in [30].

Table 9.2: Recognition Performance of FSAUs in [30]

| AUs | F-Measure |
|-----|-----------|
| 2 | .727 |
| 4 | .693 |
| 6 | .737 |
| 7 | .364 |
| 12 | .622 |
| 15 | .56 |
| 17 | .765 |
| 23 | .412 |
| 24 | .44 |
| 25 | .847 |
| 27 | .96 |

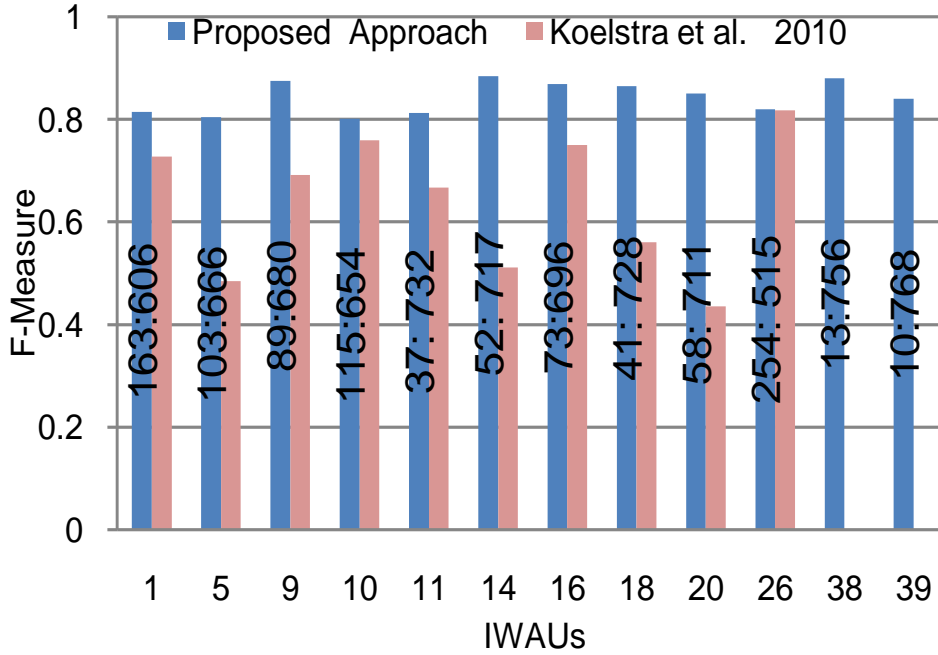


Figure 9.4: Recognition Performance of IWAUs for MMI dataset

Comparative results between the proposed approach and [30] were shown in the figure 9.4. It was found that all IWAUs achieved significant increase in F-Measure. The average F-measure for the proposed approach was 0.8429 whereas [30] got 0.6404. The F-measure of the difficult AUs AU 11, 14, 18, and 20 were reasonable enough compared to the state-of-the-art techniques. Furthermore, the proposed approach outperformed the reported literature even though the training and testing datasets were different. While compared with the dynamic modeling of AUs [6], AU 1, 5, and 9 received higher TPR (0.85, 0.86, and 0.89, respectively). This also indicates a better generalization of the proposed approach.

9.3.3 Analysis with Natural Expressions

To further illustrate the utility of the proposed method in the real life scenario, an experiment was performed using the Natural Continuous Emotion dataset that contains spontaneous facial behavior. In this experiment, CK+ database was used for training while the Natural Continuous Emotion dataset was used for testing. Comparison of the performance was made with spontaneous facial expression dataset (FACS-101) collected by Mark Frank [11]. The results for FACS-101 are reported in [31]. Though different datasets have been used for comparison, the figure 9.5 presents the over all superiority of the

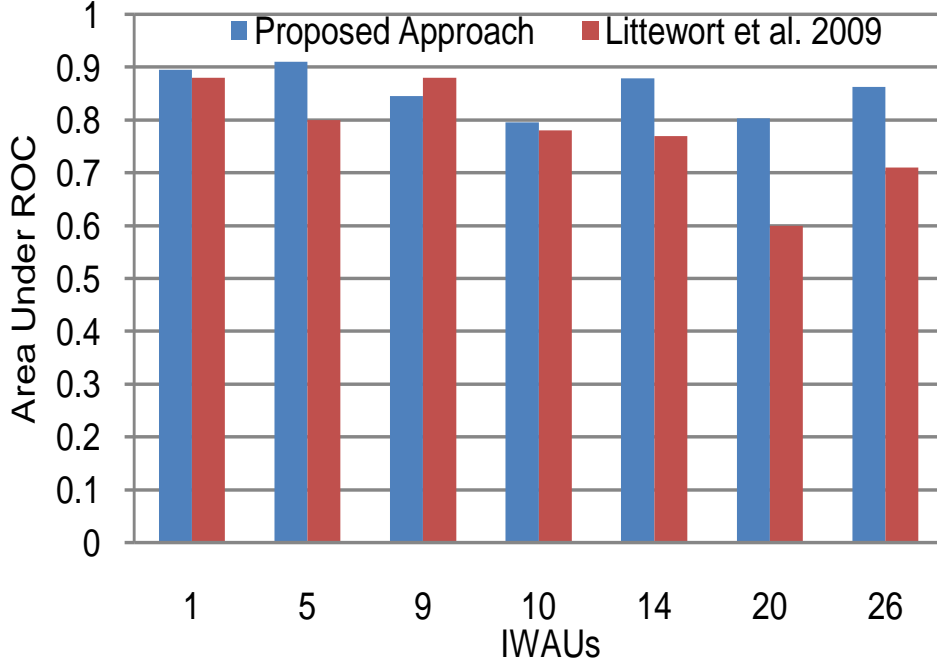


Figure 9.5: Performance evaluation for IWAUs in spontaneous datasets

proposed approach in predicting IWAUs from FSAUs using AU relations. In particular, ROCA of AU 14, 20, and 26 have been increased 14%, 33.8% and 21.45%, respectively although these AUs are considered difficult and error prone. However, AU 10 has relatively lower ROCA. The plausible explanation for this is due to the mixed emotion.

9.4 Performance Analysis with Perturbation

Though state of the art technologies have achieved significant improvement in the AU recognition, noise is added due to the inaccuracies in measurements of subtle facial deformation, pose, and out of plane head movements. A number of studies were performed to study the effect of perturbation both at the AU level as well as in predicting categorical emotion. In particular, FSAUs were perturbed to incorporate different amount of noises and then fed to the trained DBN.

9.4.1 Noise Analysis in Predicting IWAUs

In the context of significant AUs (FSAUs), these noises are expected to propagate through the proposed DBN while inferring the insignificant AUs (IWAUs). However, the dependencies of IWAUs on FSAUs are not same due to physical constraints of face. Thus, not all IWAUs are susceptible to noise in the same

scale. Experiments are expected to find out the IWAUs that are less susceptible to error (noise tolerant IWAUs). The goals of the following experiments are to determine the upper limit of noise that can be allowed in FSAU recognition. Moreover, the effect of noise has been also analyzed to illustrate the efficacy of proposed approach with tolerable amount of error in FSAU recognition.

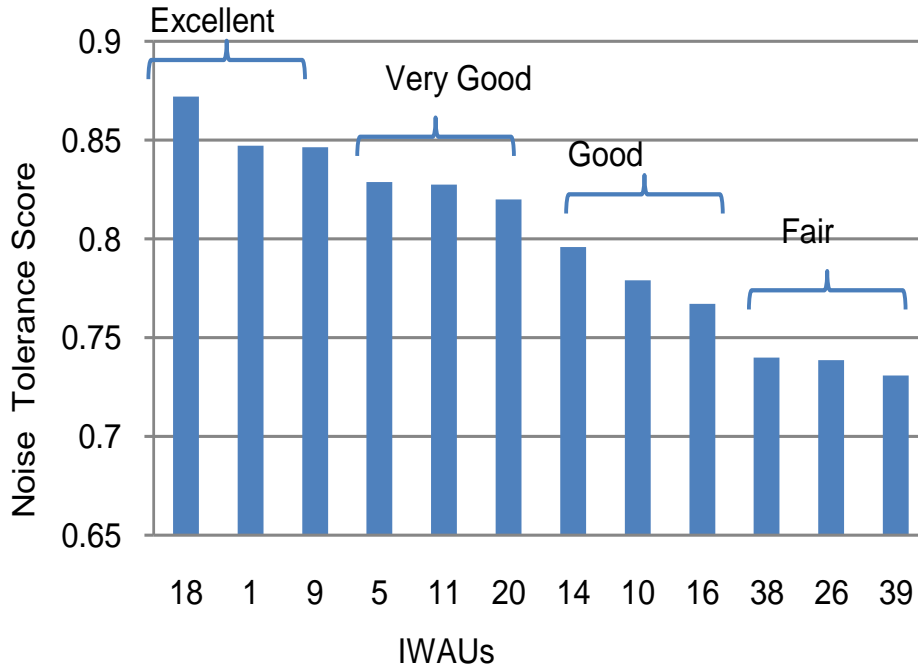


Figure 9.6: Noise Tolerance Score

Noises from identical uniform distribution were added to each of the FSAUs and subsequently noisy FSAUs were used for IWAU prediction. Figure 9.7 depicts the prediction results of IWAUs with varying degrees of noise. To provide a better insight, a measure to calculate the “noise tolerance” was introduced. The approach was to find the RMS error of the performance graph from the baseline performance (with noise 0%). After that, RMS error was subtracted from theoretically maximum possible RMS error for each IWAU. Thus, the result can be interpreted as a measure of noise tolerance where ideal score should be 1.00. To provide better insight and visualization, IWAUs were grouped into four clusters (excellent, very good, good, and fair) based on the tolerance score (fig 9.6).

It was observed that AU 18, 1, and 9 had excellent noise tolerance in figure 9.7a. Among them, AU 18 seemed to be mostly tolerant against noise. It could tolerate up to 11% noise before going down to ROCA 0.8. It was also noted that noise had linear effect on the prediction of AU 1 and 9. The degradation of their performance was gradual as the ROCA remained above 0.90 and above 0.85 while

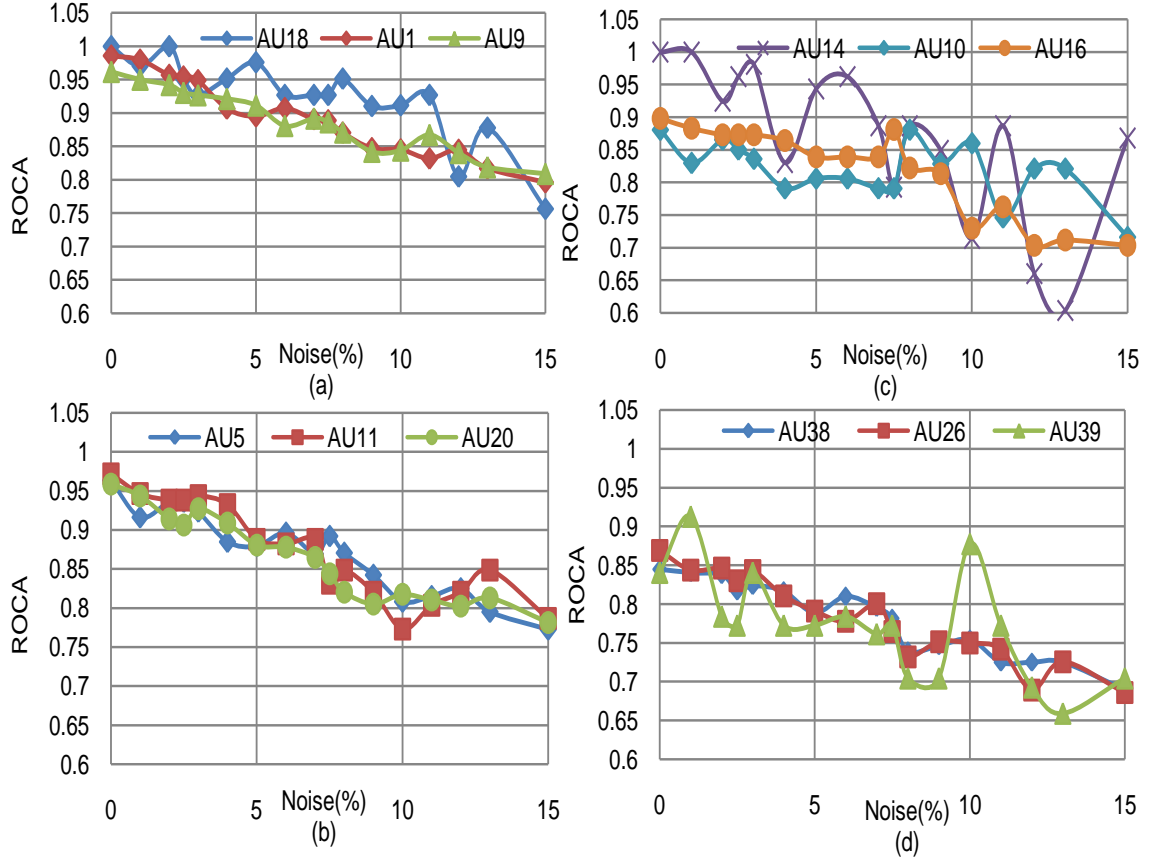


Figure 9.7: Performance of IWAU Prediction against Noise

the noise were increased to 7% and to 12%, respectively. AU 5, 11, and 20 showed very good noise tolerance as observed from the figure 9.7b. Average ROCA of this group remained above 0.9 for noise up to 4% while average ROCA went below 0.80 after 10% noise. The AU 14, 10, and 16 was found to have good tolerance (fig 9.7c) and the other AUs showed fair tolerance. For fair tolerant IWAUs (fig 9.7d), more than 4% noise resulted inferior ROCA on average.

9.4.2 Noise Analysis in Facial Expression Recognition

An additional experiment was performed to analyze the effects of perturbation/noises in FSAUs on recognition of six categorical emotions. CK+ and MMI database were used for this experiment while emotion recognition was performed on 18 AUs (11 FSAUs and 7 IWAUs). FSAUs are perturbed accordingly to generate noisy FSAUs as earlier experiments. Then, IWAUs were inferred using proposed technique while another Bayesian net [10] had been used to predict the emotions. Bayesian

net 9.8 has been used to predict the facial expressions using 18 AUs [10]. Unlike [10], parameter distributions are learned using CK+ and MMI datasets since these datasets are annotated with six basic expressions. The research works [32] describes some empirical data to relate the AUs combination to facial expressions. These data in [32] are used as prior in parameter learning process.

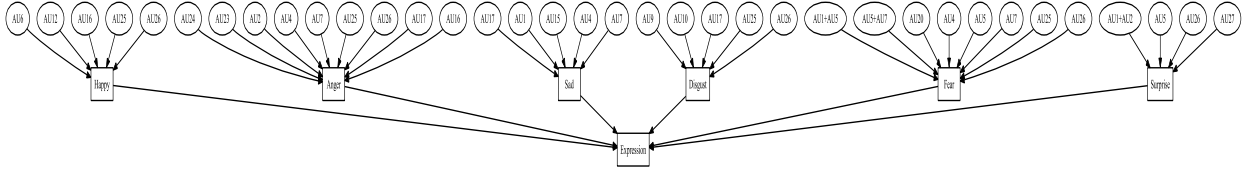


Figure 9.8: Expression Bayesian Net

It was observed from figure 9.10 that '**Happy**' emotion had the best noise tolerance against noisy FSAUs. The performance was found to be almost consistent as the ROCA remained above 0.95 up to 29% noise in FSAUs with one exception. '**Surprise**' expression showed convincing error tolerance as well. It maintained ROCA more than 0.90 across the whole noise range. Though '**Disgust**' had inferior performance after 9% noise compared to Happy and Surprise expressions, its ROCA was more than 0.85 up to 23% noise level with two exceptions only. Performance in predicting '**Anger**' found to be somewhat better than 'Disgust'. The ROCA of Anger recognition was more than 0.90 where noise increased from 0% to 7% and ROCA remained more than 0.80 up to 28% noise. '**Sad**' has reasonable noise tolerance up to 24% noise while more than 0.85 ROCA has been maintained. Therefore, though the IWAUs have less noise tolerance, expression recognition using the proposed approach is less susceptible to noise in FSAUs.

Average ROCA graph shows that the performance continues to reasonable scales (more than 0.9 ROCA) up to 10% noise. However, 24% noise is allowable to maintain the ROCA more than 0.85. Therefore, though the IWAUs have less noise tolerance, expression recognition using the proposed approach are much less susceptible to noise in FSAUs. Two layer of bayesian networks are the example of convincing use for noise canceling mechanisms in the two level of abstractions.

However, the performance of Disgust fluctuates with varied noises. The reason is that Disgust depends on 3 IWAUs and 2 FSAUs. Moreover, 3 IWAUs are AU 9, AU 10, and AU 26. Among them, AU 10 has two parents and AU 26 has only one parent. So, small number of parents are responsible for propagating errors in predicting corresponding IWAUs. Since, AU 10 and AU 26 have small number of parents, their performance are largely susceptible to noise in FSAUs. Consequently, performance of disgust is affected by noise. Another important observation is that sometimes performance increases

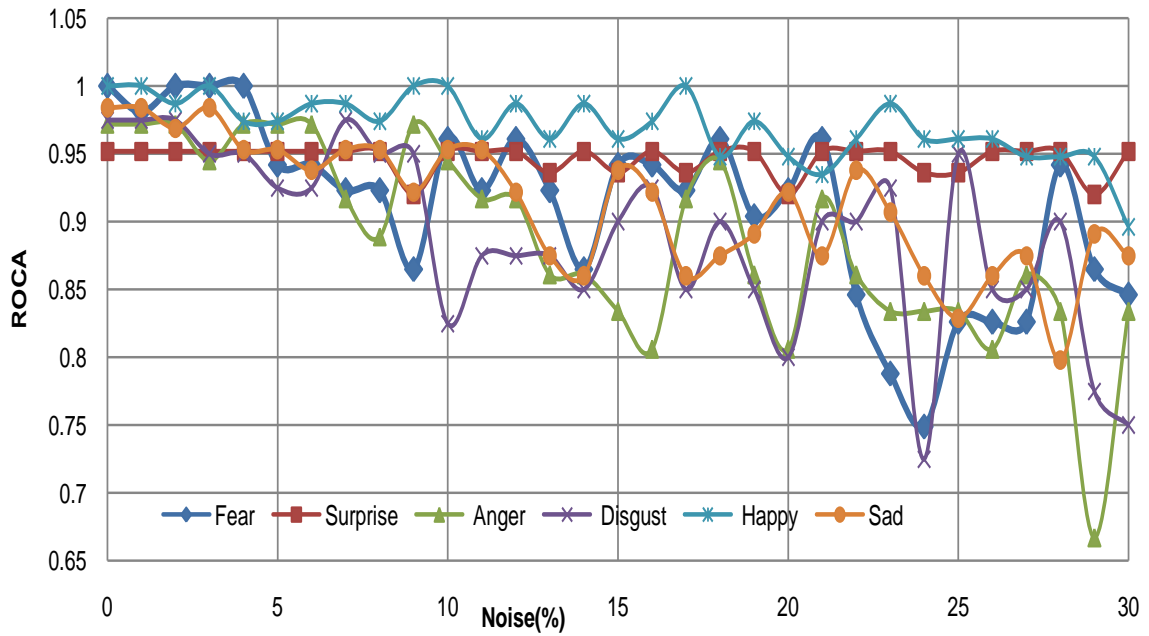


Figure 9.9: Performance of Facial Expression Recognition against Noise

sharply. Actually, noises are added randomly to 11 FSAUs by flipping them. With small probability, noisy FSAUs follow the mostly compatible instance within the relation that change the performance highly. The mostly compatible relations are responsible for correct prediction of IWAUs that in turn improves the performance sharply. The opposite event occurs when the least compatible instance is formed. However, these sharp changes are occurred with small probabilities since flipping probabilities are insignificant

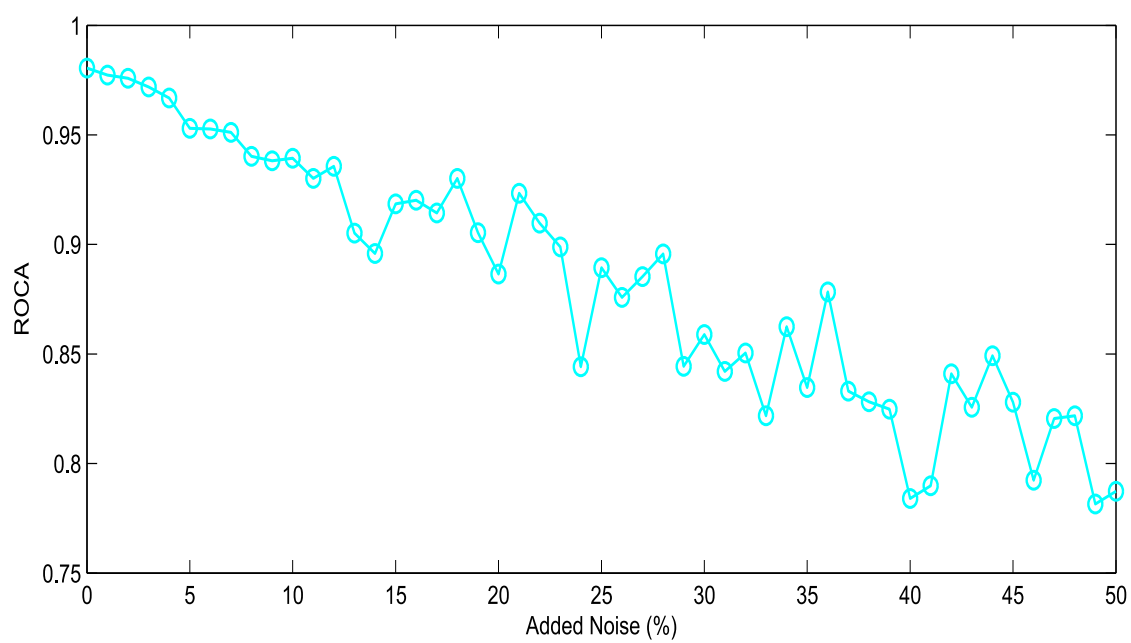


Figure 9.10: Average Performance of Facial Expression Recognition

Chapter 10

Conclusion

Reliable estimation of affective states and continuous categorical emotion spotting require all AUs or a majority of them. However, automated, robust, and real-time recognition of all AUs are computationally expensive and error prone. To address such issues, this paper proposed a spatio-temporal data driven probabilistic scoring function to divide the AUs into FSAUs and IWAUs. SBN was used to capture the relations among AUs and the DBN was used to capture their temporal evolution. A framework was implemented to predict the IWAUs on the fly from the FSAUs with very high accuracy. The proposed approach contributed in significant reduction of the run time and improved robustness in predicting IWAUs. In addition, perturbation analysis was performed to understand the effect of noise at both the AU level as well as recognition of categorical emotion. These contributions will enable real-time analysis, synthesis, and tracking of complex and natural facial behaviors.

Future Work:

1. Implementation of a real-time software system integrating FSAU recognizer, IWAU predictor, and modeling facial expressions with intensity.
2. Affective states (Confusion, Boredom, Flow, Engagement, Frustration) prediction from AUs.
3. Studying AU relations in facial expression during learning.
4. Analyzing dynamic modeling of AUs while subject is speaking
5. Clustering AU relations based on physiological constraints and localization of the proposed approach.

Bibliography

- [1] <http://www.cs.cmu.edu/afs/cs/project/face/www/facs.htm>.
- [2] P. Ekman and W. V. Friesen, "Facial action coding system: A technique for the measurement of facial movement," Palo Alto, CA, 1978.
- [3] S. D'Mello, R. W. Picard, and A. Graesser, "Toward an affect-sensitive autotutor," *IEEE Intelligent Systems*, vol. 22, pp. 53–61, 2007.
- [4] V. Alevén, B. McLaren, I. Roll, and K. Koedinger, "Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor," *Int. J. Artif. Intell. Ed.*, vol. 16, no. 2, pp. 101–128, 2006.
- [5] W. V. F. P. Ekman and J. C. Hager., "Facial action coding system," Research Nexus, Network Research Information, Salt Lake City, UT, Tech. Rep., 2002.
- [6] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Trans. PAMI*, no. 29, p. 1699, 2007.
- [7] M. Yeasin, B. Bulot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE Trans. on Multimedia*, 2006.
- [8] R. W. Picard, "Affective computing: challenges," *Int. J. Hum.-Comput. Stud.*, vol. 59, no. 1-2, pp. 55–64, 2003.
- [9] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *CVPR Workshops*, jun. 2010.
- [10] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Trans. on PAMI*, no. 27, pp. 699–714, 2005.
- [11] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.
- [12] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *ICME*. IEEE, 2005, pp. 317–321. [Online]. Available: <http://doi.ieeecomputersociety.org/10.1109/ICME.2005.1521424>
- [13] S. Lucey, A. Ashraf, and J. Cohn, "Investigating spontaneous facial action recognition through aam representations of the face," in *I-Tech Ed. and Pub.*, 2007, pp. 275–286.
- [14] Y. Tong, J. Chen, and Q. Ji, "A unified probabilistic framework for spontaneous facial action modeling and understanding," *PAMI, IEEE Transactions on*, vol. 32, no. 2, pp. 258–273, 2010.
- [15] Y. TIAN, T. KANADE, and J. COHN, "Recognizing action units for facial expression analysis," *IEEE Trans. on PAMI*, no. 23, p. 115, 2001.
- [16] J.-J. J. Lien, T. Kanade, J. Cohn, and C. Li, "Detection, tracking, and classification of action units in facial expression," *Journal of Robotics and Autonomous Systems*, July 1999.

- [17] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Trans. on PAMI*, no. 21, pp. 974–989, 1999.
- [18] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology*, no. 36, 1999.
- [19] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully automatic facial action recognition in spontaneous behavior," *Journal of Multimedia*, no. 6, pp. 22–35, 2006.
- [20] A. Kapoor, Y. Qi, and R. W. Picard, "Fully automatic upper facial action recognition," *IEEE Intl. Workshop on Analysis and Modeling of Faces and Gestures*, vol. 0, p. 195, 2003.
- [21] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Comput. Vis. Image Underst.*, vol. 91, no. 1-2, pp. 160–187, 2003.
- [22] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. de la Torre, and J. Cohn, "Aam derived face representations for robust facial action recognition," in *7th International Conference FGR*. IEEE Computer Society, 2006, pp. 155–162.
- [23] <http://face-and-emotion.com/dataface/facs/description.jsp>.
- [24] G. F. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Mach. Learn.*, vol. 9, no. 4, pp. 309–347, 1992.
- [25] K. Schmidt and J. Cohn, "Dynamics of facial expression: Normative characteristics and," in *Individual Differences, Proc. IEEE ICME*, 2001, pp. 728–731.
- [26] T. Kanade, Y. Tian, and J. F. Cohn, "Comprehensive database for facial expression analysis," in *Fourth IEEE International Conference on FGR 2000*. Washington, DC, USA: IEEE Computer Society, 2000, p. 46.
- [27] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition and Emotion*, no. 9, pp. 87–108, 1995.
- [28] A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Biometrics and identity management," B. Schouten, N. C. Juul, A. Drygajlo, and M. Tistarelli, Eds. Berlin: Springer-Verlag, 2008, ch. Bosphorus Database for 3D Face Analysis, pp. 47–56. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-89991-4_6
- [29] S. Savran, B. Sankur, and M. T. Bilge, "Facial action unit detection: 3d versus 2d modality," in *IEEE CVPR Workshop on Human Communicative Behavior Anal.*, USA, 2010.
- [30] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *PAMI, IEEE Trans. on*, vol. 32, no. 11, 2010.
- [31] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image Vis. Comput.*, vol. 27, pp. 1797–1803, Nov. 2009. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1621144.1621315>
- [32] M. Pantic and L. Rothkrantz, "Expert system for automatic analysis of Facial Expression," *Image and Vision Computing Journal*, vol. 18, no. 11, pp. 881–905, August 2000. [Online]. Available: <http://pubs.doc.ic.ac.uk/Pantic-IVCJ00/>