The Thesis committee for Suyog Dutt Jain
Certifies that this is the approved version of the following thesis

Facial Expression Recognition with Temporal Modeling of Shapes

APPROVED BY

SUPERVISING COMMITTEE:

J. K. Aggarwal, Supervisor

Kristen Grauman

Facial Expression Recognition with Temporal Modeling of Shapes

by

Suyog Dutt Jain, B.E.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master Of Science in Computer Science

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2011

Dedicated to my family

Acknowledgments

This work has been possible because of the following special people who have always been a source of inspiration for me to do research and pursue my dreams.

I offer my greatest regards to Professor J. K. Aggarwal for providing me the opportunity to work with him. His ideas have formed the foundations of this work and his continuous motivation, guidance and vision enabled me to explore the problem in right perspective and find solutions.

I also deeply thank Professor Kristen Grauman for introducing me to the world of computer vision research. Her teaching and discussions helped me in building a strong background which is invaluable for my research.

I would like to thank Dr. Changbo Hu, Birgi Tamersoy, Jong Taek Lee and all other Computer & Vision Research Center members who made my stay here a memorable one. I will always cherish the numerous philosophical discussions on evolution and existence which were a part of most lunches we had together.

I owe my wonderful time in Austin as a graduate student to my friends Yashesh, Nikhil, Akash, Deepak, Sameer, Pooja, Akanksha, Gouri, Lakshmi and Anushree who were always there with their support, love and encouragement.

I thank Esha for being there always, making my life beautiful. Without you this wouldn't be possible.

Words cannot describe the role that my parents have played in my life with their continuous support and blessings. I attribute all my success to them.

Facial Expression Recognition with Temporal Modeling of Shapes

Suyog Dutt Jain, M.S.C.S.
The University of Texas at Austin, 2011

Supervisor: J. K. Aggarwal

Conditional Random Fields (CRFs) is a discriminative and supervised approach for simultaneous sequence segmentation and frame labeling. Latent-Dynamic Conditional Random Fields (LDCRFs) incorporates hidden state variables within CRFs which model sub-structure motion patterns and dynamics between labels. Motivated by the success of LDCRFs in gesture recognition, we propose a framework for automatic facial expression recognition from continuous video sequence by modeling temporal variations within shapes using LDCRFs. We show that the proposed approach outperforms CRFs for recognizing facial expressions. Using Principal Component Analysis (PCA) we study the separability of various expression classes in lower dimension projected spaces. By comparing the performance of CRFs and LDCRFs against that of Support Vector Machines (SVMs) and a template based approach, we demonstrate that temporal variations within shapes are crucial in classifying expressions especially for those with small facial motion like anger and sadness. We also show empirically that only using changes in facial appearance

over time without using the shape variations fails to obtain high performance for facial expression recognition. This reflects the importance of geometric deformations on face for recognizing expressions.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Facial expressions are universal across all human ethnics and cultures. They can be defined as the facial changes which occur in response to a person's internal emotional states, intentions or social communications [26]. We as humans are generally very good at understanding a person's emotional state just by looking at the facial expressions of a person. It serves as an important component in our interaction with another person. We generally stress a lot on understanding a person's facial expression in order to understand the effects of our words and behavior on a person's emotional state. Dr. Paul Ekman introduced six basic expressions [17] and established that these expressions are universal. These are anger, disgust, fear, happiness, sadness and surprise (Figure 1.1 shows an example of these basic expressions). Human brain is effectively trained to recognize these expressions very quickly but it still remains a challenging problem for machines.

The automated recognition of facial expressions is a necessary first step for meaningful interactions between humans and computers. A reliable Automatic Facial Expression Recognition (AFER) system will improve the way in which humans interact with machines. Bruce [10] strongly argues that hu-

(a) Anger (b) Disgust (c) Fear

(d) Happiness (e) Sadness (f) Surprise

Figure 1.1: Examples of six basic facial expressions.

man face is key for social interaction. He explains that in order to have good human computer interfaces, its important for machines to understand human expressions and gestures. Mehrabian [31] indicates that facial expressions contribute around 55% to the effect that a message has on a person. The actual words only contribute 7%. which shows the importance of facial expressions in effective communication. Apart from designing better interfaces between human and computers, automated facial expression recognition find uses in many other applications like designing serious games for treating certain medical conditions like autism, for automobile safety, patient monitoring etc.

Facial expressions in humans are inherently dynamic in nature, consist-

ing of onset, peak and offset phases. The entire event from onset to offset is usually very short in duration and often the muscle motions on the face are very subtle. This makes the problem of recognizing facial expressions challenging. In this work, we consider the problem of recognizing facial expressions from video sequences and formulate it as a supervised sequence labeling problem, where we try to label every frame with the correct facial expression or neutral state. One way to approach this problem is to consider each image individually and train classifiers using them. The main limitation of this approach is that it does not consider the temporal dependencies between image features. These temporal dependencies have been shown to be of significant importance especially in identifying those facial expressions which are not characterized by large changes of shapes or appearance on the face[2].

We propose a new approach for recognizing six basic expressions (anger, disgust, fear, happiness, sadness, surprise) along with neutral state, by modeling temporal dynamics of face shapes. Our approach uses discriminative Latent-Dynamic Conditional random fields (LDCRFs) [32] and we show that incorporating hidden states in traditional Conditional Random Fields (CRFs) [25] model is an effective way to model the subtle changes which happen over time in face shapes. This helps in distinguishing between facial expressions which have very similar motion patterns.

We also empirically show that classifiers which use temporal variations between shapes outperform those which do not consider this information for the task of facial expression recognition. Finally we compare the ability of

appearance variability in time to recognize facial expressions with that of the shape variability. We show that variations in shape are much more important than appearance for facial expression recognition from continuous video.

The remainder of the thesis is organized as follows: We discuss the related work in the Chapter 2. Chapter 3 gives a detailed description of the features used in our work along with the various classification methods that were used in our work. In Chapter 4 we present all the experiments which were conducted along with a discussion on our results. We conclude in Chapter 5 and discuss possible future works.

Chapter 2

Related Work

Facial expressions have been the basis of human communication for centuries. Understanding the origins and causes of this strong medium of interaction with fellow species has been a topic of interest among the scientists. The oldest scientific work on this subject dates back to Charles Darwin [14]. He laid down the principles behind expressions and gave accounts to show their universality. Darwin, also made some detailed observations about the muscle deformations which occur on a human face during an expression.

Among the more recent works, the work of psychologist Paul Ekman lays the foundation for the modern day understanding of facial expressions in humans. Darwin's claim that facial expressions are universal was based on his theory of evolution but Ekman conducted several cross-cultural studies to establish that facial expressions are indeed similar across cultures [18] [19] [16]. Though he proposed the six basic expressions in humans, he also established that human face shows several other subtle/non-subtle expressions. These expressions are characterized with motions in various regions on the face. Ekman and Friesen [18] also proposed Facial Action Coding System (FACS) which describes various motion patterns on the face in terms of action unit codes. Each

action unit (AU), in their system is associated with one or more muscles which brings about a particular motion on the face (e.g. AU1 is associated with raising of inner eyebrows). The combination of various action unit activations on a face can be mapped to basic human expressions.

Among the computer vision researchers, Automated Facial Expression Recognition has been an area of interest for several decades. A comprehensive survey of some of these techniques may be found in [33]. The problem of facial expression recognition using machine vision presents various challenges. A robust system must be able to handle variability among subjects, variability in illumination and must be capable to handle occlusions and pose variations [35]. Several researchers have addressed some of these challenges individually but an automated system which is robust against all these challenges is yet to be developed, hence facial expression recognition remains an open research problem today.

A relatively new development in facial expression research has been the distinction between posed and spontaneous expressions. Posed expressions are artificial expressions shown by a person when instructed to do so while spontaneous expressions are an outcome of some natural event. It has been shown that spontaneous expressions are usually subtle and differ a lot with posed expressions which are exaggerated [19]. Much of the research has focused on recognizing posed expressions, the results of which may not be transferable to natural expressions. The main reason for this is the lack of datasets which have spontaneous expressions. As noted by Sebe et al. [39] such datasets are

usually difficult to build under controlled settings. The controlled environment and the knowledge that one is been photographed affects the manner in which a subject exhibits an expression. Sebe et al. [39] also propose a solution for this by making people interact with a kiosk which tries to induce emotions and capture their activity by hidden webcams. One of the interesting observation which came out from their study was the fact that under normal conditions, it was very difficult to induce a wide variety of expressions among the subjects as compared to controlled environment. Also, the way people show certain expressions varies with a substantial degree when compared with posed expressions. In this work, we focus on recognizing posed expressions, but it is important to acknowledge that to develop robust human computer interaction systems based on facial expressions, its imperative to address this challenge.

Most of the proposed methods either try to recognize basic expressions or the action units present in the images for the purpose of facial expression recognition. In the earlier works, the emphasis was usually on recognizing them from static images [46] [5] [40]. These static images are generally the peak displays of expressions. Such systems have less practical application in comparison with systems which can classify continuous stream of images. There has been a strong interest in recognizing expressions from image sequences recently. Some systems do this by considering each image of the sequence in isolation while others try model the temporal dynamics of facial expressions by looking at the entire sequence. The psychological experiments carried out in [2] [7] strongly suggest that modeling temporal variations is a crucial factor

in discriminating facial expressions. In the following subsections we provide details of some of the techniques used by researchers for face representation followed by a review of some of the important contributions in recognizing facial expressions from static images or image sequences.

2.1 Face Representation and Facial Features

For every AFER system, a suitable face representation is necessary to extract facial features in order to build classifiers. Some researchers take a holistic approach by using face detectors like Viola Jones [49] and then extract features from the entire face image [5] [40] while others try to localize important fiducial points, also known as facial landmarks on the face [46] [24] [12] [39].

The more successful approach has been to localize facial landmarks in order to obtain precise information about the face. These landmark points can then be used to derive geometric or appearance features. Geometric features encode the shape information of the face by using the localized landmark points or by computing various distances and angles between them [46] [24] [39]. The appearance features can be computed over the entire face or for regions around the landmark points. The appearance features usually represent the texture changes on the face in form of wrinkles or furrows which appear during an expression. Several appearance features like Gabor Filters [5] [27] and Local Binary Pattern features [40] [53] have been used to recognize facial expressions successfully.

Sebe et al. [39] use Piecewise Bezier Volume Deformation (PBVD)

tracker [45] to localize fiducial points. This tracker uses a model based approach and constructs a 3D wire-frame model of the face. The first frame is manually labeled with the locations of the landmark points. A generic face model consisting of 16 surface patches, is then warped to fit these labeled points. Littlewort et al. [27] first use a face detector based on Viola Jones [49] to get a face estimate. They further use similar feature specific detectors (mouth detector, nose detector etc.) along with linear regression to localize 10 landmark points on the face.

Lucey et al. [28] employ Active Appearance Based Models (AAM) [13] to track the face and extract visual features. AAM based models use a training procedure using a set of labeled images to model linear shape and appearance variation. It performs a gradient search at the time of tracking to fit the shape and appearance components on an unseen image. Kanaujia et al. [24] use active shape models with localized Non-negative Matrix Factorization in order to perform the localization of landmark points.

The results of various methods show that precise location of landmark points is important for an AFER system. But, it may not be necessary that all the landmark points contain important information to recognize expressions. The large amount of motion around the mouth and eyebrow region provides much more information for recognizing expressions as compared to other regions. Pards and Bonafonte [36] report very high accuracies for happiness and surprise. Both these expressions have a clear and distinct movement of mouth, hence are relatively easier to recognize. Bourel et al. [9] analyzed the effect of

occlusions on the ability to recognize expressions and their work also reports that occlusion around mouth and eyebrow region causes more performance degradation. Our results also validate these two observations. We constantly find that the system works well for both surprise and happiness but does not perform as good for other expressions.

2.2 Facial expression recognition in static images

Facial expression recognition was previously focused on analyzing static images. Much progress has been made in this field as reported in some earlier works. Pantic and Rothkrantz [34] proposed a fully automatic facial expression recognition system. Their method uses multiple feature detectors (eye, eyebrows, nose, mouth) to detect 19 fiducial points in both frontal and profile face views. From these 19 points they extract 30 facial features (various distances and angles). They then employ a rule based approach using the FACS manual [18] to recognize action units. They further map these action units to basic expressions as well.

The first work to show the usefulness of Gabor filter response for facial expression recognition was done by Lyons et al. [29]. They manually labeled 34 points on the face and then compute Gabor filter responses at each of these points to get a feature vector. They further employ principal component analysis to reduce the dimensionality of these vectors and use LDA to build binary classifiers for each expression.

Hong et al. [21] make use of elastic graph matching [52] to fit a labeled

graph which they call as General Face Knowledge (GFK) to localize landmark points on the face. They use filter responses of Gabor wavelets [29] to define jets which is an array of these responses at every point. To perform the facial expression recognition they fit a labeled graph to the test image and then match it with all the images in the dataset using elastic graph matching again. Their dataset contains only 262 images, still the average recognition time was about 8 seconds which makes this approach unsuitable for large datasets.

Huang and Huang [22] use a point distribution model (PDM) for facial expression analysis. They generate this PDM by labeling 90 points manually on the face. They further generate mouth contours using three parabolic curves and sample points from the contour to obtain the final model. They compute 10 features termed as action parameters from the model. This is done by computing the difference of the point locations in the expression image with the neutral image. After applying principal component analysis they use a 3-nearest neighbor classification to recognize expressions.

A relatively recent work by Bartlett et al. [6] addresses the challenge of recognizing spontaneous facial expressions. Their proposed system is fully automatic and works in real-time. They take a holistic approach by using the entire face image from a face detector. This face image is first rescaled and then passed through a bank of Gabor filters containing 8 orientations and 9 frequencies. They train support vector machines based classifiers after performing a feature selection using AdaBoost for recognizing 20 action units on the face.

Motivated by the success of Local Binary Pattern (LBP) features for face recognition [1], Shan et al. [40] proposed an approach for recognizing facial expressions using boosted LBP features. They report their results using Cohn-Kanade dataset [23] which consist of image sequences but they take only last 3 images into consideration. They first compare the performance of LBP features vs. Gabor features using template matching and support vector machines and show that LBP performs slightly better than Gabor features. They further use AdaBoost with LBP features and train support vector machines with boosted features. The approach with boosted features shows a further improvement in the performance. For the purpose of comparison between shape and appearance features we follow this work closely to perform facial expression recognition using appearance features.

2.3 Facial expression recognition in image sequences

In this section we review work that is closely related with recognizing facial expressions in image sequences. Essa and Pentland [20] proposed simple motion energy detectors for recognizing facial expression using motion. They use the view-based and modular eigenspace method of [37] to localize position of eyes, nose and mouth. Using the optical flow computation proposed by [41] they estimate 2D spatio-temporal motion energy between two consecutive frames. Using this motion energy representation they generate spatio-temporal templates for six basic expressions. The difference between the stored templates and the motion energy observed in the test image is used

for classification.

Black and Yacoob [8] use local parameterized models of image motion for recognizing facial expressions. Their method is capable of recognizing expressions even in presence of significant head motion. They use planar-affine model for the motion of eyes and planar-curvature model for the motion of eyebrows and mouth. These parameters are estimated using robust regression methods. In their work, they obtain some mid and high level descriptions of facial expressions. For e.g. rightward movement of mouth, curving of eyebrows etc. can be seen as some of these descriptions which are dependent on several parameters. They formulate rules for recognizing facial expressions in terms of these parameters and based on several observations.

Cohen et al. [11] [12] use a tree-augmented-naive Bayes classifier (TAN) for continuous videos to learn the correlation between motions of different facial regions and expressions. They use Piecewise Bezier Volume Deformation (PBVD) tracker [45] to localize and track landmark points. The continuous tracking of landmark points enables them to estimate the amount of motion taking place in various regions of the face. They use 12 different motion magnitudes as an input for their classifiers. They perform comparisons for their proposed approach using TAN with Naive-Bayes algorithm. Their approach outperforms Naive-Bayes considerably.

The authors in [54] propose a method using moment invariants as features along with Hidden Markov Models (HMMs) to analyze facial expressions. Their system is capable of recognizing 4 expressions: anger, disgust, happiness

and surprise. The features are normalized by subtracting the features for every frame with that of the first frame. They use 3 state discrete left-right HMM for the purpose of recognition. A very high overall accuracy is reported on their own dataset. Its difficult to see how well their method will perform in presence of more expressions and other datasets.

Littlewort et al. [27] recently released their Computer Expression Recognition Toolbox (CERT). Their system is capable of recognizing 19 different facial action units and 6 basic facial expressions. This can be done in continuous video streams but they consider each image in isolation and ignore temporal dynamics between frames. They use Gabor filters at 8 orientations and 9 spatial frequencies as features for training a support vector machines. The interesting thing about their work is that they train their detector using images from multiple datasets and report an overall accuracy of 80% on CK+ [28] dataset. We also perform our experiments on the CK+ [28] dataset but lack of detailed experimental details in their paper prevents us from doing a direct comparison.

Dhall et al. [15] use both shape and appearance features in form of pyramid histogram of gradients (PHOG) and local phase quantization (LPQ) features to recognize facial expressions in image sequences. They employ constraint local model (CLM) tracking to localize faces in the image. After computing the relevant features, they perform principal component analysis and then train a simple support vector machine classifier for recognition. The experiments were performed on GEMEP-FERA [47] dataset which is very new

and challenging.

In [53] a volume based appearance descriptor as an extension to usual Local Binary Pattern (LBP) is proposed to recognize facial expressions in image sequences. The proposed descriptor combines motion and appearance in single descriptor. The results are reported on Cohn-Kanade dataset [23] by training support vector machine classifier. The overall accuracy is reported to be 97.3% which is shown to be much better than many previous results. The main issue with their approach is that the authors consider a given image sequence as a whole and classify the entire sequence into one of the expression classes. For a practical application, a facial expression system must be able to classify images as they come, therefore a solution which can model the transition between various expressions and label each image continuously is more desirable.

Conditional Random Fields (CRFs), which provide one such solution, were introduced in [25]. These are discriminative models which define a conditional probability p(Y|X) over label sequences Y given a particular observation sequence X. The primary advantage of CRFs over generative models like HMMs [38] comes from the fact that models like HMMs try to define a joint probability distribution p(X,Y) over observation sequences X and their label sequences Y [44]. To make the model computationally feasible, strong independence assumptions among both observations and labels are required. In case of CRFs, the independence assumption has to be made for labels and not for observations. This gives them the ability to model the dependencies

between features from several frames in the sequence. Hence CRFs prove to be more robust in comparison [25].

Sminchisescu et al. [42] have shown the effectiveness of CRFs in recognizing several human motions like walking, running etc. Their method outperforms HMMs and even provides good results for differentiating between subtle motion patterns like normal walk vs. wander walk. The authors in [24] use CRFs to classify facial expressions from image sequences. They use localized active shape models for tracking face shapes across the image sequences. They also perform feature normalization by assuming the first frame to be neutral and subtracting neutral shape from all other shapes. Their work aims at designing a complete facial expression recognition system but does not provide a detailed analysis on the importance of using temporal information for performing this task. In our work, we show that the dynamics of shape contain much richer information to recognize expressions in comparison with analyzing each shape individually. We also use CRFs as one of the underlying discriminative classifiers to compare the performance of our proposed approach.

The variants of CRFs which include hidden states have been successfully applied for gesture recognition. It has been shown that these approaches are good at capturing subtle motion patterns using hidden states. One such approach known as Hidden Conditional Random Fields (HCRFs) [51] is commonly used to assign a single class label to the entire sequence. The main disadvantage of this method is the need of manual segmentation of a continuous sequence before it can be classified. This itself is a challenging task and may

need special algorithms. Another approach which can automatically segment the video sequences and assign every frame with the appropriate class label is Latent Dynamic Conditional Random Fields (LDCRFs) [32]. Morency et al. [32] demonstrated that modeling sub-structure motion for a gesture class using hidden states, helps in distinguishing between different gestures more robustly. They experimentally show that their proposed latent dynamic conditional random fields approach outperforms SVMs, HMMs and traditional CRFs easily for gesture recognition. Their work specifically deals with recognizing full body gestures. In our work we use the LDCRFs along with other techniques like procrustes analysis and PCA to recognize expressions. In our knowledge, there has been no prior work which reflects the usefulness of using LDCRFs for this task. Our proposed approach using LDCRFs is also more robust in modeling facial expressions as compared to CRFs which shows that capturing subtle facial motion is very essential in differentiating between facial expressions.

Chapter 3

Methods

In this chapter we explain the features and classification methods used in our work. Section 3.1 & Section 3.2 gives an overview of the shape and appearance features which were used while Section 3.3 and Section 3.4 discuss the classification methods which are used for static and dynamic facial expression recognition respectively.

3.1 Shape Features

A 2D face shape for our work is represented by a set of 68 landmark points which are basically located around the contours of eyebrows, eyes, nose, chin, inner lips and outer lips. The distribution of these landmark points on the face can be seen in Figure 3.1.

Mathematically, we can represent a 2D n-landmark face shape with a 2n size vector as shown in the following equation:

$$\mathbf{x} = [x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n]^T \tag{3.1}$$

In order to perform a robust shape analysis for different expression shapes, it's important to obtain their true shapes by removing the effects of

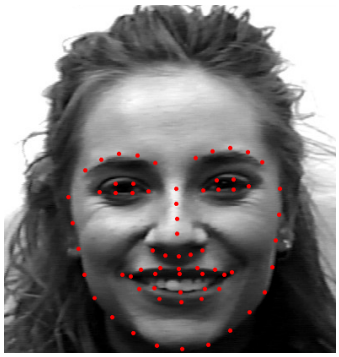


Figure 3.1: Distribution of 68 landmark points on the face. The figure shows that locations of various fiducial points used for our work on the human face. These points are located around the contours of eyebrows, eyes, nose, chin, inner lips and outer lips.

rigid geometric transformations such as translation, scale and rotation between them.

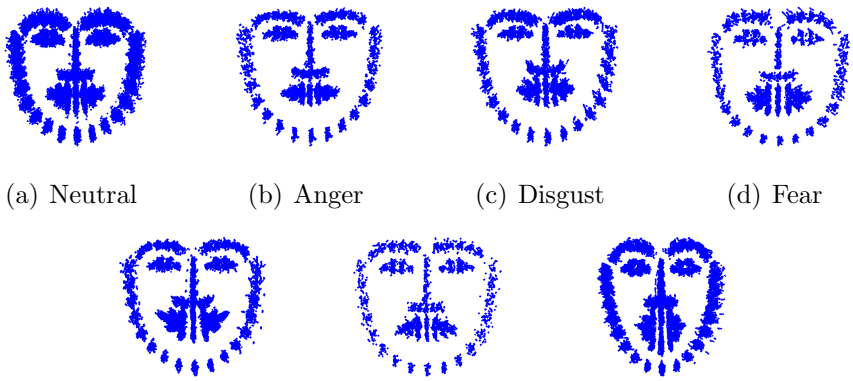
We use Generalized Procrustes Analysis for this task [43], which tries to minimize the sum-of-squared distances between landmark points of all the shapes w.r.t. the rigid transformations. Following are the steps for Generalized Procrustes Analysis algorithm for N shapes:

1. Choose an initial estimate for the mean shape (for e.g. first shape).

2. Apply a similarity transform on all the shapes to align them with the mean shape.

3. Recompute the mean shape estimate by averaging the aligned shapes from step 2 as follows:

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x_i} \tag{3.2}$$



(a) Neutral (b) Anger (c) Disgust (d) Fear

(e) Happiness (f) Sadness (g) Surprise

Figure 3.2: Shapes of various Facial expressions after Generalized Procrustes Analysis. The 5643 face shapes from the dataset grouped by expression labels and aligned using Generalized Procrustes Analysis can be seen here. This figure demonstrates that shapes for some of the expressions like surprise and happiness are easily distinguishable while for other expressions there is a considerable similarity in the face shape.

4. Repeat step 2-3 until the change in the mean shape becomes negligible.

Figure 3.2 shows the aligned face shapes after Generalized Procrustes Analysis is applied to all the shapes in the dataset. It can be observed that some expressions such as surprise have very distinct shapes while others such as anger and sadness show a certain degree of similarity. After performing the alignment, the true shape which remains gives us a 136 dimensional feature vector.

We apply Principal Component Analysis (PCA) to reduce the dimensionality to 18 by retaining 95% of the variance. It was observed that facial expressions are typically characterized by motion in various parts of the face

especially mouth region, eyes and eyebrows. The regions around chin portion do not contribute a lot in generating different expressions. This observation was validated by PCA analysis as the dominant variations were found around the mouth, eyebrows and eye regions as compared to other locations on the face. Figure 3.3 shows the first 3 principal components for the six basic expressions. It can be seen that some expressions like anger, surprise and fear are easily separable in the PCA space while sadness shows a lot of overlap with other expressions. This is understandable because sadness is generally exhibited with a very small movement on the face and thus does not bring very visible changes in the face shape. The expressions disgust and happiness look to have some overlap in the PCA space but our results show that its easy to recognize happiness in comparison with other expressions.

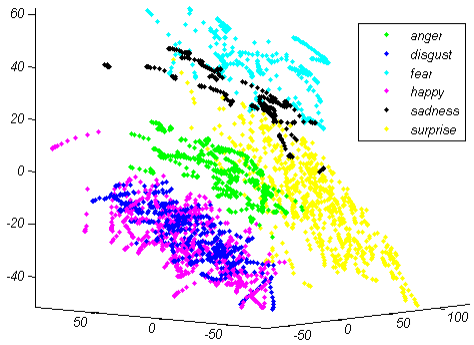


Figure 3.3: First 3 Principal Components (Shape Features) for six basic expressions. The figure shows that expressions like anger, surprise and fear are easily separable from other expressions while sadness shows a significant overlap with other expressions. Disgust and happiness look to have some overlap but our results show that happiness is easier to recognize as compared to other expressions.

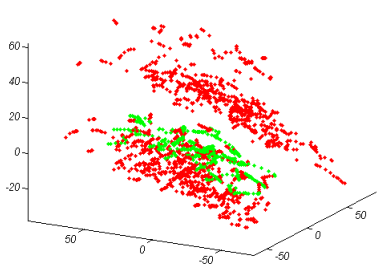
For practical applications, its important to consider even the neutral state while designing classifiers for facial expressions. Introducing the neutral state makes the task of recognizing facial expressions more difficult because of two reasons:

1. Many subtle expressions like anger and sadness do not cause a lot of movement on face, therefore they are difficult to differentiate with neutral state.

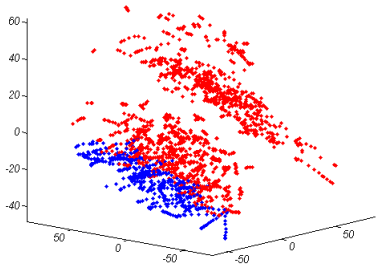
2. It is very difficult to define a clear distinction between the end of neutral state and the onset of an expression even for humans. It makes the task of ground-truth labeling very challenging.

These issues are clearly highlighted in Figure 3.4. For all the expressions, neutral shows some overlap with the actual expressions. These shapes mostly correspond to the transition phase where its difficult to tell if a shape belongs to the neutral state or to the actual expression. The plots corresponding to the anger and sadness show a lot of overlap with neutral shapes in the PCA space which makes it difficult to recognize these expressions in presence of neutral shapes.

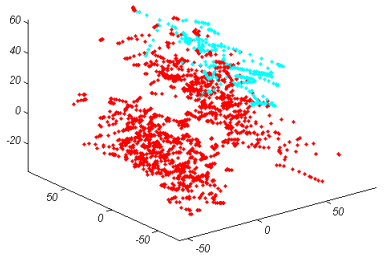
It is important to note that Figure 3.3 and Figure 3.4 show just the first 3 principal components for the purpose of analysis. To actually build the classifiers first 18 principal components were used based on the amount of variance they can retain. The other principal components also contribute towards reducing the overlap that is observed in these figures.



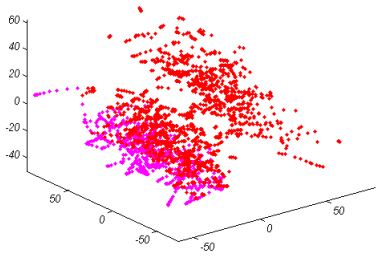
(a) Neutral(Red) vs. Anger(Green)



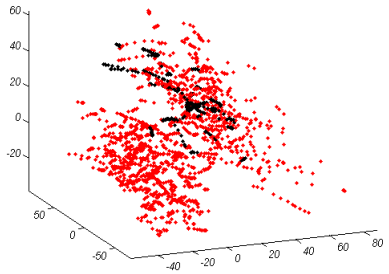
(b) Neutral(Red) vs. Disgust(Blue)



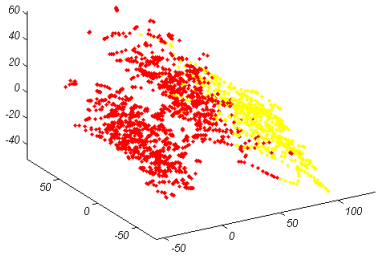
(c) Neutral(Red) vs. Fear(Cyan)



(d) Neutral(Red) vs. Happy(Magenta)



(e) Neutral(Red) vs. Sadness(Black)



(f) Neutral(Red) vs. Surprise(Yellow)

Figure 3.4: Comparison between first 3 Principal Components (Shape Features) for neutral images with other expressions. This figure demonstrates that some neutral images are very close to the expression images in the PCA space. Those neutral images mostly correspond to the transition phase from the neutral to the actual expression. Especially for anger and sadness there is a significant overlap which makes these expressions difficult to recognize in presence of neutral images.

3.2 Appearance Features

One of the aims of this work is to experimentally show the importance of temporal variations in shape as compared to the temporal variations in appearance for facial expression recognition. Several appearance features have been successfully applied for recognizing static facial expressions.

We use histogram based Uniform Local Binary Pattern (U-LBP) [40] features which are commonly used for facial expression recognition to conduct our experiments. In this method the LBP operator is applied on a pixel by thresholding its circular neighborhood with the intensity value of the pixel and representing it in binary form (1 if the intensity value of the neighboring pixel is greater than the current pixel, 0 otherwise). The patterns which contain at most two bitwise transitions from 0 to 1 or vice versa are called uniform local binary patterns. It was observed that uniform patterns form the majority of the observed patterns [40]. Hence, to construct the histogram, all unique uniform patterns are binned separately while all non-uniform patterns are assigned to a single bin. We use a 8 pixel neighborhood which gives us a 59 bin histogram.

It has been shown[40] that using a single histogram for the entire image is not a good technique for facial expression recognition, hence the cropped face image is subdivided into 42 regions using a 6 x 7 grid (see Figure 3.5). Then a separate histogram is computed for each sub-region which gives us a feature vector of length 2478. Principal Component Analysis (PCA) is applied to reduce the dimensionality to 59 by retaining 95% of the variance.

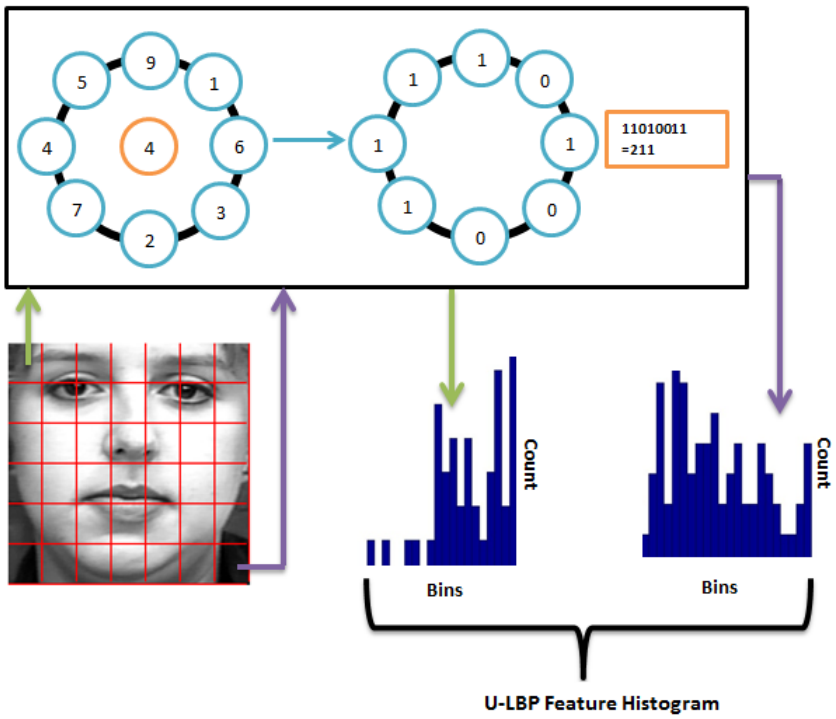


Figure 3.5: Compuation of Uniform Local Binary Pattern (U-LPB) Histogram. The face image is divided into 6 x 7 grid and then a separate U-LBP histogram is computed within each grid by applying the ULBP operator on every pixel in the grid as shown.

3.3 Static Facial Expression Recognition

Static Facial Expression Recognition is usually referred as the task of classifying static images which are not a part of any video or an image sequence and have to be recognized individually. Traditionally researchers have been trying to develop methods which work on single expression images [34] [29] [21] [6] . Most of these approaches have been shown to be successful for the

cases where the static image shows the peak expression. Our goal is not to obtain the highest accuracy for the task of classifying in such cases, we are rather interested in classifying images from continuous image sequences which contain the full extent of an expression from onset to offset.

Hence for comparison with our proposed method, we define Static Facial Expression Recognition as the task of labeling all the images in a given image sequence individually without considering the dynamics between them. Therefore, for the purpose of training we will consider each image in isolation, even though they are a part of an image sequence. This section discusses the two techniques used in our work for Static Facial Expression Recognition:

3.3.1 Facial Expression Templates

The manner in which facial expressions are exhibited on a human face vary from person to person. These differences can be attributed to the person's physical characteristics or the emotional state at the time of the expression. It forms a very interesting proposition to see if it's possible to generate templates representing different variations which the same expression can exhibit (for e.g. different kind of smiles, which may vary in intensity). These templates can then be used to classify facial expressions.

As mentioned in the previous section, after applying Procrustes Analysis we are left with the true shapes of the faces. To generate facial expression templates we cluster these true shapes using K-means clustering for every expression class individually. The motivation behind this approach is that similar

face shapes within an expression class will occupy same clusters. The cluster centers will then represent distinct templates for that expression class. Figure 3.6 shows an example of the kind of templates we obtain using this technique. It can be seen that these templates are able to capture the variations within an expression class effectively (for e.g. the mouth shape variations can be observed within surprise and sadness templates).

To classify unseen face shapes we use K-Nearest Neighbor classification (KNN) approach by computing the sum-squared distances between the test shape and expression templates. It was found that using 8 templates per expression and K=3 for KNN gives the best classification accuracy.

We compare the performance for static classification of facial expressions obtained using the template based approach with the performance of CRFs and LDCRFs based classifiers which are dynamic in nature.

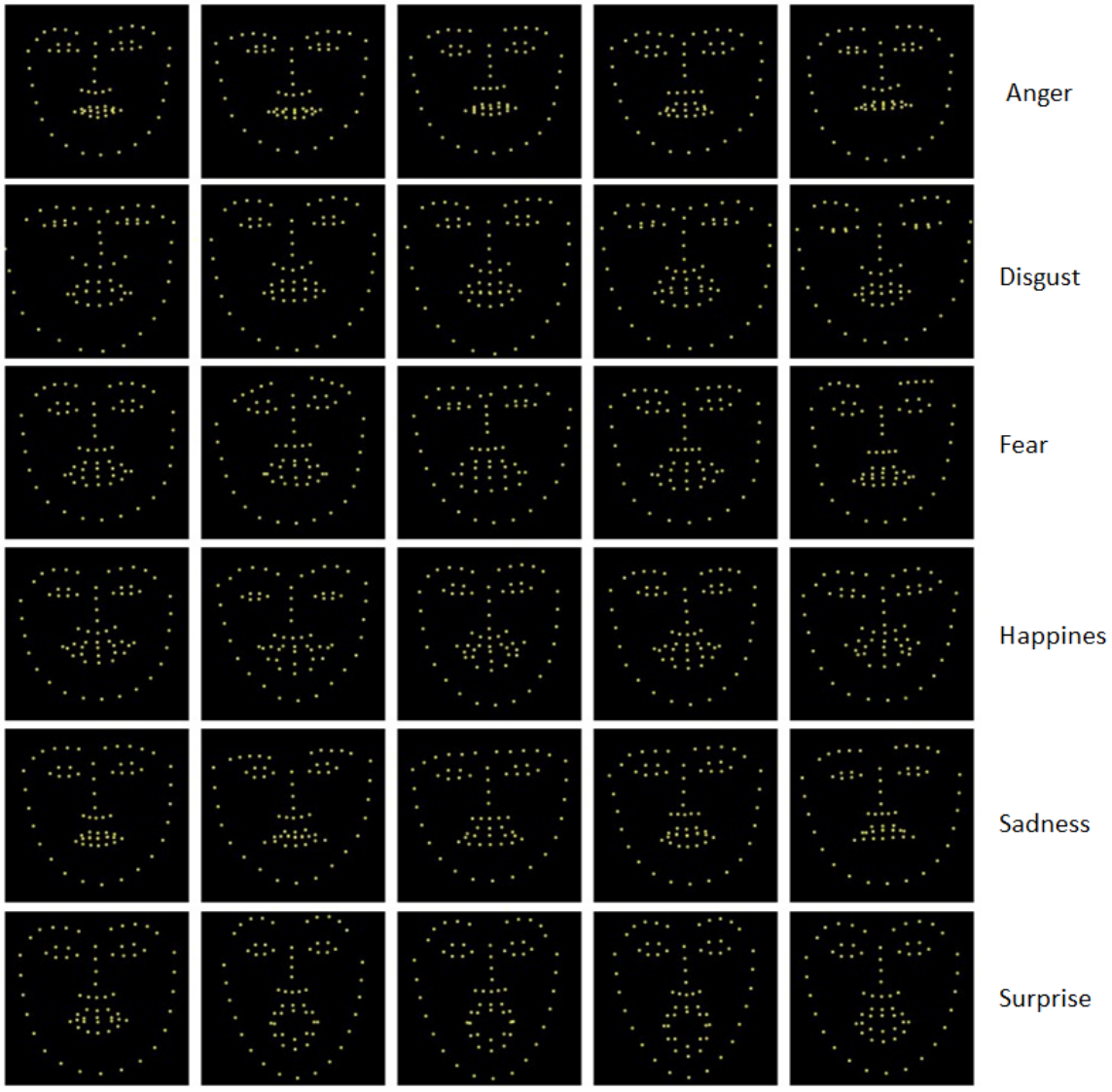


Figure 3.6: Facial Expression Templates. This figure shows various templates for facial expressions obtained using K-means Clustering. Different variations of the same expressions are effectively represented by different templates. The small changes in eyebrows, nose, mouth shapes for the same expression can be seen.

3.3.2 Support Vector Machines (SVMs)

SVMs have been applied successfully for many computer vision problems including facial expression recognition [3] [4] [48] [40]. SVMs try to find a separating hyperplane with maximal margin after projecting the data into a higher dimensional feature space where its more easy to separate.

SVMs make binary decisions, so for our work we use the shape features derived using Generalized Procrustes Analysis and PCA to train 6-class (without neutral) and 7-class (with neutral) one-against-all SVMs. We assign the largest output from each binary classification as the probable class label. The Radial Basis Function (RBF) kernel along with a grid search using 10-fold cross validations was used to find the best values for C (penalty term) and γ (RBF kernel parameter) while training the SVMs.

In this work the results obtained using SVM classifiers which amounts to static classification of facial expressions will be compared with the performance of CRFs and LDCRFs based classifiers which are dynamic in nature.

3.4 Dynamic Facial Expression Recognition

The temporal dynamics between features has proved to be an important consideration while trying to design classifiers for supervised sequence labeling problems. Classification by modeling temporal dynamics between features has been applied successfully for many problems in speech recognition, natural language processing and gesture recognition. Some of the psychological exper-

iments [2] [7] and encouraging results by some of the researchers [27] [53] [24] [11] [12] reflects the importance of modeling temporal motion patterns for the task of facial expression recognition. This section discusses two approaches which were used in this work for modeling temporal variations between shape and appearance features for the purpose of facial expression recognition.

3.4.1 Conditional Random Fields (CRFs)

CRFs provide a highly discriminative and probabilistic method [25] to model the variation of shapes in time. For comparison with our proposed approach, we use the basic linear chain CRFs for the task of facial expression recognition. It can be viewed as an undirected graphical model in which the nodes represent the class labels and feature observations while the edges represent transition probabilities between them (Figure 3.7). CRFs are conditioned on observations i.e. the independence assumptions are made for class labels and not for observations. Thus they are able to model complex features for observations while remaining computationally tractable.

Mathematical Formulation:

The problem of supervised sequence labeling requires us to learn a classifier from training data consisting of a set of labeled sequences. For notational simplification, we refer the observation sequences (X_1,X_2,X_3,.....,X_T) as X and label sequences (Y_1,Y_2,Y_3,.....,Y_T) as Y for T frames to be labeled. Here each X_i, i $\in (1, 2...T)$ is a random variable representing either the shape or the ap-

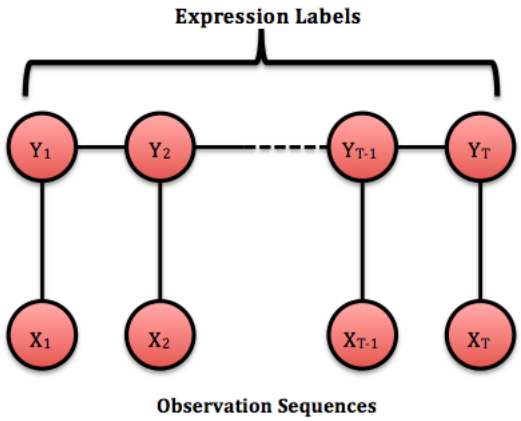


Figure 3.7: Linear Chain Conditional Random Fields. Here each X_i is the observation vector for each frame and each Y_i is the class label for each frame in the sequence. A linear chain structure means that a first-order Markov assumption is made for class labels.

pearance features and each Y_i, i $\in (1, 2...T)$ is a random variable representing the expression label or neutral state.

A CRF model for T image frames is formulated as follows:

$$P(Y|X;\theta) = \frac{1}{Z(X,\theta)} exp\left(\sum_j \theta_j F_j(Y,X)\right) \tag{3.3}$$

where,

$$F_j(Y,X) = \sum_{t=1}^{T} f_j(Y_{t-1}, Y_t, X, t) \tag{3.4}$$

$$Z(X,\theta) = \sum_Y exp\left(\sum_j \theta_j F_j(Y,X)\right) \tag{3.5}$$

Here, $Z(X,\theta)$ is the normalization factor and each $f_j(Y_{t-1}, Y_t, X, t)$ is either a state function $st_j(Y_t, X, t)$ which evaluates the interaction between features or

a transition function $tr_j(Y_{t-1}, Y_t, X, t)$ which models the temporal dependencies among features [50].

Training (Parameter Estimation):

Given a set of N labeled training samples the objective of the training procedure is to estimate the set of weights θ^* which maximizes the conditional log likelihood (i.e. $\theta^* = \text{argmax}_\theta \text{ L}(\theta)$) by optimizing the conditional log likelihood function given by equation (3.6).

$$L(\theta) = \sum_{k=1}^{N} \left[\sum_j \theta_j F_j(Y^{(k)}, X^{(k)}) - log \frac{1}{Z(X^{(k)}, \theta)} \right] \tag{3.6}$$

For our work, we use Broyden Fletcher Goldfarb Shanno (BFGS) [30] gradient ascent technique for optimizing the log likelihood function. The optimization procedure also involves a regularization term which is decided using cross validation with values ranging from 10^{-3} to 10^3 during training. The training procedure converges in less than 100 iterations.

Inference:

To classify an unseen test sequence X_{test} we want to find the most likely labels Y^* for the sequence. For the purpose of inference we can ignore the denominator as well as the exponential in equation 3.3. Using the learned parameters θ^* from the training data we can simply compute:

$$Y^* = \underset{Y}{\text{argmax}}\, P(Y|X_{test}; \theta^*) = \underset{Y}{\text{argmax}} \sum_j \theta_j^* F_j(Y, X_{test}) \tag{3.7}$$

The model outputs the marginal probabilities for each class label. The class label with the highest probability for each frame is used as the predicted label for that frame.

3.4.2 Latent-Dynamic Conditional Random Fields (LDCRFs)

CRFs provide a strong discriminative framework to model the transitions between facial expressions. They consider the features for a given class in isolation and learn the dynamics of those features from one class to another. Since the structure within an expression class is considered holistically, they fail to model the subtle facial motions within an expression which may be important to differentiate between visually similar facial expressions.

In [32] a variant of traditional CRFs known as Latent-Dynamic Conditional Random Fields (LDCRFs) was proposed which captures the subtle motion patterns within a class along with inter-class motion patterns by associating a set of hidden states with each class label. These hidden states can model the internal sub-structure for different facial expressions and contribute in the overall likelihood for recognition. Each hidden state can be treated in a similar manner as a CRF and the overall likelihood can simply be the sum of individual likelihoods from the hidden states.

Mathematical Formulation:

The LDCRF model uses an additional set of hidden variables H = (H_1,H_2,H_3,.....,H_T) apart from X and Y for every sequence (Figure 3.8). The

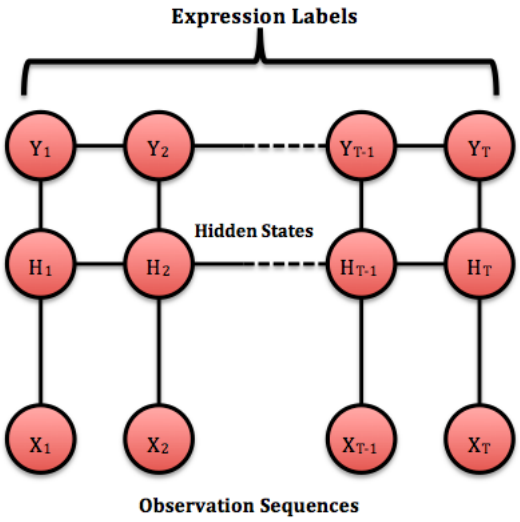


Figure 3.8: Latent Dynamic Conditional Random Fields. Here each X_i is the observation vector for each frame, each H_i is the hidden state associated with every X_i while Y_i is the class label for each frame in the sequence. The variables associated with hidden states are not directly observed and have to be estimated from the training data.

model can then be defined over parameters θ as:

$$P(Y|X;\theta) = \sum_{H} P(Y|H,\theta)P(H|X,\theta) \tag{3.8}$$

The LDCRF model imposes a restriction that sets of hidden states for each class label needs to be disjoint. This implies that for a given class label Y_j the set of possible hidden states H_j is constrained to a subset H_{Yj} of all possible hidden states. This assumption gives the following deterministic relationship between Y and H:

$$P(Y|H,\theta) = \begin{cases} 1 & \forall\ H_j \in H_{Yj} \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

Hence equation (3.8) can be refined as:

$$P(Y|X;\theta) = \sum_{H:\forall H_j \in H_{Yj}} P(H|X,\theta) \tag{3.10}$$

P(H|X,θ) is then defined exactly as P(Y|X,θ) is defined in the previous section.

$$P(H|X;\theta) = \frac{1}{Z(X,\theta)} exp\left(\sum_j \theta_j F_j(H,X)\right) \tag{3.11}$$

where,

$$F_j(H,X) = \sum_{t=1}^{T} f_j(H_{t-1}, H_t, X, t) \tag{3.12}$$

$$Z(X,\theta) = \sum_{Y} exp\left(\sum_j \theta_j F_j(H,X)\right) \tag{3.13}$$

As in CRFs, $Z(X,\theta)$ is the normalization factor and each $f_j(H_{t-1}, H_t, X, t)$ is either a state function $st_j(H_t, X, t)$ or a transition function $tr_j(H_{t-1}, H_t, X, t)$ [50]. The parameter estimation and inference can then be performed in a similar manner as in CRFs. For our work, we use Broyden Fletcher Goldfarb Shanno (BFGS) [30] gradient ascent technique for optimizing the log likelihood function.

The optimal number of hidden states and the regularization term were found using cross-validation during training. It was observed that 5 hidden states give the best results. The training procedure converges in less than 100 iterations.

Chapter 4

Experiments & Results

This chapter gives details about the dataset used for experiments, followed by an overview of the experiments that were conducted. We then present the results of various experiments and compare the facial expression recognition performance for all the techniques from various aspects.

4.1 Overview of the dataset

The experiments for our work were conducted on the Extended Cohn-Kanade Dataset (CK+) [28] which contains 593 sequences from 123 subjects. These are not fixed length sequences and the duration varies from 10 to 60 frames (Figure 4.1 shows some example sequences). All the sequences start from the neutral pose to the peak formation of the expression. The locations of facial landmarks are provided along with the dataset. Out of the 593 sequences in the dataset only 309 were labeled as one of the six basic expressions (see [28] for details). Table 4.1 gives the detailed statistics for the portion of the dataset that was used. The expression onset for all sequences takes place after certain number of neutral frames; hence we manually label each frame in a sequence to be either neutral or belonging to the expression class (Figure 4.2).

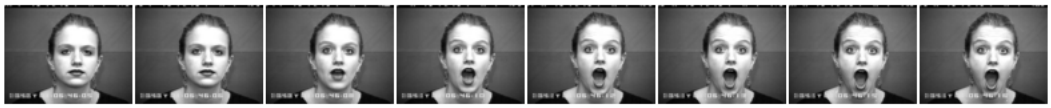
(a) Anger

(b) Disgust

(c) Fear

(d) Happiness

(e) Sadness



(f) Surprise

Figure 4.1: Example from the dataset of a person exhibiting various facial expressions. The neutral, onset and apex phases for facial expressions can be observed in all the example sequences.

Expression	No. of Sequences	Total No. of Images
Anger	45	1022
Disgust	59	868
Fear	25	546
Happiness	69	1331
Sadness	28	547
Surprise	83	1329
Total	309	5643

Table 4.1: Overview of the dataset.

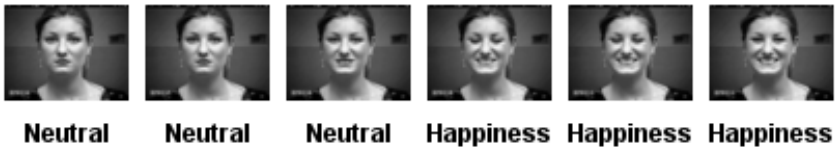


Figure 4.2: Example of a labeled sequence. All the image sequences were manually labeled as either belonging to neutral state or the given expression class.

4.2 Experiment Details

We perform experiments to show that modeling temporal variation between shapes helps in recognizing those facial expressions which are otherwise difficult to recognize using classifiers which do not model temporal dependencies. For this, we compare the recognition performance of Expression Templates based technique and Support Vector Machines (SVMs) classifier against the performance of CRFs and our proposed method of using LDCRFs for recognizing facial expressions. We also show by experiments that modeling

variation in shape across time is much more important than modeling variation in appearance across time for recognizing facial expressions. To show this, we train both CRFs and LDCRFs classifiers using appearance features and compare the performance with shape features.

All the experiments were conducted using 4-fold cross-validation and the results were averaged over all the folds. We evaluate the recognition performance in all the experiments for two cases (6-class vs. 7-class). First we consider only those frames in the dataset which belong to one of the six expression classes (Anger, Disgust, Fear, Happiness, Sadness, Surprise) to see how well we can discriminate between the expression classes. It's important to note that although we are removing neutral frames from consideration here, the remaining sequence still contain the dynamics of an expression from onset to the apex phase. For the other case, we consider all the frames including the neutral ones.

4.3 Static Shape Analysis vs. Dynamic Shape Analysis

In this section we present a comparison between the performance of static facial expression recognition and dynamic facial expression recognition for image sequences using shape features. We empirically show that for both 6-class and 7-class classification, dynamic recognition performs better then static recognition especially for those expressions which have very subtle shape variations.

4.3.1 Classification Results without Neutral State (6-class)

This section discusses the performance of various classifiers trained using images belonging only to one of the expression classes. The results in Table 4.2 for 6-class classification show that happiness and surprise are two expressions which are much easier to recognize as compared to other expressions. The recognition performance for both static shape analysis and dynamic shape analysis is high for these two expressions. It is an intuitive result as these expressions bring a large amount of change in the shape of the face especially the mouth region and thus are relatively easy to recognize. The precision and recall statistics for Happiness (Table 4.3 & Table 4.4) are both high. The precision for surprise is slightly low because other expressions like fear are sometimes confused with surprise. The fear expression is usually exhibited by tightening of lips along with raised eyebrows. But sometimes, it can cause the mouth to open widely as well, which is usually a characteristic of surprise expression causing some confusion. This also results in low recall rates for fear.

	An	Di	Fe	Ha	Sa	Su	Avg
Kanade[28]	75.00	94.70	65.20	**100.00**	68.00	96.00	83.15
Templates	91.50	85.46	89.32	90.24	78.98	90.62	87.69
SVM	74.70	87.13	88.77	98.43	63.70	91.67	84.06
CRF	96.41	97.60	**92.51**	99.41	83.83	97.86	94.60
LDCRF	**97.91**	**97.86**	90.52	99.55	**90.08**	**98.87**	**95.79**

Table 4.2: Recognition Rates without Neutral expression using Shape Features (6-class classification) [Anger(An), Disgust(Dn), Fear(Fe), Happiness (Ha), Sadness(Sa), Surprise(Su), Average(Avg)]

	An	Di	Fe	Ha	Sa	Su	Prec.	Rec.
An	**96.4**	1.8	0.0	0.0	0.0	1.8	0.99	0.97
Di	0.0	**97.6**	0.0	2.4	0.0	0.0	0.97	0.97
Fe	0.0	0.0	**92.5**	0.0	1.0	6.5	0.90	0.91
Ha	0.0	0.6	0.0	**99.4**	0.0	0.0	0.98	0.99
Sa	1.3	0.0	9.2	2.1	**83.8**	3.6	0.96	0.84
Su	0.0	0.0	1.3	0.0	0.9	**97.9**	0.93	0.98

Table 4.3: Confusion Matrix (percentage) and Precision-Recall Statistics for 6-class classification using CRFs. [Anger(An), Disgust(Dn), Fear(Fe), Happiness (Ha), Sadness(Sa), Surprise(Su)]

	An	Di	Fe	Ha	Sa	Su	Prec.	Rec.
An	**97.9**	1.8	0.0	0.3	0.0	0.0	0.98	0.98
Di	0.0	**97.9**	0.0	2.1	0.0	0.0	0.97	0.98
Fe	0.0	0.0	**90.5**	0.0	0.4	9.1	0.97	0.90
Ha	0.0	0.4	0.0	**99.6**	0.0	0.0	0.97	0.99
Sa	2.8	0.0	2.2	1.5	**90.1**	3.3	0.98	0.90
Su	0.0	0.0	0.3	0.0	0.9	**98.9**	0.92	0.99

Table 4.4: Confusion Matrix (percentage) and Precision-Recall Statistics for 6-class classification using proposed method based on LDCRFs. [Anger(An), Disgust(Dn), Fear(Fe), Happiness (Ha), Sadness(Sa), Surprise(Su)]

For other expressions such as anger, disgust and sadness which do not cause a lot of deformation on the face, the temporal shape modeling performs much better than static shape analysis. The performance for SVMs and Expression Templates based method is very low for sadness. The sadness expression causes very little deformation on the face and hence is very difficult to recognize by looking at shape in isolation. The temporal modeling using CRFs improves the performance. But, there is a lot of overlap in the motion patterns for sadness and other expressions hence just learning

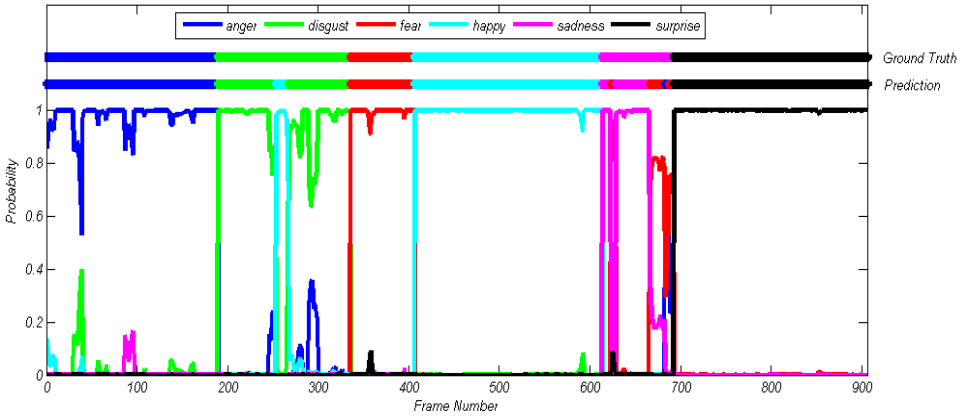
the transitions from one expression to another is not sufficient. The proposed approach using LDCRF successfully models these overlapping patterns using hidden states and captures the subtle differences which improves the accuracy significantly. The confusion matrices in Table 4.3 & Table 4.4 clearly show that other expressions are usually not confused with sadness, giving it a high precision value.

For Anger & Disgust both CRF and LDCRF based classification outperform the static approaches significantly. The precision and recall is high for these expressions which shows that in absence of neutral images, these two expressions can be recognized robustly using temporal modeling. Our approach also gives equivalent results for Happiness and Surprise with the ones reported by Kanade et al. [28] and performs significantly better for other expressions. In their work Kanade et al. [28] use the same subset of data that was used for our experiments for expression recognition. But their experiments use only the last frame of the sequence for classification training and testing. This simplifies the problem further since the last frame exhibits the peak of the expression. Hence, our method which even though tries to label every sequence in the frame shows significant performance increments for the difficult expressions in comparison.

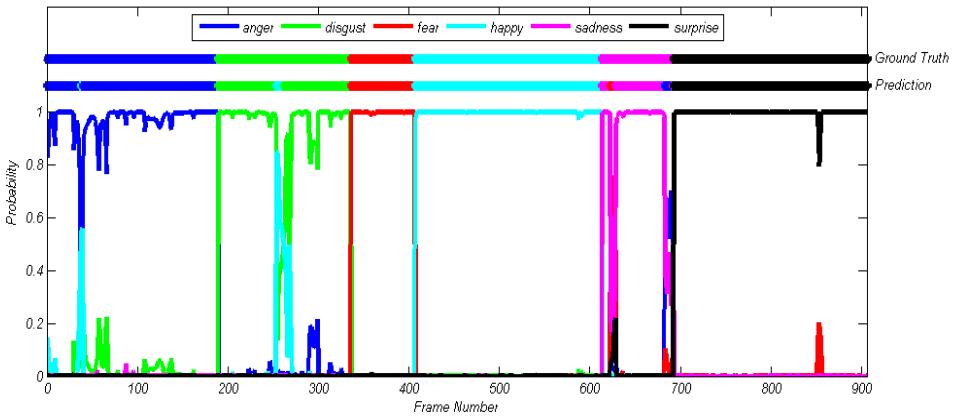
Figure 4.3 shows the predicted probabilities for a random subset of sequences from the dataset. It can be seen that expressions with large movement i.e. Happiness and Surprise are recognized correctly with near certainty by both CRF and LDCRF. It can also be seen that for some cases where CRF

model fails to recognize expressions correctly (e.g. for sadness) the LDCRF model predicts the correct class label.

Overall, the dynamic approaches for facial expression recognition perform significantly better than static approaches for 6-class classification. The proposed approach using LDCRFs gives better accuracy than the traditional CRFs based approach and substantially improves the performance of sadness expression.



(a) 6-class classification using Conditional Random Fields (CRF)



(b) 6-class classification using Latent-Dynamic Conditional Random Fields (LDCRF)

Figure 4.3: Recognition Performance for 6-class classification for a random subset of sequences by temporal modeling of shapes (Horizontal Lines on the top show the ground truth and predicted labels).

4.3.2 Classification Results with Neutral (7-class)

The case for 7-class classification where we consider neutral frames as well is relatively difficult. Some expressions that have very little facial movement have very similar shapes as the neutral shape which makes it hard to discriminate them. The difficulty in differentiating between neutral shapes and the expression shapes during the transition phase from neutral to expression makes it more challenging.

The temporal dynamics between shapes become much more important in this situation and the results in Table 4.5 show that this is indeed true. As expected, the recognition performance of happiness and surprise is very high for this case also using either the static shape analysis or the dynamic shape analysis. The recall statistics in Table 4.6 & Table 4.7 are very high for both happiness and surprise, which means that shapes corresponding to surprise and happiness are very distinct as compared to other expressions and neutral shapes. The lower precision value can be attributed to some overlap between neutral shapes and these expressions especially at the onset of an expression.

	Ne	An	Di	Fe	Ha	Sa	Su	Avg
Templates	66.67	65.61	61.64	76.06	87.92	54.43	96.26	72.66
SVM	71.32	**77.09**	82.77	75.81	96.92	56.15	97.45	79.64
CRF	72.17	73.54	**85.62**	**94.37**	98.06	69.62	**99.06**	84.64
LDCRF	**73.46**	76.71	81.51	**94.37**	**98.55**	**77.22**	**99.06**	**85.84**

Table 4.5: Recognition Rates with Neutral expression using Shape Features (7-class classification). [Neutral(Ne), Anger(An), Disgust(Dn), Fear(Fe), Happiness (Ha), Sadness(Sa), Surprise(Su), Average(Avg)]

The confusion matrix in Table 4.6 & Table 4.7 show that neutral frames cause a lot of confusion with subtle expressions such as anger and sadness and make these expressions difficult to recognize. The static shape analysis gives very low performance for fear and sadness expressions. The temporal shape analysis improves the recognition rate significantly for these expressions. It can be seen that the proposed LDCRFs based method is capable of better discriminating between the neutral and sadness shapes as compared to static approaches and CRFs. The average recognition rate for CRFs is 84.64% while for LDCRFs is 85.84%. Though this is a small performance increase but for a difficult expression like sadness, LDCRFs outperforms CRFs by more than 7%. For anger and disgust, SVMs, CRFs and LDCRFs all perform in a comparable manner.

	Ne	**An**	**Di**	**Fe**	**Ha**	**Sa**	**Su**	**Prec.**	**Rec.**
Ne	**72.2**	6.1	2.6	1.6	2.4	10.5	4.5	0.87	0.72
An	20.6	**73.5**	0.0	0.0	0.0	5.8	0.0	0.74	0.74
Di	2.7	6.8	**85.6**	0.0	4.8	0.0	0.0	0.88	0.86
Fe	0.0	0.0	0.0	**94.4**	0.0	5.6	0.0	0.86	0.94
Ha	0.5	1.0	0.5	0.0	**98.1**	0.0	0.0	0.90	0.98
Sa	29.1	0.0	0.0	1.3	0.0	**69.6**	0.0	0.41	0.70
Su	0.9	0.0	0.0	0.0	0.0	0.0	**99.1**	0.88	0.99

Table 4.6: Confusion Matrix (percentage) and Precision-Recall Statistics for 7-class classification using CRFs. [Neutral(Ne), Anger(An), Disgust(Dn), Fear(Fe), Happiness (Ha), Sadness(Sa), Surprise(Su)]

As noted above and shown with the principal component analysis, introduction of the neutral state brings a lot of overlap with certain expressions. The performance results using our proposed approach reflect the same. The

	Ne	**An**	**Di**	**Fe**	**Ha**	**Sa**	**Su**	**Prec.**	**Rec.**
Ne	**73.5**	6.0	1.6	1.9	2.6	9.2	5.2	0.88	0.73
An	20.6	**76.7**	1.1	0.0	1.6	0.0	0.0	0.75	0.77
Di	2.7	6.2	**81.5**	0.0	9.6	0.0	0.0	0.91	0.82
Fe	0.0	0.0	0.0	**94.4**	0.0	4.2	1.4	0.84	0.94
Ha	0.5	1.0	0.0	0.0	**98.6**	0.0	0.0	0.86	0.99
Sa	21.5	0.0	0.0	1.3	0.0	**77.2**	0.0	0.50	0.77
Su	0.9	0.0	0.0	0.0	0.0	0.0	**99.1**	0.87	0.99

Table 4.7: Confusion Matrix (percentage) and Precision-Recall Statistics for 7-class classification using proposed method based on LDCRFs. [Neutral(Ne), Anger(An), Disgust(Dn), Fear(Fe), Happiness (Ha), Sadness(Sa), Surprise(Su)]

accuracies for happiness and surprise are still high for both 6-class and 7-class classification. Performance for other expressions like anger and sadness suffer when neutral state is present. The recognition performance for anger falls down from 97.91% to 76.71% and for sadness from 90.08% to 77.22%.

It was observed that the misclassification usually occur during the transition phase from one expression to either neutral or to other expression. Since we formulated the task of facial expression recognition as a supervised sequence labeling problem, we have reported the results based on the classification/misclassification of each frame and not considering the dominant expression within a certain window of time. It makes the problem harder because as mentioned previously, it's very difficult to label the ground-truth for frames which lie in the transition phase. For practical applications, it is certainly not necessary for the system to label each and every frame correctly.

4.4 Dynamic Shape Analysis vs. Dynamic Appearance Analysis

The appearance on face changes in form of wrinkles and furrows which appear when an expression is exhibited. In this section we show that in contrast with shape features, temporal variations in appearance alone is not sufficient to recognize facial expressions with high accuracy. Using CRF and LDCRF techniques we model Uniform Local Binary Pattern (ULBP) based appearance features which are known for their ability to capture these micro patterns (e.g. wrinkles and furrows) on the face and have been successfully used for static facial expression recognition.

The results in Table 4.8 clearly show that except for Happiness and Sadness the performance is much lower in comparison with the performance of dynamic shape analysis. The interesting thing here is that the performance becomes worse on introducing the neutral state (Table 4.10). The reason for this is that for expressions like Anger, Disgust and Sadness the small amount of facial motion does not bring a significant change in the appearance in comparison with the neutral face which results in a considerable overlap of appearance features between them. This makes it difficult to distinguish these expressions from the neutral state using just the appearance. These experiments show that, dynamics of shape and the ability to capture the subtle motion patterns on the face is very important for robust facial expression recognition.

Several researchers (refer to the survey in [33]) have presented a view that a combination of shape and appearance may be a better way to design

facial expression recognition systems. Our initial experiments with the combination of both shape and appearance features did not show any conclusive improvements. Also, there has been no prior work which models the temporal variations of appearance in isolation and analyze their usability for recognizing facial expressions. The work of Littlewort et al. [27] which is considered to be state of the art, uses only appearance features (Gabor filters) along with SVM based classifiers thus considering each image in isolation. A detailed comparison of their approach over the dataset used in this work will be an useful analysis to support our findings regarding the use of appearance features.

	An	Di	Fe	Ha	Sa	Su
LBP + CRF	70.42	85.54	67.61	90.90	60.20	89.15
LBP + LDCRF	76.28	86.01	80.05	90.01	58.68	87.81
Shape + CRF	96.41	97.60	**92.51**	99.41	83.83	97.86
Shape + LDCRF	**97.91**	**97.86**	90.52	**99.55**	**90.08**	**98.87**

Table 4.8: Comparison of Recognition Rates without Neutral expression between Shape and Appearance Features (6-class classification). [Anger(An), Disgust(Dn), Fear(Fe), Happiness (Ha), Sadness(Sa), Surprise(Su), Average(Avg)]

	Avg
LBP + CRF	77.30
LBP + LDCRF	79.81
Shape + CRF	94.60
Shape + LDCRF	**95.79**

Table 4.9: Comparison of Average Recognition Rates without Neutral expression between Shape and Appearance Features (6-class classification). [Anger(An), Disgust(Dn), Fear(Fe), Happiness (Ha), Sadness(Sa), Surprise(Su), Average(Avg)]

	Ne	An	Di	Fe	Ha	Sa	Su
LBP + CRF	**87.80**	61.59	65.22	47.20	87.84	49.37	91.28
LBP + LDCRF	85.41	62.73	66.80	55.43	84.28	51.17	93.76
Shape + CRF	72.17	73.54	**85.62**	**94.37**	98.06	69.62	**99.06**
Shape + LDCRF	73.46	**76.71**	81.51	**94.37**	**98.55**	**77.22**	**99.06**

Table 4.10: Comparison of Recognition Rates with Neutral expression between Shape and Appearance Features (7-class classification). [Neutral(Ne), Anger(An), Disgust(Dn), Fear(Fe), Happiness (Ha), Sadness(Sa), Surprise(Su), Average(Avg)]

	Average
LBP + CRF	70.05
LBP + LDCRF	71.36
Shape + CRF	84.64
Shape + LDCRF	**85.84**

Table 4.11: Comparison of Average Recognition Rates with Neutral expression between Shape and Appearance Features (7-class classification). [Neutral(Ne), Anger(An), Disgust(Dn), Fear(Fe), Happiness (Ha), Sadness(Sa), Surprise(Su), Average(Avg)]

Chapter 5

Conclusion & Future Work

We have presented a new approach for facial expression recognition from video sequences using Latent-Dynamic Conditional Random Fields (LDCRFs). The results of our approach show that the expressions such as surprise and happiness which bring significant changes in face shapes are relatively easy to recognize. For other more subtle expressions, classification methods which do not consider the temporal variation between shapes fail to achieve a good recognition rate. Sadness and Anger are two most difficult expressions to classify especially in presence of neutral frames. The proposed method was able to perform better as compared to other techniques for these expressions. This shows the importance of modeling small facial motions effectively for recognizing facial expressions.

Our experiments also show that shape provides much richer information as compared to appearance and modeling appearance changes in isolation without considering shape changes is not sufficient for robust facial expression recognition.

Facial expression recognition is an active research field among the vision community. There are several open possibilities for enhancing our current

work. Though our current approach performs better than several other methods, still the performance for some of the expressions like anger and sadness is relatively low in presence of neutral shapes. In future, one of the things we want to focus on is improving the performance for these expressions.

We also want to evaluate the performance of our approach by training and testing it across various datasets. Most of the results in the field of facial expression recognition have been reported on either standard datasets or data collected in controlled environments. We wish to go beyond that and see if we can extend the current work to handle real world issues like pose and illumination variations, recognizing expressions from continuous video streams like web-cams etc. We also want to analyze 3D face shapes and see if temporal modeling of 3D data can give us better results for recognizing facial expressions.

Bibliography

[1] Timo Ahonen, Student Member, Abdenour Hadid, Matti Pietikinen, and Senior Member. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:2037–2041, 2006.

[2] Zara Ambadar, J. Schooler, and Jeffrey Cohn. Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions. *Psychological Science*, 2005.

[3] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *In CVPR Workshop on CVPR for HCI*, 2003.

[4] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior, 2005.

[5] Marian Stewart Bartlett, Gwen Littlewort, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In *Proc. IEEE Intl Conf. Systems, Man and Cybernetics*, pages 592–597, 2004.

[6] Marian Stewart Bartlett, Gwen C. Littlewort, Mark G. Frank, Claudia Lainscsek, Ian R. Fasel, and Javier R. Movellan. Automatic recognition of facial actions in spontaneous expressions, 2006.

[7] J. N. Bassili. Facial motion in the perception of faces and of emotional expression. 1978.

[8] Michael J. Black and Yaser Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25:23–48, 1997.

[9] Fabrice Bourel, Claude C. Chibelushi, and Adrian A. Low. Recognition of facial expressions in the presence of occlusion. In *British Machine Vision Conference*, 2001.

[10] Vicki Bruce. What the human face tells the human mind: some challenges for the robot-human interface. *Advanced Robotics*, 8:341–355, 1993.

[11] Ira Cohen, Ashutosh Garg, and Thomas S. Huang. Emotion recognition from facial expressions using multilevel hmm. In *Neural Information Processing Systems*, 2000.

[12] Ira Cohen, Nicu Sebe, Larry Chen, Ashutosh Garg, and Thomas S. Huang. Facial expression recognition from video sequences: Temporal and static modelling. In *Computer Vision and Image Understanding*, pages 160–187, 2003.

[13] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:681–685, 2001.

[14] Charles Darwin. *The expression of the emotions in man and animals /.* New York ;D. Appleton and Co.,, 1916. http://www.biodiversitylibrary.org/bibliography/482

[15] Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. Emotion recognition using phog and lpq features. In *FG*, pages 878–883, 2011.

[16] P. Ekman. Strong evidence for universals in facial expressions: A reply to russell"s mistaken critique. 2000.

[17] P. Ekman and W. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17:124–129, 1971.

[18] P. Ekman and W. V. Friesen. Facial action coding system investigator"s guide. 1978.

[19] P. Ekman and E. Rosenberg. What the face reveals. 1997.

[20] Irfan Essa and Alex Pentland. Facial expression recognition using image motion, 1997.

[21] Hai Hong, Hartmut Neven, and Christoph Von Der Malsburg. Online facial expression recognition based on personalized galleries. In *Proc. IEEE FG*, pages 354–359.

[22] C.L. Huang and Y.M. Huang. Facial expression recognition using model-based feature extraction and action parameters classification. In *Journal of Visual Communication and Image Representation*, pages 278–290, 1999.

[23] Takeo Kanade, Jeffrey F. Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *in Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53.

[24] Atul Kanaujia and Dimitris N. Metaxas. Recognizing facial expressions by tracking feature shapes. In *International Conference on Pattern Recognition*, pages 33–38, 2006.

[25] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, San Fransisco, 2001. Morgan Kaufmann.

[26] Ying li Tian, Takeo Kanade, and Jeffrey F. Cohn. Chapter 11. facial expression analysis.

[27] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian R. Fasel, Mark G. Frank, Javier R. Movellan, and Marian Stewart Bartlett. The computer expression recognition toolbox (cert). In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 298–305, 2011.

[28] P. Lucey, J.F. 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 *CVPR4HB10*, pages 94–101, 2010.

[29] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. pages 200–205, 1998.

[30] Andrew McCallum. Efficiently inducing features of conditional random fields, 2003.

[31] A. Mehrabian. *Communication without words*, pages 51–52. 2 edition, 1968.

[32] Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. Latent-dynamic discriminative models for continuous gesture recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2007.

[33] M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[34] M. Pantic and L. Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing Journal*, 18(11):881–905, 2000.

[35] Maja Pantic and Lon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1424–1445, 2000.

[36] Montse Pards and Antonio Bonafonte. Facial animation parameters extraction and expression recognition using hidden markov models. *Signal Processing-image Communication*, 17:675–688, 2002.

[37] Alex Pentland, Baback Moghaddam, and Thad Starner. View-based and modular eigenspaces for face recognition. In *in Proceedings of Fourth IEEE International Conference on Computer Vision and Pattern Recognition*, 1994.

[38] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

[39] Nicu Sebe, Michael S. Lew, Ira Cohen, Yafei Sun, Theo Gevers, and Thomas S. Huang. Authentic facial expression analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 517–522, 2004.

[40] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vision Comput.*, 27:803–816, May 2009.

[41] Eero P. Simoncelli. Distributed representation and analysis of visual motion, 1993.

[42] Cristian Sminchisescu, Atul Kanaujia, Zhiguo Li, and Dimitris Metaxas. Conditional models for contextual human motion recognition. In *In Intl Conf. on Computer Vision*, pages 1808–1815, 2005.

[43] M. B. Stegmann and D. D. Gomez. A brief introduction to statistical shape analysis, mar 2002. Images, annotations and data reports are placed in the enclosed zip-file.

[44] Charles Sutton and Andrew Mccallum. *Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2006.

[45] Hai Tao and Thomas S. Huang. Explanation-based facial motion tracking using a piecewise bzier volume deformation model. In *Computer Vision and Pattern Recognition*, pages 1611–1617, 1999.

[46] Ying-Li Tian, Takeo Kanade, and Jeffrey Cohn. Recognizing lower face action units for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, pages 484 – 490, March 2000.

[47] Michel Valstar, Bihan Jiang, Marc Mehu, Maja Pantic, and Klaus Scherer. The first facial expression recognition and analysis challenge. In *Proceedings of IEEE International Conference on Automatic Face and Gesture*

Recognition, Workshop on Facial Expression Recognition and Analysis Challenge, 2011.

[48] Michel F. Valstar, I. Patras, and Maja Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data, 2005.

[49] Jones Viola. Rapid object detection using a casacde of simple features. In *Computer Vision and Pattern Recognition*, 2001.

[50] Hanna M. Wallach. Conditional random fields: An introduction. Technical Report MS-CIS-04-21, University of Pennsylvania, Philadelphia, 2004.

[51] Sy Bor Wang, Ariadna Quattoni, Louis philippe Morency, David Demirdjian, and Trevor Darrell. Hidden conditional random fields for gesture recognition. In *Computer Vision and Pattern Recognition*, pages 1521–1527, 2006.

[52] Laurenz Wiskott, Jean-Marc Fellous, Norbert Kruger, and Christoph von der Malsburg. Face recognition by elastic bunch graph matching. In *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, pages 355–396, 1999.

[53] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:915–928, 2007.

[54] Y. Zhu, Liyanage C. De Silva, and Chi Chung Ko. Using moment invariants and hmm in facial expression recognition. *Pattern Recognition Letters*, 23(1-3):83–91, 2002.

Vita

Suyog Dutt Jain was born in India on 02 May 1985, the son of Sunil Dutt Jain and Sangeeta Jain. He received the Bachelor of Engineering Degree from Manipal University, Manipal in 2008. During his undergraduate studies, he spent six months as a research intern at Indian Statistical Institute, Kolkata working on medical images. He also worked as a research assistant at Indian Institute of Technology, Bombay in 2008-09 developing tools for weather data visualization. Subsequently he joined University of Texas at Austin in Fall 2009 for his graduate studies in computer science.

Permanent address: 20/291 Shanti Niketan, Plaza Road
Ajmer, India 305001

This thesis was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.