Simultaneously tracking and recognition of facial features and facial expressions

A Thesis Submitted in Partial Fulfilment

Of the Requirements for the Award of the Degree of

Dual Degree [B.Tech. & M. Tech.]

In

Electrical Engineering

By

SRINATH KOILAKONDA

(Roll No.710EE1111)

May, 2015

Department of Electrical Engineering

National Institute of Technology

Rourkela-769008

Simultaneously tracking and recognition of facial features and facial expressions

A Thesis Submitted in Partial Fulfilment

Of the Requirements for the Award of the Degree of

Dual Degree [B.Tech. & M. Tech.]

In

Electrical Engineering

By

SRINATH KOILAKONDA

(Roll No.710EE1111)

May, 2015

Under the Guidance of

Prof. Dipti patra

Department of Electrical Engineering

National Institute of Technology

Rourkela-769008

National Institute of Technology

Rourkela

CERTIFICATE

This is to certify that the thesis entitled, "**Simultaneously tracking and recognition of facial features and facial expressions"** submitted by **SRINATH KOILAKONDA** in partial fulfillment of the requirements for the award of Dual Degree B. Tech. and M. Tech. in Electrical Engineering with specialization in "**ELETRONIC SYSTEMS & COMMUNICATION** "during 2014 - 2015 at the National Institute of Technology, Rourkela is an authentic work carried out by him under my supervision and guidance.

To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University / Institute for the award of any Degree or Diploma.

Date ……

Prof. Dipti Patra

Department of Electrical Engineering

National Institute of Technology

Rourkela-769008

Acknowledgement

I am indebted to many people who contributed through their support, knowledge and friendship, to this work and the years at NIT Rourkela.

I am grateful to my guide **Prof. DIPTI PATRA** for giving me the opportunity to work on this area with vast opportunities. His valuable guidance made me learn some of the advanced concepts during my work. I sincerely appreciate the freedom Prof. DIPTI PATRA provided me to explore new ideas in the field of my work. He supported and encouraged me throughout the project work.

I am thankful to our Head of the Department, Prof. A.K Panda, for providing us the facilities required for the research work.

My hearty thanks to all my friends, for their help, co-operation and encouragement.

I render my respect to all my family members for giving me mental support and inspiration for carrying out my research work.

SRINATH KOILAKONDA

Roll No.710EE1111

Contents

CHAPTER-1

INTRODUCTION

CHAPTER I

INTRODUCTION

1.1 INTRODUCTION

Face recognition is the most important applications of image analysis. It's a true challenge to build an automated system which equals human ability to recognize faces. Even though human beings are good identifying known faces, we are not capable when we must deal with a large amount of unidentified faces. The computers, with an almost limitless memory and computational speed, should overcome human's limitations.

Affective state shows a fundamental role in humanoid interactions, influencing cognition, opinion & even rational choice making. This element has enthused the research field of "affective computing" which ambitions at enabling computers to identify, interpret & simulate affects [17]. Such systems can contribute to humanoid computer communication and to applications such as learning environment, entertainment, customer service, computer games, security/surveillance, and educational software as well as in safety critical application such as driver monitoring [9]. To make human computer interaction (HCI) more normal & inviting, it would be helpful to give PCs the capacity to perceive affects the same way a human does. Since speech and vision are the essential faculties for human expression and recognition, critical exploration exertion has been centered around creating canny frameworks with sound and feature interfaces.

Facial recognition remains as an unresolved problem and essential technology. There are various industry territories that keen on what it could offer. A few illustrations incorporate feature reconnaissance, human-machine cooperation, photograph cameras, virtual reality or law requirement. Face recognition is a related subject in pattern recognition, neural systems, PC representation, picture handling and brain science. Some applications are hovering the interest on face recognition. It is narrow initial application area is being expanded. Some applications show as table 1.

1.2 Face recognition system structure

Face Recognition is a term that includes numerous sub-problems.

1.2.1 A generic facial recognition system

The input (I/P) of a facial recognition system every time is an image or video stream. The output of the system is an identification or verification of the subject or subjects that appear in the image or video. Some methods [30] define a face recognition system as a three step procedure - see Figure 1.1. From this point of view, the Facial Detection and facial Feature Extraction stages can run separately.

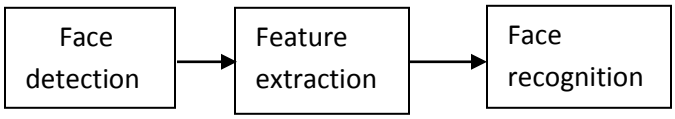


Figure 1.1: A generic Face recognition system [32].

Facial detection is characterized as the procedure of extracting faces from divisions or scenes. Thus, the framework emphatically distinguishes a certain picture locale as a face. This system has numerous applications like face following, posture estimation or pressure. The following step -highlight extraction- includes acquiring applicable facial elements from the information. These elements could be sure face districts, variations, Angles or measures, which can be human important (e.g. eyes dividing) or not. This stage has different applications like facial element following or expression recognition. In final step, the framework does recognize the face.

1.3 Face detection

Nowadays some applications of Face Recognition don’t require face detection. In some cases, face images stored in the data bases are already normalized. There is a standard image input format, so there is no need for a detection step. An example of this could be a criminal data base. There, the law authorization organization stores countenances of individuals with a criminal report. In the event that there is new subject and the police has his or her identification photo, face discovery is redundant. Nonetheless, the routine information picture of PC vision frameworks is not that suitable. They can contain numerous things or countenances. In these cases face identification is obligatory. It's likewise unavoidable in the event that we need to build up a robotized face following framework. Case in point, feature observation frameworks attempt to incorporate face location, following and perceiving. Along these lines, its sensible to expect face discovery as a major aspect of the more abundant face acknowledgment issue. Face identification must manage a few no doubt understood difficulties [31]. They are normally present in pictures caught in uncontrolled situations, for example, observation feature frameworks. These difficulties can be credited to a few factors:

- Posture variation: The perfect situation for face discovery would be the one in which just front images were included. But, as indicated, this is improbable all in all uncontrolled conditions. Moreover, the execution of face recognition calculations drops extremely when there are expansive stance varieties. It's a noteworthy examination issue. Posture variety can happen because of subject's activities or camera's angle..
- Feature occlusion: The manifestation of elements like beards, glasses or hats introduces high variability. Faces can also be moderately covered by objects or other faces.
- Facial expression: Facial features also vary significantly because of different facial gestures.
- Imaging conditions: cameras and surrounding conditions can influence the nature of a images, aggravating the appearance of a face.

There are a few issues firmly identified with face discovery other than highlight extraction and face order. Case in point, face area is a disentangled methodology of face identification. It is objective as to focus the area of a face in a picture where there's one and only face. We can

separate between face recognition and face area, since the last is a disentangled issue of the previous. Systems like finding head limits were initially utilized on this situation and after that sent out to more muddled issues [7]. Facial element location concerns identifying and finding some significant components, for example, nose, eyebrow, lips, ears, and so forth. Some element extraction calculations are in light of facial component location. There is much writing on this point, which is talked about later. Face following is other issue which in some cases is an outcome of face recognition. Much framework's objective is to distinguish a face, as well as to have the capacity to find this face progressively. At the end of the day, feature reconnaissance framework is a decent illustration.

1.4 Approaches to face detection

It’s not very easy to give categorization of face detection approaches. There is not an all-inclusive recognizing gathering rule. They more often than not to blend and cover. In this method, two arrangement criteria will be introduced. One of them separates between unmistakable situations. Contingent upon these situations diverse methodologies may be required. The other criteria divide the detection algorithms into four categories.

- Controlled environment: It is the most straightforward case. Photos are taken under controlled light, foundation, and so on. Basic edge identification systems can be utilized to recognize faces.
- Colour images: The distinctive skin colors also can be used to find out faces. They can be feeble if light conditions change. Also, human skin shading changes a considerable measure, from white to verging on dark. Yet, a few studies demonstrate that the significant contrast lies between their forces, so chrominance is a decent component [31]. It's not simple to set up a strong human skin shading representation. On the other hand, there are activities to construct hearty face identification scheming that take into account of skin shading.
- Images in motion: Real time feature gives the opportunity to utilize movement discovery to restrict faces. These days, most business frameworks must find confronts in features there is a proceeding with test to accomplish the best recognizing results with the best

conceivable execution [13]. Another methodology in light of movement is eye squint location, which has numerous uses beside face recognition.

1.5 Recognition Problems

Facial expression recognition systems, in particular, have matured to a level where automatic detection of small number of expressions in posed and controlled displays can be achieved with reasonably high accuracy. Automatic analysis of human affective behavior has been extensively studied in past several decades. These emotions are often deliberate and exaggerated displays. However, the deliberate and spontaneous behavior differs in their visual appearance, audio profile and the timing between the two modalities. Detecting these expressions in less constrained settings during spontaneous behavior is still a challenging problem. The research shift towards this direction suggests utilizing the multimodal data analysis approaches.

1.6 Motivation

Multimodal systems, specifically with audio and visual modalities, have shown several interesting interactions between the two modalities. For example, audio-visual speech recognition (AVSR), also recognized as spontaneous lip-reading or speech reading goes for enhancing programmed discourse acknowledgment by investigating the visual methodology of the speaker's mouth district [14]. Not surprisingly, it has outperformed audio alone ASR system particularly in noisy conditions. Similarly, the well-known perceptual phenomenon, McGurk effect [10], which demonstrates an interaction between hearing and vision in speech perception. Furthermore, Munhall *et al* suggests that rhythmic head movements are correlated with the pitch and amplitude of speaker’s voice and that visual information can improve speech intelligibility by 100% over that possible using auditory information only [11].

In the field of affect recognition, there have been number of efforts to exploit audio-visual information as well and our framework can utilize these methods. However, above examples, where visual modality improves audio alone system, are motivated us to ask the fundamental question of how does audio modality influence visual perception, in particular, for the task of

facial expression recognition. It is evident that speech generation influences facial expression. Also, for expression recognition the coupling between these two modalities is not so tight unlike the case in audio-visual speech recognition task.

1.7 Objective of the work

A novel face expression recognition system using bimodal information. Our framework explicitly models the cross-modality data correlation while allowing them to be treated as asynchronous streams. To recognize the key emotion of an image sequence, the proposed framework seeks to summarize the emotion using one single image derived from hundreds of frames contained in the video. We also show that the framework can improve the recognition performance while significantly reducing the computational cost by avoiding redundant or insignificant frame processing using auditory information.

1.8 Organization of the thesis

Present thesis is organized in seven chapters as discussed below:

Chapter 1: This chapter includes general introduction, recognition problems, motivation and objective of the work. Finally, this thesis chapter explains organization of the thesis.

Chapter 2: This chapter explains literature reviews of various journals and conference publications.

Chapter 3: This chapter includes block overview and determines audio visual data approach with details.

Chapter 4: This chapter includes database experimental analysis and also determines analysis of expressions.

Chapter 5: This chapter includes synthesis and simulated results of thesis work.

Chapter 6: This chapter includes the conclusion of the work with limitations and future scope.

In this chapter introduction of this thesis, overview of this thesis discussed. Also explains as a facial expression recognition system, where automatic detection of small number of expression in posed and controlled display can be achieved with high accuracy.

CHAPTER 2

LITERATURE REVIEW

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This section summarizes the pioneer works on face expression recognition. Our long term goal is to study the cross-modal influence of the audio-visual data streams on each other for the affect recognition task. In this study, however, our focus is on face expression recognition.

2.2 Review

Firstly discuss some of the representative works for facial expression recognition and then move our discussion on existing audio-visual affect recognition approaches to highlight the challenges lies in the integration of the two modalities. For an overview of audio only, visual only and audio-visual affect recognition, readers are encouraged to study a recent survey by Zeng *et al.* [33]. Because of the significance of face in expressions and their recognition, the majority of the vision-based recognize studies center. A lot of existing outward appearance recognizers utilize different example acknowledgment approaches and are in light of 2D spatiotemporal facial components: geometric features or appearance based features. Geometric based approaches track the facial geometry information over time and classify expressions based on the deformation official feature. Chang *et al.* defined a set of points as the facial contour feature, and an Active Shape Model (ASM) is learned in a low dimensional space [4]. Lucey *et al.* employed Active Appearance Model (AAM)-derived representation while Valtar, Patras, and Pantic tracked 20 fiducial facial points on raw video using a particle filter [6] [26].

Then again, appearance-construct methodologies stress with respect to depicting the presence of facial elements and their progress. Zhao and Pietikaninen utilized the dynamic Local Binary Pattern (LBP) which has the capacity remove data along the time hub [37]. Bartlett et al. utilized a bank of Gabor wavelet channel to break down the facial surface. All the more as of late, Wu et al. used Gabor Motion Energy Filters which is likewise ready to catch the spatial-

fleeting data [32]. Yang and Bhanu created a single good image representation from a visual sequence by first registering the face image to an reference image using dense SIFT flow algorithm and extract appearance feature using Local Phase Quantization (LPQ) [32]. The method has provided the best overall emotion recognition performance till date for the GEMEP-FERA benchmark [27]. This can be derived as one of the special cases in our framework. It is important to mention that precise registration of frames is an important step otherwise single representation of image sequence using all the frames could suffer from large deviation of head pose.

Cohen *et al.* performed expression classification in video sequence using temporal and static modeling by Naive-Bayes based (‘static’) and HMM based (‘dynamic’) classifiers respectively [2]. Static classifiers outperformed dynamic ones. It is contended that dynamic classifiers are more troublesome, so they oblige additionally preparing examples and numerous more parameters to learn contrasted and the static methodology. Creator proposes that dynamic classifiers are more suited for individual ward frameworks in view of their higher affectability not simply to changes in appearance of expressions among changed individuals, also to the refinements in momentary cases. Static classifiers are less requesting to plan and realize, yet when used on a steady element gathering, they can be faulty especially for edges that are not at the peak of an expression.s this brings an important aspect of how to obtain a better and robust representation of an expression from video sequences.

Figure 2.1: A spontaneous conversation between driver and passenger during a driving task [21].

As far as automatic facial affect recognition is concerned, most of the existing efforts studied the expressions of the six basic emotions (Happy, Sad, Surprise, Fear, Anger and Disgust) due to their universal properties and the availability of the relevant training and test material. These emotions are often deliberate and exaggerated displays [22]. The deliberate behavior, however, differs in visual appearance, audio profile, and timing from spontaneously occurring behavior [19]. This has led the research field to new trends: analysis of spontaneous affective behavior and development of multimodal analysis. Multimodal analysis helps to improve the performance in challenging naturalistic setting during spontaneous behavior. Combining complementary information from the two streams can help improve the recognition performance. However, the two modalities are not tightly coupled in spontaneous naturalistic behavior as depicted in Fig. 1, Film strip shows tests of five pictures similarly dispersed in the expression. To start with a large portion of the articulation contains the discourse and later a large portion of the street commotion. Notification, nonetheless, that facial elements are more expressive after discourse substance while head motion is attending with the discourse [21].

Moreover, speech generation affects the facial expression dynamics. We present some of the works which address these two issues. In particular, how they derive various visual representations for visual channel as well as how they model asynchrony in the two streams.

One of the testing undertakings of the visual following frameworks is to manage changes fit as a fiddle of the mouth created because of discourse. Keeping in mind the end goal to manage this circumstance, Datcu et al. [3] proposed an information combination system where they depend just on the visual information in the noiseless period of the feature succession and the intertwined varying media information amid non-quiet sections. The visual methodology amid non-quiet sections just centered on the upper a large portion of the facial district to wipe out the impacts brought on by changes fit as a fiddle of the mouth. However, the result shows that full face based model performs superior than partial face. Hence an alternative strategy is requiring filtering out the influence of phonemes.

Wang *et al.* proposed a generally reasonable computational technique for visual based feeling acknowledgment which chose a solitary key edge from every varying media succession to speak to the feeling present in the whole arrangement [29]. The standard for selecting the key

edges from the varying media groupings was in view of the heuristic that top feelings are shown at the most extreme sound intensities. The visual components are removed from these key edges utilizing Gabor wavelets. An acoustic feature is then combining with derived visual feature at feature level data fusion scheme for the classification task. However, choosing one single frame from visual sequence is very restrictive and the same is clear from the performance of their visual alone system.

An important audio visual fusion scheme which aim at making use of the correlation between audio and visual data streams and relaxing the requirement of synchronization of these streams, is that of model-level fusion. Zeng *et al.*presented a Multistream Fused HMM to build an optimal connection among multiple streams from audio and visual channels according to the maximum entropy and the maximum mutual information criterion [37]. Author, however, considered tightly coupled HMMs. Song *et al.* proposed an approach for multimodal feeling acknowledgment which was particularly centered around transient investigation of three arrangements of components: 'sound just elements', 'visual just highlights' (upper 50% of facial district) and 'visual discourse highlights' (lower a large portion of facial area) utilizing a triple HMM, i.e., one HMM for each of the data modes [20]. This model was proposed to manage state asynchrony of the varying media components while keeping up the first connection of these elements over the long haul. On the other compelling is the model that permits complete asynchrony between the streams. This is, be that as it may, infeasible because of the exponential increment in the quantity of state mixes conceivable because of the asynchrony.

Our contribution in this thesis is two folds: first, we explicitly model the correlation between the two streams while allowing them to be treated as asynchronous streams; second, we assign importance to a particular frame and thereby avoiding extreme treatment (all the frames or just a single frame). More importantly this is accomplished by incorporating cross-modal models developed at the first step. The idea is that the analysis of the sequential changes can be beneficial for the facial expression recognition, however, the onset and the offset of the facial dynamics are hard to detect using visual alone modalities. Hence most of the efforts often try to classify every frame and take a majority voting in the end to come up with single expression class. If the near apex frame or a set of more representative frames can be picked up based on multimodal data, to represent an entire segment, we can restrict noisy/redundant sequential facial

feature deformations to negatively influence the recognition performance, and hence describe emotions in a reliable manner. An initial finding based on the mentioned proposition was reported in [23]. Here, we provide further in depth analysis by statistically substantiating claims and compare multi-class classification performances with existing literature.

This section concludes the pioneer works on face expression recognition. This chapter also discuss about the audio-visual fusion scheme and a facial feature formations to dynamic expression with the speech.

CHAPTER 3

VISUAL DATA ASSOCIATION APPROACH

CHAPTER 3
VISUAL DATA ASSOCIATION APPROACH

3.1 Introduction

The proposed facial activity recognition system consists of two main stages: offline facial activity model construction and online facial motion measurement and inference. Specifically,
Using training data and subjective domain knowledge, the facial activity model is constructed offline.

The various computer vision techniques are used to track the facial feature points, and to get the measurements of facial motions, i.e., AU's. These measurements are then used as evidence to infer the true states of the three level facial activities simultaneously.
Figure 3.1 sketches an overview of the proposed recognition system. Salient feature of our framework is the introduction of modal relevance feedback blocks and frame relevance measure blocks. The cross-modal relevance feedback block measures the importance of the current frame of the other modality from the analysis of its modality. The frame relevance block can potentially use cross-modal feedback and the analysis of its modality to finally assess the relevance of the current frame.

In our present work, frame relevance block utilizes only cross-modal feedback to highlight the importance of cross modal information. Also, we have focused our discussion to facial expression recognition using visual features alone. Hence classification module only utilizes visual feature.

A visual classification framework, however, can be devised to utilize standard fusion schemes (early, model-level or late-fusion). Important point to note is that the proposed method at-tempts to improve signal representation at the first place hence by reducing error propagation

which, in general, is harder to deal at later stages.

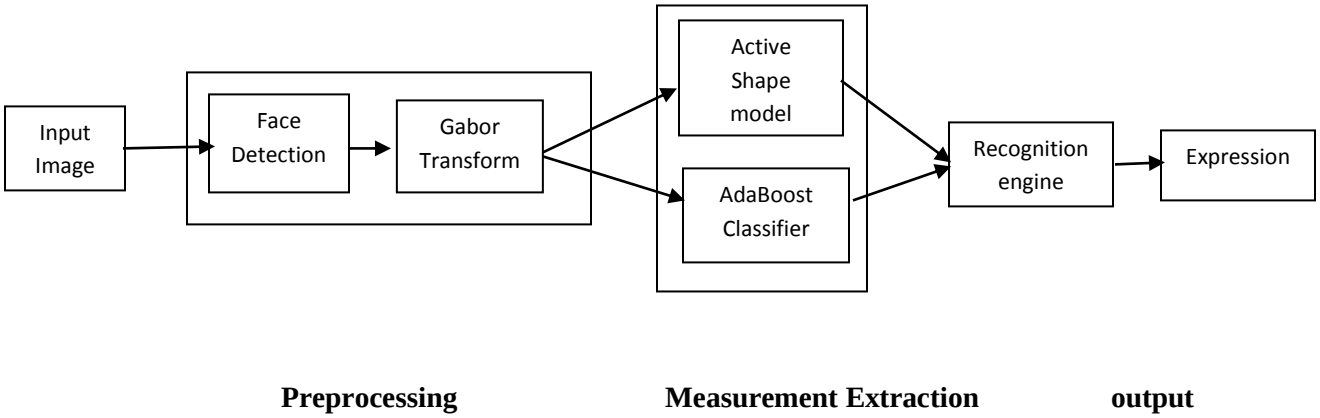


Figure 3.1: Overview of the proposer expression recognition system.

A detailed approach to condense the visual expression information into a single image representation is presented in following sections.

3.2 Face Tracking and Alignment

Many face recognition systems have a video sequence as the input. Those systems may require to be fit for recognizing as well as following appearances. Face tracking is basically a movement assessment problem. Face tracking can be achieved by utilizing a wide range of approaches, e.g., head tracking, component tracking, image based tracking, model-based tracking. These are distinctive approaches in order to classify these algorithms.

- Head tracking/Individual feature tracking: The head can be tracked all over the face element, or certain methods can follow independently and simultaneously.
- 2D/3D: Two dimensional frameworks track a face and yield a picture space where the face is found. Three dimensional frameworks, then again, perform a 3D displaying of the face. This methodology permits evaluating stance or introduction varieties.

The essential face following procedure tries to find a given picture in a photo. At that point, it needs to register the contrasts between edges to upgrade the area of the face. There are numerous issues that must be confronted: Partial obstructions, illumination changes, computational rate and facial deformations.

The first step of visual processing involves face detection and tracking. This is accomplished using constraint local model (CLM) [14]. It is based on fitting a parameterized shape model to the location landmark points of the face. The fitting process on an image I(m,n) provides a row vector P(m,n) for each sequence m and frame n containing 66 detected landmark positions.

$$P(m,n) = [x_1; y_1; x_2; y_2; \ldots\ldots\ldots\ x_l; y_l]$$

The detected landmark is normalized by appropriate scaling, rotation and translation to make center of eyes 200 pixel apart and line joining the two centers horizontal.

We denote the normalized shape vector as $P_N(m;n)$ .Further, a reference shape is calculated using Eq. 3.1

$$P^{ref} = \frac{1}{M}\sum_{m=1}^{M} \frac{1}{\mathrm{Nm}} \sum_{n=1}^{Nm} P\, a^{(m,n)} \qquad (3.1)$$

Where Nm is total number of frames in sequence m and M is the total number of image sequences. Given this Reference shape Pref, image I(m;n) is aligned using affine transform to obtain the aligned image I align(m;n). For alignment, we only considered the points which are relatively stable to track corresponding to the eyebrows, eyes, and nose and mouth regions.

Fig. 3.2 shows the reference shape obtained for the database and the points used for image alignment. Reference shape derived from the database showing 66 landmark positions along with the ones in block which are used during image alignment process.

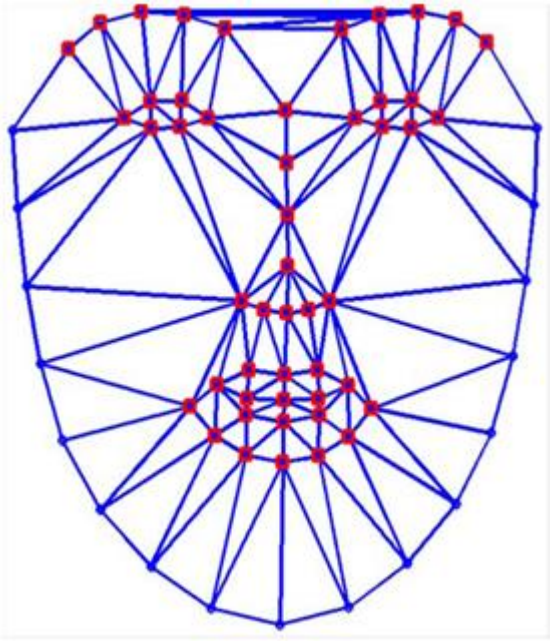


Figure 3.2: Reference shape landmark position [23]

An example of automatically tracked face and the aligned face is illustrated in Fig. 3.3. (a) An example of tracked face and the landmarks, and (b) aligned face image obtained using reference shape during image alignment step.

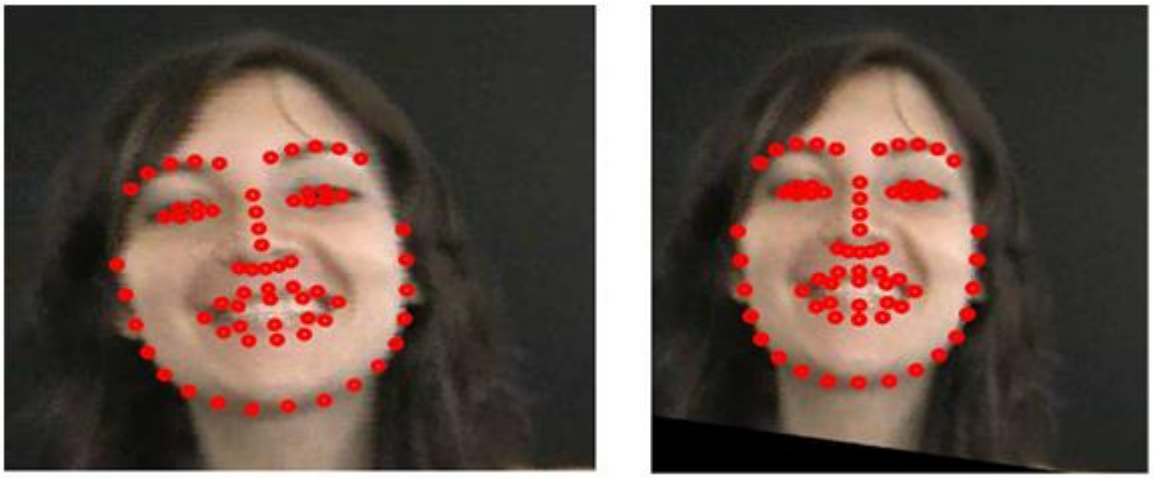


Figure 3.3: Tracked face and landmarks and Aligned face image [23]

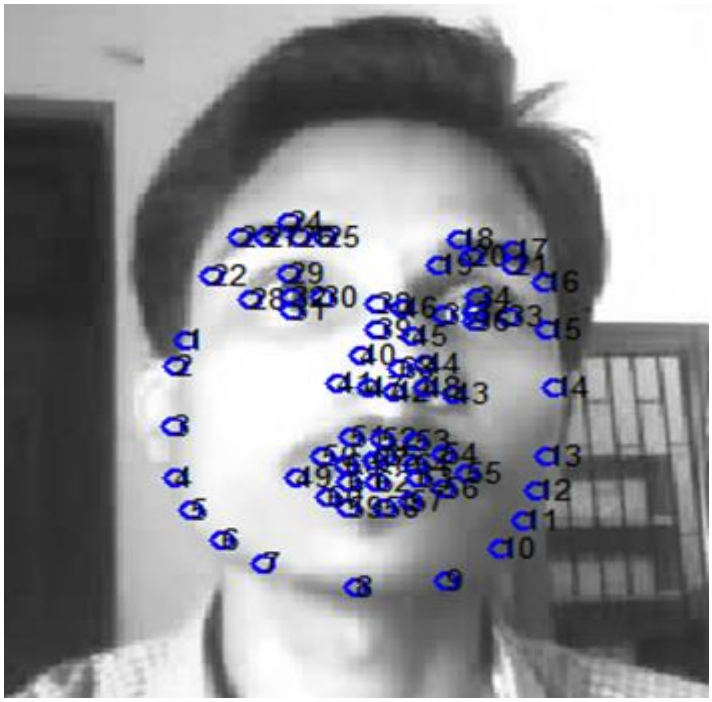


Figure 3.4: Tracked point at face

3.3 Appearance Feature Extraction

Originally proposed for texture analysis, the Local Binary Pattern (LBP) family of descriptors (LBP [15], LBP-TOP [34], LPQ [16] and LPQ-TOP [5]), in recent years, have been extensively used for static and temporal facial expression analysis, and face recognition. We use the blur insensitive LPQ (Local Phase Quantization) appearance descriptor proposed by Ojansivu *et al.* as the feature for facial expression analysis [16]. LPQ is based on computing the short-term Fourier transform on local image window. At each pixel the local Fourier coefficients are computed for four frequency points: [α0], [0α], [αα] and [α-α], where is sufficiently small number. We use in our experiment. Then phase information is recovered using binary scalar quantization of the signs of the real and imaginary part of each coefficient. The resultant eight bit binary coefficients are then represented as integers using binary coding. Finally, a histogram of these integer values from all image positions is composed and used as a 256-dimensional feature vector. We also use de-correlation process to eliminate the dependency of the neighboring pixels before quantization. In our experiment, we resize the aligned face images to 200 x 200 and further divided into non-overlapping tiles of 10 x 10 to extract local pattern. Thus the LPQ feature vector is of dimension 256 x 10 x 10 = 256.

3.4 Feature Extraction

In our prior work [24], [25] we have used prosodic and spectral features to model emotional states. We used subset of these features for cross-modal relevance calculation in the proposed framework. In particular, the pitch and intensity (energy) contours are used to derive weights w(n) for the nth frame in visual stream as described in Section 3.4.

For pitch contour calculation, we used the auto-correlation algorithm similar to [18]. The input speech signal is divided into overlapping frames with shift intervals (difference between the starting point of consecutive frames) of 10 ms. Every frame is of 60 ms in length to have the capacity to span 3 periodic times of minimum pitch value(for our situation 50 Hz). Pitch competitor over every frame is computed and a dynamic programming strategy is utilized to get the last pitch shape. Log-energy coefficients are calculated using 30 ms frames of outlines with movement interval of 10 ms. Fig. 3.5 demonstrates the inserted pitch shape and voiced fragment as well as the intensity contour.

This section defines as four major block of face expression recognition and overview of the proposed expression recognition system. This chapter also discuss about a face tacking and single image representation of the image sequence. A Bi-model approach has sequenced performed as audio visual analysis.

CHAPTER 4

DATABASE EXPERIMENTAL ANALYSIS

CHAPTER 4

DATABASE EXPERIMENTAL ANALYSIS

4.1 Introduction

The various facial expressions are driven by the muscular activities which are the direct results of emotion state and mental condition of the individual. Facial expressions are the visually detectable changes in appearance which represent the change in neuromuscular activity.Facial expressions could be identified by facial motion cues without any facial texture and complexion information [1].

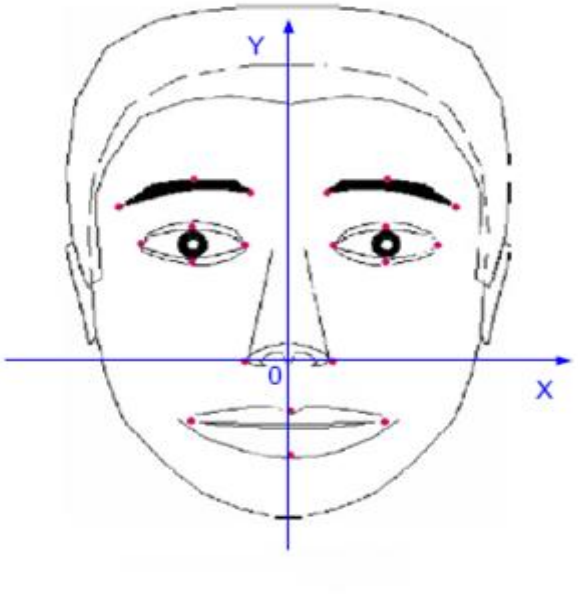


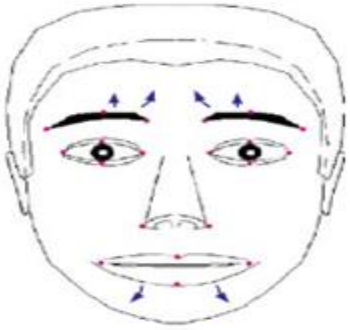
Figure 4.1: The facial coordinates [1]

4.2 Facial Movement Pattern for Different Emotions

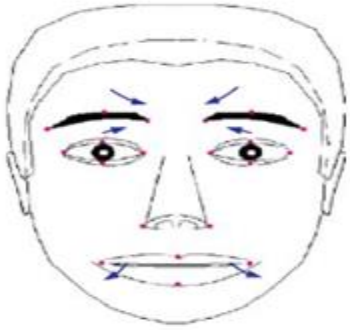
As illustrated in Figure 4.1 the principal facial motions provide powerful cues for facial expression recognition.

(a) happiness

(b) sadness



(c) fear



(d) disgust

(e) surprise

(f) anger

From Table 4.1 we can summarize the movement pattern of different facial expressions.

- When a person is happy, e.g. smiling or laughing, the main facial movement occurs at the lower half portion while the upper facial portion is kept still. The most significant feature

is that both the mouth corners will move outward and toward the ear. Sometimes, when laughing, the jaw will drop and mouth will be open.

Table 4.1: The facial movement's cues for six emotions [1].

Emotion	**Forehead & eyebrow**	**Eyes**	**Mouth & Nose**
Happiness	Eyebrows are Relaxed	Raise Upper and Lower lids slightly	Pull back and up lip corners toward the ears
Sadness	Bend together and upward the inner eyebrows	Drop down upper lids Raise lower lids slightly	Extend Mouth
Fear	Raise brows and pull together bent upward inner eye brows	Eyes are tense and alert	Slightly tense mouth and draw back may open mouth
Disgust	Lower the eyebrows	Push up lids without tense	Lips are curled and often asymmetrical
Surprise	Raise eyebrows Horizontal wrinkles	Drawn lower eyelid raise upper eyelid	Drop jaw, open mouth No tension or stretching of the mouth
Anger	Lower and draw together eyebrows vertical wrinkles between eyebrows	Eyes have a hard stare tense upper and lower lids	Moth firmly pressed Nostrils may be dilated

- When a sad expression occurs, the eyebrows will bend together and upward a bit at the inner parts. The mouth will extend. At the same time, the upper lids may drop down and lower lids may rise slightly.

- The facial moving features of the fear expression mainly occur at the eye and mouth portion. The eyebrows may raise and pull together. The eyes will become tense and alert. The mouth will also tend to be tense and may draw back and open.
- When a person is disgusted about something, the lips will be curled and often asymmetrical.
- The surprise expression has the most widely spread features. The whole eyebrows will bend upward and horizontal wrinkles may occur as a result of the eyebrow raise. The eyelids will move oppositely and the eyes will be open. Jaw will drop and mouth may open largely.
- When a person is in anger, the eyebrows are lowered and drawn together. Vertical wrinkles may appear between eyebrows. The eyes have a hard stare and both lids are tense. The mouth may be firmly pressed [1].

4.3 Visual Dataset

In our experiments, we used the visual affective database eNTERFACE'05 [12] database. It contains the six archetypal emotions: happiness (ha), sadness (sa), surprise (su), anger (an), disgust (di) and fear (fe). 42 subjects were asked to react to six different situations. The subjects were given five different answers to react to these situations. However, they were not given any instruction on how to express their emotions. Two human experts judged whether the reaction expressed the emotion in an unambiguous manner. If not, it was discarded. The database is collected in English language. Among the 42 subjects, 81% were men and remaining 19% were women. 31% of the total set wore glasses, while 17% of the subjects had a beard. The database is captured in a controlled recording environment.

Self-creatable database has captured in a general environment. In proper manner, it is a effective to capture frame. A special variance to use as audio has synchronized to visual environment. It is efficient and effective as cost.

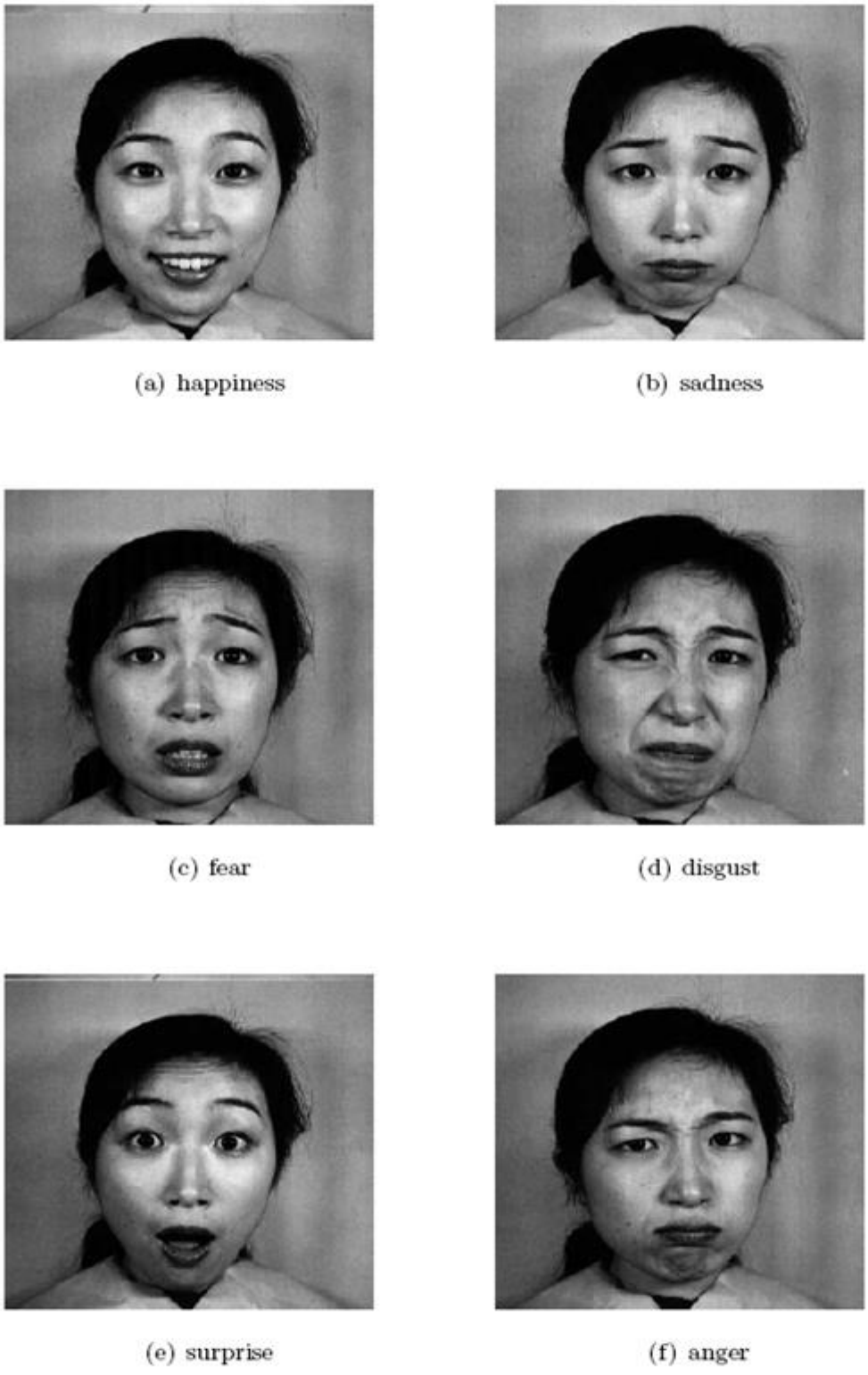


Figure 4.3: Six Universal Facial Expressions [8]

CHAPTER 5

RESULTS AND DISCUSSION

CHAPTER 5

RESULTS AND DISCUSSION

5.1 Introduction

In our experiments, we perform binary classification using Support Vector Machines (SVMs) with linear kernel and default parameters available in **MATLAB** implementation. We have 15 binary classification tasks corresponding to every possible pair of six expression classes available in the database. This is to emphasize the importance of bimodal data association in facial expression recognition using visual sequence data. Also, binary classification analysis helps us gain better in-sight on, specifically, the impact of our proposed frame-work and generally, the inherent confusion between two classes as discussed in the following section.

5.2 Result and Analysis

In this section, we present results for two classification tasks: the first one involves binary-class classification experiments and the second involves multi-class classification experiments. While, the purpose of binary classification task is to bring forth the importance of bimodal data association in facial expression recognition using visual sequence data. Also, binary classification analysis helps us gain better insight on, specifically, the impact of our proposed framework and generally, the inherent confusion between two classes. It is also worthy to note that many multi-class classification strategies inherently involve multiple binary classification and their performance is often ignored from discussion.

We perform binary classification using Support Vector Machine (SVM) with linear kernel and default parameters available in Matlab implementation. We have 15 binary classification tasks corresponding to every possible pair of six expression classes.

For subject dependent analysis, we utilize 15 fold cross validation strategy. That is the database, we are choose effective frame divided into 15 folds in stratified manner so that they contain approximately the same proportions of labels as the original database. The system is trained on 14 folds and tested on the left out fold. This is repeated 15 times each time leaving out a different fold. In the end, we obtain classification accuracy. We repeated the above procedure

15 times generating 10 accuracy figures for each of binary classification task. Mean accuracy is reported in Table 5.1. For subject independent analysis, we employ Leave-One-Subject-Out (LOSO) cross validation strategy. That is the system is trained using the data associated with all the subjects but one and tested on the left out subject. This is repeated until every subject is kept as test subject.

Firstly, it can be observed from Table 5.1 that the use of single image representation can provide high recognition accuracy. The best accuracy is obtained for the Happy/Anger binary classification with over 95% for randomized 15 folds cross validation. As expected, subject independent results show lower accuracy. Also, certain classes are more confusing in visual domain like the Sad/Fear or Surprise/Fear with recognition accuracy below 78%. It is important to point out, though, that we have not used any tuning of SVM parameters nor have we used any feature selection technique which often improves the performance greatly. Our focus is to compare the usefulness of auditory cross-modal feedback for frame selection which is also evident from the results.

Table 5.1 shows slight improvement on overall average performance by exploiting audio information. While the best improvement of 10% is obtained for binary classification task of Surprise/Fear in subject independent analysis (Table 5.1.ii). A closer look on the results suggests that emotion classes Fear and Happy have shown the most improvements. On the other hand, emotion classes Disgust and Sad may have not been benefited and even showing opposite trend in some cases. This can be attributed to our rule based weight assignment for these emotion classes. Particularly, for Sad class having low arousal profile, region corresponding to high intensity and pitch may not provide representative frames. This encourages us to learn such bimodal association automatically from audio visual data.

Another important performance metric is the computation cost. Notice that audio assisted approach utilizes maximum of 4 x 200 ms = 800ms worth of visual data corresponding to the four segments as described in Section 3.3 while using all the frames on an average requires 1 sec worth of visual frame processing. Hence using cross-modal information improved the visual computation cost by factor of 3.

Table 5.1 determines as Classification accuracy for the possible 15 different combinations of the binary classification tasks over six basic emotions: happy (ha), Sad (sa), Surprise (Su), Fear (Fe), Anger (An) and Disgust (Di). (i) Randomized 15 fold cross validation (ii) leave-one-subject-out cross validation. Note that the computation cost associated with visual processing of weighted-mean image (WMI) is at least one third than that of mean image (MI) method.

Table 5.1: Classification accuracy for the possible 15 different combination of the binary classification task over six basic emotions. These result analyses occur as matlab analysis.

Method	*Ha/Sa*	*Ha/Su*	*Ha/Fe*	*Ha/An*	*Ha/Di*	*Sa/ Su*	*Sa/Fe*
MI	93.70	88.32	90.02	95.80	93.50	81.20	65.50
WMI	92.58	92.96	90.62	95.85	93.69	78.00	73.50

Sa/An	*Sa/Di*	*Sa/Fe*	*Sa/An*	*Su/Di*	*Fe/An*	*Fe/Di*	*An/Di*
82.25	90.70	74.60	82.80	92.10	82.80	84.64	89.76
82.25	90.00	79.30	81.40	93.02	82.55	88.15	88.00
Average Accuracy (%)- MI: 85.84 and WMI: 86.79							

(i)

Method	*Ha/Sa*	*Ha/Su*	*Ha/Fe*	*Ha/An*	*Ha/Di*	*Sa/ Su*	*Sa/Fe*
MI	85.60	83.00	80.00	87.20	85.20	70.50	57.00
WMI	84.80	86.40	80.80	87.60	81.20	68.00	60.80

Sa/An	*Sa/Di*	*Sa/Fe*	*Sa/An*	*Su/Di*	*Fe/An*	*Fe/Di*	*An/Di*
69.20	86.40	61.20	72.00	88.80	71.20	81.60	88.40
68.30	86.00	71.20	71.20	87.60	74.40	80.00	84.40
Average Accuracy (%)-MI: 77.82 and WMI: 78.20							

(ii)

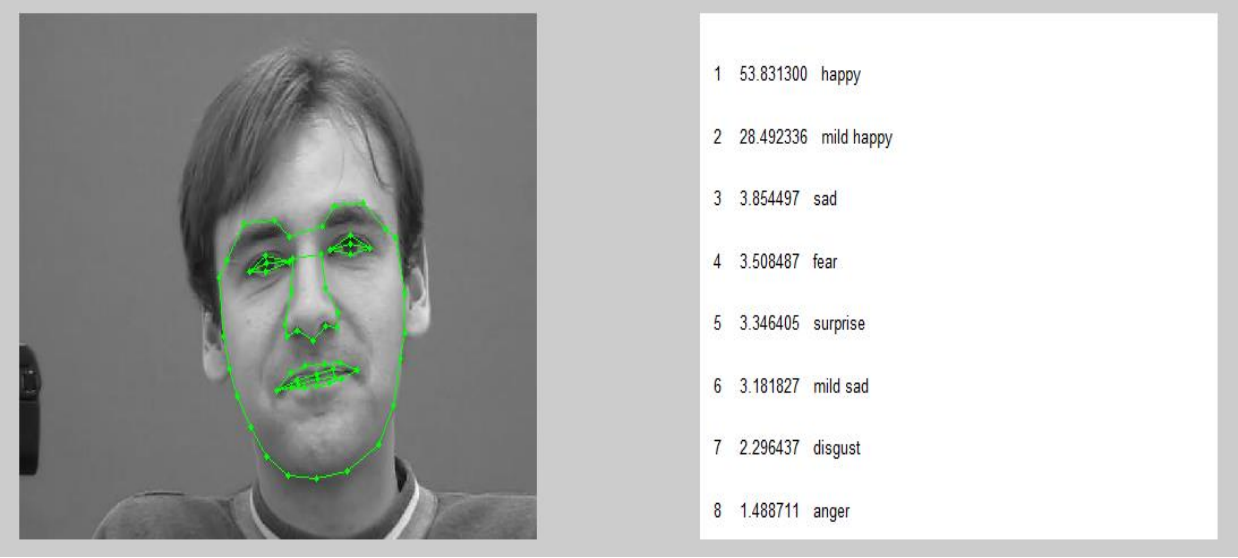


Figure 5.1: Image derived for the expression class of Ha/Mi-Ha

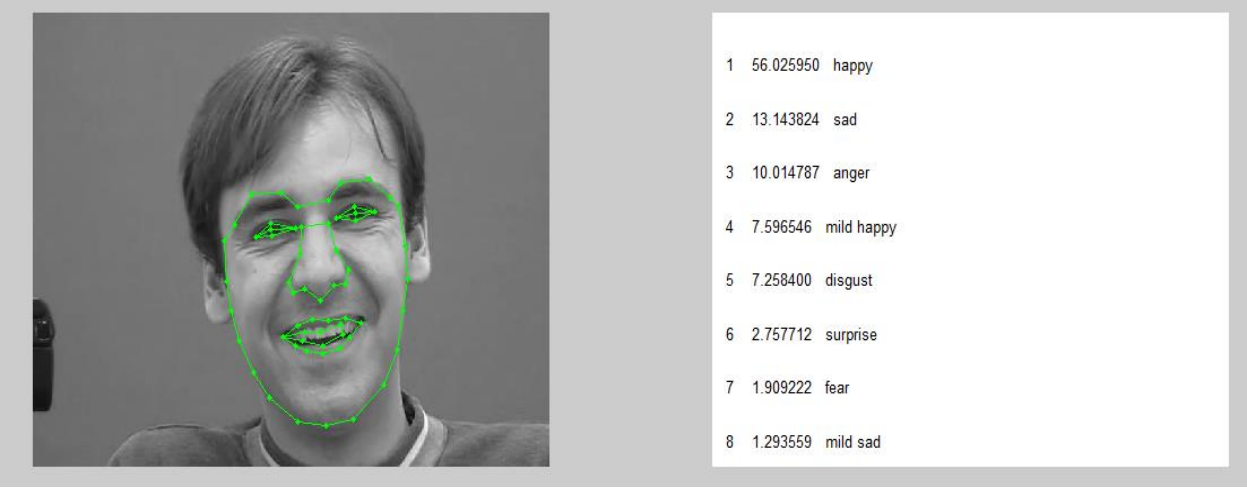


Figure 5.2: Image derived for the expression class of Ha/Sad

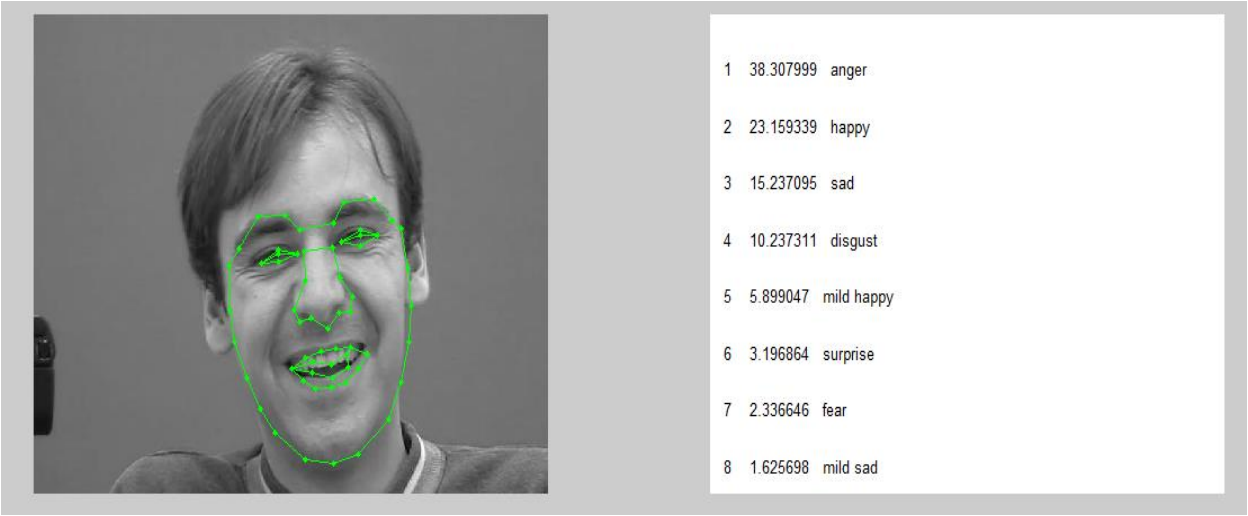


Figure 5.3: Image derived for the expression class of An/Ha

Figure 5.4: Image derived for the expression class of Ha/Mi-Sad

Figure 5.5: Image derived for the expression class of Ha/Di

Figure 5.6: Image derived for the expression class of An/Sad

Figure 5.7: Image derived for the expression class of Dis/Sad

An expression of facial recognition emotions of percentage level showed as figure 5.1, 5.2, and 5.7. These results of facial expression determine with visual setup. These module designs of fully automatic visual affect recognition system.

5.3 Comparison with other techniques results

Table 5.2: Comparison with other technique results [28, 32]

	TECHNIQUES	MEAN RESULT (In Percentage)
1	Gabar Motion Energy	78.56
2	Emotion Avtar Image + LBP	77.38
3	Emotion Avtar Image +LPQ	83.78
4	CLM+LPQ+SVM (Present Work)	85.84

We perform binary classification using Support Vector Machines (SVMs) with linear kernel and default parameters available in **MATLAB** implementation. An expression of facial recognition emotions of percentage level had been showed. Synthesis and simulated results has shown with explanation. The comparison table shows efficient and better results.

CHAPTER 6

CONCLUSION AND FUTURE WORK

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Introduction

Automatic analysis of human affective behavior has been extensively studied in past several decades. Facial expression recognition systems, in particular, have matured to a level where automatic detection of small number of expressions in posed and controlled displays can be done with reasonably high accuracy. Detecting these expressions in less constrained settings during spontaneous behavior, however, is still a challenging problem. In recent years, increasing number of efforts has been made to collect spontaneous behavior data in multiple modalities. The research shift towards this direction suggests utilizing the multimodal data analysis approaches.

We presented a novel approach of summarizing emotional content of the video frames by a single image using modal data association. We then investigated two different rule based data association approach for face expression recognition task. Our results showed that use of video data could improve the performance in terms of computation cost (since in general visual processing is costlier than audio processing) as well as recognition accuracy. Unlike various data fusion strategies, our approach attempted to better represent signal at feature extraction level by weighting frames by it is importance based on cross-relevance feedback.

6.2 Limitations

1. One limitation of the current system is that it can detect only one front view face looking at the camera.
2. Limitation of computers to detect as human face expression.

3. Audio should be less effective to detect emotion comparison with face expression.
4. Environment should be effective of emotion data base.

6.3 Future scope

In future work, explore data driven approach to learn better and more realistic cross-modal relevance measure as opposed to simple uniform weights used in present study. We will also incorporate audio modality in classification module and examine the multi-class classification approach for the design of fully automatic visual affect recognition system.

REFERENCES

[1] Bassili,J, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face," *Journal of Personality Social Psychology*, vol. 37, pp. 2049–2059, 1979.

[2] Cohen.I, .N Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Comput. Vision Image Understand.*, vol. 91, pp. 160–187, Jul. 2003.

[3] Datcu D. and Rothkrantz L., "Semantic audio-visual data fusion for automatic emotion recognition," in *Proc. Euromedia'2008 Porto*, J. Tavares and R. N. Jorge, Eds., Ghent, Belgium, Apr. 2008, pp. 58–65, Eurosis.

[4] Hu .C, Chang .Y, Feris .R, and Turk .M, "Manifold based analysis of facial expression," in *Proc. Conf. Computer Vision and Pattern Recognition Workshop*, Jun. 2004, p. 81.

[5] Jiang .B, Valstar.M, and Pantic.M, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recognition Workshops*, Mar. 2011, pp. 314–321.

[6] Lucey.S, Ashraf .A.B, and Cohn J.F, "Investigating spontaneous facial action recognition through aam representations of the face," in *Face Recognition, Delac*, 2007, pp. 275–286.

[7] Lam .K.M and Yan.H Fast algorithm for locating head boundaries.Journal of Electronic Imaging, 03(04):351–359, October 1994.

[8] Lyons M.J, Akamatsu .S, Kamachi.M, and Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. of the Third IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 200–205, April 1998.

[9] Murphy.E-Chutorian and Trivedi .M.M, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 300–311, 2010.

[10] Macdonald .J and McGurk .H, "Visual influences on speech perception processes," *Attention, Percept., Psychophys.*, vol. 24, pp. 253–257, 1978.

[11] Munhall .K, Jones .J, D. E. Callan, T. Kuratate, and E. Vatikiotis- Bateson, "Visual prosody and speech intelligibility," *Psychol. Sci.*, vol. 15, no. 2, pp. 133–137, 2004.

[12] Martin. O, Kotsia .I, B. Macq, and Pitas .I, "The enterface'05 audiovisual emotion database," in *Proc. 22nd Int. Conf. Data Engineering Workshops*, 2006, p. 8, IEEE Computer Society.

[13] Nechyba. M. C, Brandy.L, and Schneiderman .H. Lecture Notes in Computer Science, volume 4625/2009, chapter PittPatt Face Detection and Tracking for the CLEAR 2007 Evaluation, pages 126–137. Springer Berlin / Heidelberg, 2009.

[14] Neti.C, Potamianos.G, J. Luettin, I.Matthews, H.Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," in *Final Workshop 2000 Report.*, Baltimore, MD, USA, 2000, The Johns Hopkins Univ., Center for Language and Speech Processing.

[15] Ojala.T, Pietikäinen .M, and Mäenpää.M, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[16] Ojansivu.V and Heikkilä .J, "Blur insensitive texture classification using local phase quantization," in *Image and Signal Processing*, A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, Eds., 2008, vol. 5099, pp. 236–243.

[17] Picard. R. W, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[18] Paul .B, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Inst. Phonet. Sci. 17*, 1993, pp. 97–110.

[19] Sebe.N, Lew . M., Cohen I., Sun, T. Gevers. Y., and T. Huang, "Authentic facial expression analysis," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, May 2004, pp. 517–522.

[20] Song. M., Chen. C., and You. M., "Audio-visual based emotion recognition using tripled hidden Markov model," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, May 2004, vol. 5, pp. 877–880.

[21] Tawari.A. and Trivedi. M. M., "Speech emotion analysis: Exploring the role of context," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 502–509, Oct. 2010.

[22] Tian.Y., Kanade.T, and Cohn.J, "Facial expression analysis," in*Handbook of Face Recognition*, S. Li and A. Jain, Eds. New York, NY, USA: Springer, 2005, pp. 247–276.

[23] Tawari.A. and Trivedi. M. M., "Audio-visual data association for face expression analysis," in *Proc. Int. Conf. Pattern Recognition*, 2012.

[24] Tawari.A. and Trivedi. M. M., "Speech emotion analysis in noisy real world environment," in *Proc. Int. Conf. Pattern Recognition*, 2010.

[25] Tawari.A. and Trivedi. M. M., "Speech based emotion classification framework for driver assistance system," in *Proc. Intelligent Vehicles Symp. (IV)*, 2010, pp. 174–178.

[26] Mansoorizadeh.M. and Moghaddam.N. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," *Multimedia Tools Applicat.*, vol. 49, pp. 277–297, 2010.

[27] Valstar.M, Jiang.B, Mehu,M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition, Workshop on Facial Expression Recognition and Analysis Challenge*, 2011.

[28] Wu.T, Bartlett.M, and Movellan.J, "Facial expression recognition using Gabor motion energy filters," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2010, pp. 42–47.

[29] Wang.Y and Guan.L, "Recognizing human emotional state from audiovisual signals," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 936–946, Aug. 2008.

[30] Witten .I. H and Frank.E., *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*, ser. Morgan Kaufmann Series In Data Management Systems, 2Nd ed. SanMateo, CA, USA: Morgan Kaufmann, Jun. 2005.

[31] Yang. M.-H., Kriegman.D, and Ahuja.N. Detecting faces in images: A survey. IEEE Transactions on Pattern Analysis and Machine Intel- ligence, 24(1):34–58, January 2002.

[32] Yang.S and Bhanu.B, "Facial expression recognition using emotion avatar image," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recognition Workshops*, Mar. 2011, pp. 866–871.

[33] Zeng.Z, Pantic.M, Roisman.G, and Huang.T, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *PAMI*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[34] Zeng.Z, Tu.J., Pianfetti.B., Liu.M , Zhang.T, Zhang, Huang, and S. Levinson, "Audio-visual affect recognition through multi-stream fused hmm for hci," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, Jun. 2005, vol. 2, pp. 967–972.

APPENDIX

METLAB PROGRAMMING

DEMO FILE

```
close all

load('Model.mat')
% SVM C parameter
options.SvmC=1e-3;
% Size of canvas size for patch generation
options.CanvasSize=[400 400];
% Show debug images
options.Verbose=false;
% Patch size
options.PatchSize = [32 32];


% Number of search iterations
options.Iterations=25;
% Search area
options.SearchArea = [16 16];     % 1/2 patch size
% Shape constraints weight
options.ShapeConstraintsW = 0.005;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Set debug options
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% show image and feature points
options.dbgShowFeaturePts = false;
% show Patch of each image
options.dbgShowPatches = false;
% write training patches to files
options.dbgOutputPatches = 0;
% write search results to AVI file
options.WriteAvi = 1;


display=[100 100 800 400];
handle=figure('name','graph', 'position', display);
```

```
expr=strvcat('mild happy','happy','mild sad','sad','anger','disgust','surprise','fear');

%%%%%%%%%%%%%%%%%%%%
% Open video file
%%%%%%%%%%%%%%%%%%%%
k=0;
try

      options.AviFile = close(options.AviFile);

catch
   'nothing to do'
end
if(options.WriteAvi)
   AviFileName = sprintf('Demo_%d.avi',k);

   options.AviFile = avifile(sprintf('Demo_%d.avi',k));

end


kk=0;

load('svm_exp','model','MI_feature');

expression=1;

TestFiles=dir('images\*.jpg')
NumTestFiles = size(TestFiles, 1);

for kk = 1:NumTestFiles
   ImgFileName = TestFiles(kk, :).name;
   ImgFileName
   Itest=(im2double(imread(['images\' ImgFileName])));
   GrayITest = double(rgb2gray(Itest));


   if kk==1



      out=face_crop1(GrayITest);
      x=(out(1,1)+out(2,1))/2;
      y=(out(3,1)+out(4,1))/2;
```

```
    theta = -pi/20;
    %      width = 270;
    %      height = 270;
    width = abs(out(1,1)-out(2,1));
    height = abs(out(3,1)-out(4,1));
    %%%%%%%%%%%%%%%%%%%%%%%%
    % Make initial guess
    %%%%%%%%%%%%%%%%%%%%%%%%

    Si = MakeInitialGuess(Model, GrayITest, x, y, theta, width, height, options);

    end
hfig = subplot(1,2,1);
imshow(GrayITest);

hold on;
Draw_FS(Si.XY(1:2:end), Si.XY(2:2:end), 'r');
xlim([1 700]);
ylim([1 550]);


if(options.WriteAvi)
    frame = getframe(handle);
    try
        options.AviFile = addframe(options.AviFile, frame);
    catch
        kk
    end
    % options.AviFile = addframe(options.AviFile, frame);
end




for iter = 1:options.Iterations
    Si = Search(Model, GrayITest, Si, options);
    hfig = subplot(1,2,1);

    imshow(GrayITest);
    Draw_FS(Si.XY(1:2:end), Si.XY(2:2:end),'g');
    xlim([1 700]);
    ylim([1 550]);

    if(options.WriteAvi)
        frame = getframe(handle);
        try
            options.AviFile = addframe(options.AviFile, frame);
```

```
        catch
            kk
        end
        % options.AviFile = addframe(options.AviFile, frame);
        % options.AviFile = addframe(options.AviFile, frame);
    end


end

x=Si.XY(1:2:end);
y=Si.XY(2:2:end);
width=max(x)-min(x);
for jx=1:length(x)-1
    test_dataset(1,2*jx-1)=( (x(jx)-x(length(x))) )/width;
    test_dataset(1,2*jx)=( (y(jx)-y(length(x))) )/width;

end
dataset=[];

kx=0;
for jx=1:length(MI_feature)
    if MI_feature(jx)==1
        kx=kx+1;

        dataset(1,kx)=test_dataset(1,jx);
    end
end



[predict_label, accuracy, dec_values] = svmpredict(expression, dataset, model,'-b 1');

subplot(1,2,2);imshow(ones(550,700).*255)
xlim([1 700]);
ylim([1 550]);

[dec_values,idx]=sort(dec_values,'descend');
for iy=1:length(dec_values)
    yy=(iy)*65
    str=sprintf('%d    %2.6f   %s',iy,dec_values(iy)*100,expr(idx(iy),:))
    text(20,yy,str);
    xlim([1 700]);
    ylim([1 550]);
    hold on;
end
```

```
    hold off;
    subplot(1,2,1);
    expression=predict_label;
    if mod(kk,500)==0


        if(options.WriteAvi)
            options.AviFile = close(options.AviFile);
            k=k+1;
            options.AviFile = avifile(sprintf('Demo_%d.avi',k));

        end

    end

    options.Iterations=3;
end

if(options.WriteAvi)
    options.AviFile = close(options.AviFile);
end
```

SVM TRAINING

```
function training_dataset()
% SVM C parameter
options.SvmC=5e-3;
% Size of canvas size for patch generation
options.CanvasSize=[256 256];
% Show debug images
options.Verbose=false;
% Patch size
options.PatchSize = [32 32];
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% Set search options
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Number of search iterations
options.Iterations=20;
% Search area
options.SearchArea = [16 16];     % 1/2 patch size
% Shape constraints weight
options.ShapeConstraintsW = 0.005;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Set debug options
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% show image and feature points
options.dbgShowFeaturePts = false;
% show Patch of each image
options.dbgShowPatches = false;
% write training patches to files
```

```
options.dbgOutputPatches = 0;
% write search results to AVI file
options.WriteAvi = 0;
UseExistingModel = 0;
options.BuildShapeModelOnly = 0;
fid=fopen('aaa.txt','r');
a=fgets(fid);
k=0;
while a~=-1
   b=sscanf(a,'%d');
   for i=b(1):b(2)
      if i<10
         im_name=sprintf('franck_0000%d.jpg',i);
      else
         if i<100
            im_name=sprintf('franck_000%d.jpg',i);
         else
            if i<1000
               im_name=sprintf('franck_00%d.jpg',i);
            else
               im_name=sprintf('franck_0%d.jpg',i);
            end
         end
      end

      k=k+1
      TrainData(k).im_name=im_name;
```

```
        TrainData(k).exp=b(3);
    end
    a=fgets(fid);
end
fclose(fid);
TrainData=Load_pts(TrainData);
TrainingData = Preprocess_D(TrainData, options);
save('TrainingData','TrainingData');
```

FACE CROP PROGRAM

```
function out=face_crop1(x)


x=double(x);%make sure the input is double format
%[output,count,m,svec]=facefind(x);%full scan
[output,count,m,svec]=facefind(x,50,size(x,2),1,1,5);

area=abs(output(3,:)-output(4,:)).*abs(output(1,:)-output(2,:));
j=find(area>=max(max(area)),1,'first');
out(:,1)=[output(1,j),output(2,j),output(3,j),output(4,j)];
```

MAKE INITIAL GUESS

function Si = MakeInitialGuess(Model, image, x, y, theta, width, height, options)

ShapeModel = Model.ShapeModel;

MeanShape = ShapeModel.MeanShape;
NumPts = ShapeModel.NumPts;

```
MeanX = MeanShape(1:2:end);
MeanY = MeanShape(2:2:end);
MeanX = MeanX - mean(MeanX);
MeanY = MeanY - mean(MeanY);


ShapeW = max(MeanX) - min(MeanX);
ShapeH = max(MeanY) - min(MeanY);

%%%%%%%%%%%%%%%%%%%%%%%%%%
% calculate scale factor
%%%%%%%%%%%%%%%%%%%%%%%%%%
scf = width/ShapeW;
if(ShapeH*scf>height)
    scf = height/ShapeH;
end


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Calculate template x,y.
% This is not intended to be accurate,
% just barely enough for initial guess.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
TransM = [cos(theta) sin(theta) 0;
        -sin(theta) cos(theta) 0;
        0           0          1];
tXY = scf*TransM * [MeanX'; MeanY'; ones(1, size(MeanY, 1))] + repmat([x y
1]', 1, NumPts);


%%%%%%%%%%%%%%%%%%%%%%%%
% assemble initals
%%%%%%%%%%%%%%%%%%%%%%%%

Si = struct;
Si.XY = zeros(NumPts*2, 1);
Si.XY(1:2:end) = tXY(1, :)';
Si.XY(2:2:end) = tXY(2, :)';
```

CROP IMAGE BY SVM

```
function [Si CropImage] = CropImageSVM(Model, Image, Si, options)


ShapeModel = Model.ShapeModel;
PatchModel = Model.PatchModel;


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 1. Align to mean shape using procrustes analysis
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
MeanShape = ShapeModel.MeanShape;
NumPts = ShapeModel.NumPts;

MeanX = MeanShape(1:2:end);
MeanY = MeanShape(2:2:end);



CurrX = Si.XY(1:2:end);
CurrY = Si.XY(2:2:end);

[d AlignXY tform] = procrustes([MeanX MeanY], [CurrX CurrY],
'Reflection',false);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 2. Calculate center and transform matrix
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Translatexy = -1/tform.b*tform.c*tform.T';
Translatexy = Translatexy(1, :);

transM = [1/tform.b*tform.T Translatexy'];

Si.TransM = transM;
Si.AlignXY = zeros(NumPts*2, 1);
Si.AlignXY(1:2:end) = AlignXY(:, 1);
Si.AlignXY(2:2:end) = AlignXY(:, 2);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 3. Calculate bounding rectangle
%    and crop images.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
searchW = options.SearchArea(1);
searchH = options.SearchArea(2);
patchW = PatchModel.PatchSize(1);
patchH = PatchModel.PatchSize(2);
CropW = searchW + patchW;
CropH = searchH + patchH;

CropImage = zeros(CropH, CropW, NumPts);
for i=1:NumPts
    xy = AlignXY(i, :);
```

```
    tminx = xy(1) - searchW/2 - patchW/2;
    tmaxx = tminx + CropW -1;
    tminy = xy(2) - searchH/2 - patchH/2;
    tmaxy = tminy + CropH -1;

    targetXY = [ tminx tmaxx tminx tmaxx;
                 tminy tminy tmaxy tmaxy;
               ];

    %%%%%%%%%%%%%%%%%%%%%%%%
    % Crop image
    %%%%%%%%%%%%%%%%%%%%%%%%
    CropImage(:, :, i) = Quad2Box(Image, targetXY, transM);

end
```

OPTIMIZATION BY QUAD METHOD

```
function SiNew = Optimize(ShapeModel, H, f, Si, options)

numPts = ShapeModel.NumPts;
mean_xy = ShapeModel.MeanShape;

aligned_xy = Si.AlignXY;


debug = 0;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Find sum(bj^2/lambda_j), should be simple...
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
EvalueMat = repmat(sqrt(ShapeModel.Evalues'), size(ShapeModel.Evectors, 1),
1);
BMat = ShapeModel.Evectors./EvalueMat;
Evecs = ShapeModel.Evectors;

w0 = options.SearchArea(1)/2;
h0 = options.SearchArea(2)/2;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% now do quadratic programming.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Word of caution:
%   below assumes w0 == h0...
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
lb = ones(numPts*2, 1);
ub = ones(numPts*2, 1)*w0*2;


sub = 4*ones(size(ShapeModel.Evalues, 1), 1);

a1 = options.ShapeConstraintsW;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% This is a little messy when using quadprog...
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
basexy = -w0 + aligned_xy - mean_xy;

i_ee = (eye(numPts*2) - Evecs*Evecs');
W = i_ee'*i_ee;

H_n = 2*(a1*W - H);
F_n = (2*a1*W*basexy - f);


% Weighted version... for future extension.
%H_n = 2*(a1*D*W*D - H);
%F_n = (2*a1*D*W*basexy - f);

qops = optimset('LargeScale', 'off', 'Display', 'off');
```

```
x = quadprog(H_n, F_n, [BMat'; -BMat'], [sub - BMat'*basexy; sub +
BMat'*basexy], [], [], lb, ub, [], qops);

debug = 0;
if debug

    % plug in values here
    tbw = [-1.79383896534101    -1.70327125644485   -0.404600622497097  -
2.34375039017130   -0.405900258889102  -0.181308749606067  0.804114996656464
-2.07126003375068];

    www = tbw.*sqrt(ShapeModel.Evalues')
    revecs = ShapeModel.Evectors*www';
    xsave = x;
    x = revecs - basexy;
end

if 0
    % for test...
    RR = -x'*H*x - f'*x;

    error = (x+basexy) - Evecs*(Evecs'*(x+basexy));
    error2 = a1*error'*error;

    Errors = [RR error2]
    b_over_lambda = (BMat'*(x+basexy))'

end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Adjust output x from
% quadratic programming.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
x = x + aligned_xy;

new_x = x(1:2:numPts*2)- w0;
new_y = x(2:2:numPts*2)- h0;

%%%%%%%%%%%%%%%%%%%%%%
% Align back to current
% image coordinate:
%%%%%%%%%%%%%%%%%%%%%%
tform = Si.TransM;
newxy = [new_x'; new_y'; ones(1, numPts)] ;
txy = tform*newxy;

SiNew = Si;
SiNew.XY(1:2:end) = txy(1, :)';
SiNew.XY(2:2:end) = txy(2, :)';
```

END