# FACIAL EXPRESSION IMITATION FOR HUMAN ROBOT INTERACTION

## CHEN WANG

*(B.Eng. Beijing University of Aeronautics and Astronautics,*

*Beijing, China)*

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER

ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2008

# Acknowledgements

First and foremost, I would like to take this opportunity to express my sincere gratitude to my supervisors, Professor Shuzhi Sam Ge and Chang Chieh Hang, for their inspiration, encouragement, patient guidance and invaluable advice, especially for their selflessly sharing their invaluable experiences and philosophies, through the process of completing the whole project.

I would also like to extend my appreciation to Ms Pan Yaozhang, Mr Yang Chenguang, Mr Yang Yong, Ms Ren Beibei, Mr Tao Peyyuen, Dr Fua Chengheng, Dr Guan Feng and Mr Hooman Aghaebrahimi Samani for their help and support.

I am very grateful to National University of Singapore for offering the research scholarship.

Finally, I would like to give my special thanks to my parents, Wang Chaozhi and Hao Jin, and all members of my family for their continuing support and encouragement during the past two years.

Wang Chen

June 2008

# Contents

# Contents

# Abstract

As social robots become more and more interactive and communicated, it is crucial that they can understand, perceive and imitate the human emotions appropriately in the social environment. We propose an interactive system consisting of two key components, facial expression recognition and robot imitation. Within the recent decade, facial expression recognition has become a hot topic. But the existing 3D face mesh for facial expression recognition is based on the assumption of linear mass-spring model which can not simulate the facial muscle movements effectively. Thus in the system,a nonlinear mass-spring model is employed to simulate twenty two facial muscles' tensions during facial expressions, and then the elastic forces of these tensions are grouped into a vector which is used as the input for facial expression recognition. The experimental results show that the nonlinear facial mass-spring model coupled with the MLPs classifier is effective to recognize the facial expressions. For the robot imitation, we introduce the mechanism of our robot on imitating the facial expressions. Experimental results of imitating facial expressions demonstrate that our robot can imitate six kinds of facial expressions effectively.

# List of Tables

# List of Figures

# Chapter 1

# Introduction

As robot and people begin to co-exist and cooperatively share a variety of tasks, "natural" human-robot interaction with an implicit communication channel and a degree of emotional intelligence is becoming increasingly important. For a robot to be emotionally intelligent it should clearly have a two-fold capability - the ability to understand human emotions and the ability to display its own emotion just like human beings (usually by using facial expressions). There has been a stunningly vast amount of improvement in the basic capabilities of robotic entities - robots are getting smarter, more mobile, more aesthetically appealing to the masses, and subsequently, more widely accepted in the modern society. The incursion of robots into our everyday lives is unavoidable, and in most cases becoming indispensable. This explosion of intelligence robot also poses challenging problems of detecting, recognizing and imitating human emotions. Thus there is a growing demand for new techniques to efficiently recognize human facial expressions and for advanced robots to imitate human facial expressions.

## 1.1 Background

In recent years there has been a growing interest in developing more intelligent interface between humans and robots, and improving all aspects of the interaction. The emerging field on multi-modal/media human robot interface (HRI) has attracted the attention of many researchers from several different scholastic tracks, i.e., computer science, engineering, psychology, and neuroscience[1]. The main characteristics of human communication are: multiplicity and multi-modality of communication channels. A channel is a communication medium while a modality is a sense used to perceive signals from the outside world. Examples of human communication channels are: auditory channel that carries speech, auditory channel that carries vocal intonation, visual channel that carries facial expressions, and visual channel that carries body movements. Facial expression analysis could bring facial expressions into man-machine interaction as a new modality. Facial expression analysis and recognition are essential for intelligent and natural HRI, which presents a significant challenge to the pattern analysis and human-robot interface research community. Facial expression recognition is a problem which must be overcome for the future prospective application such as: emotional interaction, interactive video, synthetic face animation, intelligent home robotics, 3D games and entertainment[2].

Facial expression plays an important role in our daily activities. The human face is a rich and powerful source which is full of communicative information about human behavior and emotion. The most expressive way that humans display emotions is through facial expressions. Facial expression includes a lot of information about human emotion. It can provide sensitive and meaningful cues about emotional response and plays a major role in human interaction and nonverbal communication[3]. Facial expression analysis originates from Darwin in the 19th century when he proposed the concept of universal facial expressions in Man and

Animals. According the psychological and neurophysiological studies, there are six basic emotions-happiness, sadness, fear, disgust, surprise, and anger. Each basic emotion is associated with one unique facial expression[4]. Research on facial expression recognition and analysis in robot has been a hot research topic in the affective science of robotics. A large number of methods have been developed for facial expression analysis. There are some key problems need to be solved: detecting a human face in an image, extracting the facial features and classifying the feature-based facial expressions into different categories.

For the robot to express a full range of emotions and to establish a meaningful communication with a human being, nonverbal communications such as body language and facial expressions is vital. The ability to mimic human body and facial expressions lays the foundation for establishing a meaningful nonverbal communication between humans and robots [5].

Successful research and development in the area of social robots has important implications in several aspects of human society [6]. Intelligent robots which are capable of participating in meaningful interactions with humans around them have great potential in the following applications:

- Companions. Social robots, equipped with high level artificial intelligence and adaptive behaviours, will act as capable companions to users from diverse age groups. For children, these social robots can provide valuable companionship and act as babysitters that help parents monitor their children. Such interactive toys also serve to spark off creativity and can be a great source of information (via content/ information delivery from internet information sources) for children, able to answer their questions intelligently. In the case of adults, these robots act as personal assistants that can help manage the appointments and work commitments of the working adult. For the elderly, these robots serve as companions, combating loneliness amongst the elderly,

which is currently a major cause of depression and suicide and is expected to become more severe in the coming years. In addition to fulfilling the role of an able companion, intelligent social robots can also act as a conduit for bridging the distance between users, where emotions and gestures can be transmitted and manifested on the social robots on either end with humanistic robots serving as realistic personifications of loved ones. Furthermore, persistent wireless connectivity to the world wide web (which is fast becoming a standard feature on even the most basic digital device) and being equipped with intelligent filtering and information recognition tools, the social robot can act as a valuable one-point information source, in addition to a remote personal assistant.

- Entertainment. These robots will serve as interactive guides, realistic actors for exhibits, and even competent service providers. Currently, robots have already been actively employed in entertainment venues and theme parks. However, the majority of these robots are still limited to simple tasks, scripted actions and responses, heavily user initiated interactions, and limited learning. The use of social robots, with high level artificial intelligence and adaptive behaviours, will bring the concept of entertainment robotics to a new level and greatly enhance the consumer's experience. For example, sociable robotic agents will play significant roles in museums as guides, leading visitors on tours around the museum, providing oral accounts and multimedia presentations related to the display pieces. Robotic and human guides can work in tandem, with the robot handling the repetitive and mentally exhaustive task of giving oral accounts of the exhibits and the answering of common questions from the visitors, reducing the workload of their human counterparts. While human guides will handle questions from visitors that are beyond the AI of the robotic guides. The immense knowledge capacity

of robots makes it a suitable candidate for providing the detailed and accurate information on the exhibits to visitors. In addition, the robot can be equipped with features not available to human guides such as visual displays and wireless connections.

- Education. Interactive and intelligent robots capable of participating actively in the educational process will stimulate creativity within the young minds of students. In addition, the robot will provide new and valuable tools for teachers in both classroom-based learning and excursions. The near limitless information that can be contained within a robot will complement the teacher's knowledge base. Inspiring creativity is a major consideration in the development of interactive edutainment robot. Current robot programs in the schools focus on the design and development of low level robots. Although this encourages creativity through active participation in the design process, the hardware restrictions of these low level developmental kits limits creative exploration. An alternative to these educational robotic systems is to provide an advanced robotic platform, incorporating a variety of sensor systems and actuators, with high level software developmental kits (SDK). The readily available array of sensor systems and easy usage through a high level SDKs provides flexibility in the design and developmental stage, allowing imagination and creativity to flow. This approach will motivate the students to become creative thinkers through providing hands on experience and active participation in robot design. In addition, the SDKs provided will help to maintain the students' interest in robotic design by providing fast results for their efforts compared with low level robotic design where the process can be tedious and bogged down by hardware technicalities. Apart from inspiring creativity and facilitating the teaching process, the interactive robots can trigger significant learning across broad educational themes that

extend well beyond science, technology, engineering and mathematics, and into the associated lifelong learning skills of problem-solving, collaboration and communication through team-based development projects using open ended architecture.

## 1.2 Motivation of Thesis

The objective of our research is to develop a video-based human robot interaction system consisting of human facial expression recognition and imitation. Most existing systems for human robot interaction, however, suffer the following shortcomings:

- Facial expression in a video is a dynamic process or expression sequence. Most of the current techniques adopt the facial texture or shape information for expression recognition [7], [8]. There are more information stored in the facial expression sequence compared to the facial shape information. Its temporal information can be divided into three discrete expression states in an expression sequence: the beginning, the peak, and the ending of the expression. But those techniques often ignore such temporal information.

- The existing 3D face mesh for facial expression recognition is based on the assumption of linear mass-spring model. As discussed in [9], the simple linear mass-spring models can not simulate the real issue muscles accurately. The facial muscle motivation is a nonlinear mass-spring model, and the facial feature is also controlled by the nonlinear spring motivation which can simulate the elastic dynamics of real facial skin.

- A facial expression consists of not only its temporal information, but also a great number of AU combinations and transient cues. The HMM can model

uncertainties and time series, but it lacks the ability to represent induced and nontransitive dependencies. Spatio-temporal approaches allow for facial expression dynamics modeling by considering facial features extracted from each frame of a facial expression video sequence.

## 1.3 Contributions

The main contributions of this thesis can be summarized as follows:

1. A nonlinear mass-spring model is implemented to describe the facial muscles' elasticity in facial expression recognition. We study facial muscles' temporal transition characteristics of different expressions and propose a novel feature to represent the facial expressions based on non-linear mass-spring model.

2. We build up a human-robot interactive system for recognizing and imitating human facial expressions by integrating our proposed feature. The experimental results showed that our proposed nonlinear facial mass-spring model coupled with the Multi-layer Perceptrons (MLPs) classifier is effective to recognize the facial expressions compared with the linear mass-spring model. A social robot was designed to make artificial facial expressions. Experimental results of facial expression generation demonstrated that our robot can imitate six types of facial expressions effectively.

## 1.4 Thesis Organization

The remainder of this paper is organized as follows:

In Chapter 2, a general framework for facial expression imitation system in human robot interaction is introduced. The methods of face detection, facial features

extraction and facial expression classification are discussed. Representative facial expression recognition system and interactive robot expression animation system are described finally.

In Chapter 3, the face detection and facial features extraction methods are discussed. Face detection can fix a range of interests, decrease the searching range and initial approximation area for the feature extraction. Vertical and horizontal projection methods are conducted to automatically detect and locate face area. And then facial features are extracted by using deformable templates to get precise positions.

In Chapter 4, we discuss the nonlinear mass-spring model which can be used to simulate the muscle's tension during the expression. It takes advantage of the optical ow method which tracks the feature points' movement information. For each expression we use the typical patterns of muscle actuation, as determined using our detailed physical analysis, to generate the typical pattern of motion energy associated with each facial expression.

In Chapter 5, we present how to classify the facial expressions and summarize the experimental results. Both integration-based approach and action units-based approach are discussed. Mlps are employed for static facial expression classification.

Chapter 6 describes the proposed human-robot interaction application.From its concept design, the robotic face's affective states are triggered by the emotion generator engine. It's facial features can give a vivid animation according to the tester's expression. This occurs as a response to its internal state representation, captured through multimodal interaction.

In Chapter 7, we give some conclusions and discuss our future work.

# Literature Review

This Chapter introduces a general facial expression framework, and then discusses each module in this framework, including face acquisition, feature extraction and representation, facial expression classification. Then we describe some state-of-the-art facial expression recognition systems. Some social interactive robots and their applications in the field of facial emotion expression imitation are also discussed. Finally, our system description and assumption are introduced.

## 2.1 A General Framework of Facial Expression Imitation System in Human Robot Interaction

There are two key components for most existing facial expression imitation systems. One is for facial expression recognition, and the other is for facial expression imitation.
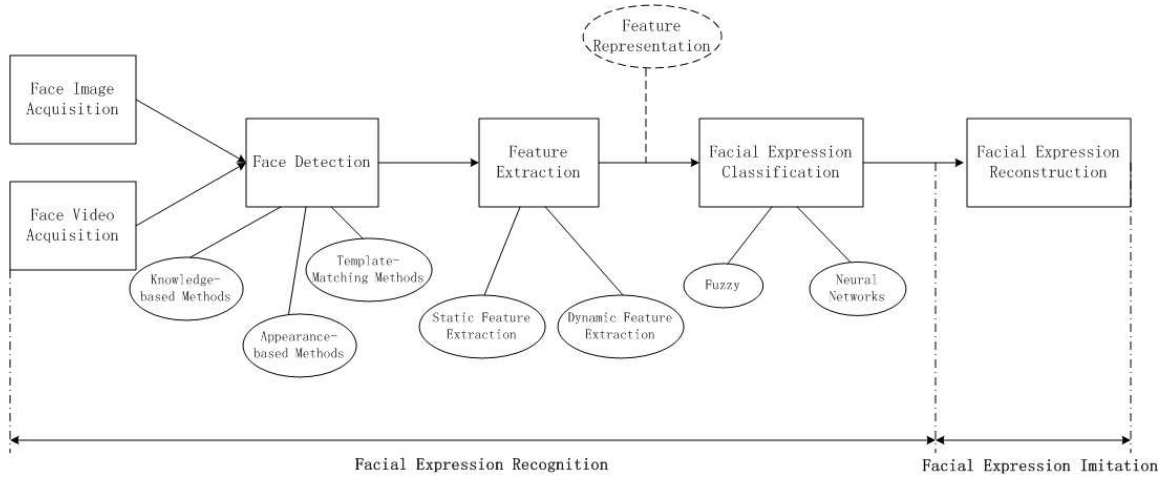
Figure 2.1: Robot imitates human facial expression.

As shown in Fig. 2.1, the recognition component is composed of four modules: face acquisition, facial feature extraction, facial feature representation and facial expression classification. Given a facial image, the face acquisition module is used to segment the face region in this image. Then the module of the facial feature extraction includes locating the positions and shapes of the eyebrows, eyes, nose, mouth, and extracting facial features in a still image of human face. The module of facial feature representation postprocesses the extracted facial features and preserve all the information for further classification. Finally based on the post-processed facial features, the module of facial expression classification is used to classify the given facial image into the predefined emotion class. In the reminder of this chapter, we will have a closer look at the individual module of this general framework. Finally, the module of artificial emotion generation can control a social robot to imitate the facial expression in response of the user's expression.

## 2.2 Face Acquisition

An ideal module of face acquisition should feature an automatic face detector that allows to locate faces in complex scenes with cluttered backgrounds [10]. Certain face analysis methods need the exact position of the face in order to extract facial features of interest while others work, if only the coarse location of the face is available. This is the case with e.g. active appearance models [11]. Hong et al. [12] used the PersonSpotter system by Steffens et al. [13] in order to perform realtime tracking of faces. The exact face dimensions were then obtained by 0tting a labeled graph onto the bounding box containing the face previously detected by the PersonSpotter system. Essa and Pentland [14] located faces by using the view-based and modular eigenspace method of Pentland et al. [15]. To As far as we know, face analysis is still complicated due to face appearance changes caused by pose variations and illumination changes. It might therefore be a good idea to normalize acquired faces prior to their analysis:

1. Pose: The appearance off facial expressions depends on the angle and distance at which a given face is being observed. Pose variations occur due to scale changes as well as in-plane and out-of-plane rotations off aces. Especially out-of-plane rotated faces are difficult to handle, as perceived facial expression are distorted in comparison to frontal face displays or may even become partly invisible. Limited out-of-plane rotations can be addressed by warping techniques, where the center positions of distinctive facial features such as the eyes, nose and mouth serve as reference points in order to normalize test faces according to some generic face models e.g. see Ref. [14]. Scale changes off aces may be tackled by scanning images at several resolutions in order to determine the size of present faces, which can then be normalized accordingly

[16].

2. Illumination: A common approach for reducing lighting variations is to filter the input image with Gabor wavelets or model facial colour and identity with Gaussian mixtures see Ref. [17] The problem of partly lightened faces is still an open research problem which is very difficult to solve.

## 2.3 Feature extraction and Representation

A facial expression involves simultaneous changes of facial features on multiple facial regions. Facial expression states vary over time in an image sequence and so do the facial visual cues. For a particular facial activity, there is a subset of facial features that is the most informative and maximally reduces the ambiguity of classification. In general, there are three kinds of approaches to extract facial features.

### 2.3.1 Deformation based approaches

Deformation of facial features are characterized by shape and texture changes and lead to high spatial gradients that are good indicators for facial actions and may be analyzed either in the image or the spatial frequency domain. The latter can be computed by high-pass gradient or Gabor wavelet-based filters, which closely model the receptive field properties of cells in the primary visual cortex [18, 19]. They allow to detect line endings and edge borders over multiple scales and with different orientations. These features reveal much about facial expressions, as both transient and intransient facial features often give raise to a contrast change with

regard to the ambient facial tissue. Gabor filters remove most of the variability in images that occur due to lighting changes. They have shown to perform well for the task of facial expression analysis and were used in image-based approaches [20, 21, 22] as well as in combination with labeled graphs [12, 23, 24].

### 2.3.2 Muscle based approaches

Muscle-based frameworks attempt to interfere muscle activities from visual information. This may be achieved e.g. by using 3D muscle models to describe muscle actions [25, 26]. Modeled facial motion can hereby be restricted to muscle activations that are allowed by the muscle framework, giving control over possible muscle contractions, relaxation and orientation properties. However, the musculature of the face is complex, 3D information is not readily present and muscle motion is not directly observable. For example, there are at least 13 groups of muscles involved in the lip movements alone [27]. Mase and Pentland [28] did not use complex 3D models to determine muscle activities. Instead they translated 2D motion in predefined windows directly into a coarse estimate of muscle activity. As discussed in [29], the actual facial expressions can be generated by the dynamics of the facial muscles which are under the skin.

### 2.3.3 Motion based approaches

Among the motion extraction methods that have been used for the task of facial expression analysis we find feature point tracking and difference-images.

1. Feature point tracking: Here, motion estimates are obtained only for a selected set of prominent features such as intransient facial features [30, 31, 32].

In order to reduce the risk of tracking loss, feature points are placed into areas of high contrast, preferably around intransient facial features as is illustrated on the right-hand side of Fig. 6. Hence, the movement and deformation of the latter can be measured by tracking the displacement of the corresponding feature points. Motion analysis is directed towards objects of interest and therefore does not have to be computed for extraneous background patterns. However, as facial motion is extracted only at selected feature point locations, other facial activities are ignored altogether. The automatic initialization of feature points is difficult and was often done manually. Otsuka and Ohya [33] presented a feature point tracking approach, where feature points are not selected by human expertise, but chosen automatically in the first frame of a given facial expression sequence. This is achieved by acquiring potential facial feature points from local extrema or saddle points of luminance distributions. Tian et al. [31] used different component models for the lips, eyes, brows as well as cheeks and employed feature point tracking to adapt the contours of these models according to the deformation of the underlying facial features. Finally, Rosenblum et al. [34] tracked rectangular, facial feature enclosing regions of interest with the aid of feature points.

Note that even though the tracking of feature points or markers allows to extract motion, often only relative feature point locations, i.e. deformation information was used for the analysis of facial expressions, e.g. in [35] or [31]. Yet another way of how to extract image motion are difference-images: Specifically for facial expression analysis, difference-images are mostly created by subtracting a given facial image from a previously registered reference image, containing a neutral face of the same subject. Compared with difference-images, feature point tracking approach could be more robust to the subtle changes of face positions. Thus we employ the feature tracking approach to extract facial features in our system.

## 2.4   The measurement of facial expression

Facial expressions are generated by contractions off facial muscles, which results in temporally deformed facial features such as eye lids, eye brows, nose, lips and skin texture, often revealed by wrinkles and bulges. Typical changes of muscular activities are brief, lasting for a few seconds, but rarely more than 5 s or less than 250 ms. We would like to accurately measure facial expressions and therefore need a useful terminology for their description. Of importance is the location off facial actions, their intensity as well as their dynamics. Facial expression intensities may be measured by determining either the geometric deformation of facial features or the density of wrinkles appearing in certain face regions. For example the degree of a smiling is communicated by the magnitude of cheek and lip corner raising as well as wrinkle displays. Since there are inter-personal variations with regard to the amplitudes off facial actions, it is difficult to determine absolute facial expression intensities, without referring to the neutral face of a given subject. Note that the intensity measurement of spontaneous facial expressions is more difficult in comparison to posed facial expressions, which are usually displayed with an exaggerated intensity and can thus be identi0ed more easily. Not only the nature of the deformation of facial features conveys meaning, but also the relative timing off facial actions as well as their temporal evolution. Static images do not clearly reveal subtle changes in faces and it is therefore essential to measure also the dynamics off facial expressions. Although the importance of correct timing is widely accepted, only a few studies have investigated this aspect systematically, mostly for smiles [36]. Facial expressions can be described with the aid of three temporal parameters: onset (attack), apex (sustain), o¡set (relaxation). These can be obtained from human coders, but often lack precision. Few studies relate to the problem of automatically computing the onset and offset off facial expressions, especially when

not relying on intruding approaches such as Facial EMG [37]. There are two main methodological approaches of how to measure the aforementioned three characteristics of facial expressions, namely message judgment based and sign vehicle-based approaches [38]. The former directly associate specific facial patterns with mental activities, while the latter represent facial actions in a coded way, prior to eventual interpretation attempts.

### 2.4.1 Judgment-based approaches

Judgment-based approaches are centered around the messages conveyed by facial expressions. When classifying facial expressions into a predefined number of emotion or mental activity categories, an agreement of a group of coders is taken as ground truth, usually by computing the average of the responses of either experts or non-experts. Most automatic facial expression analysis approaches found in the literature attempt to directly map facial expressions into one of the basic emotion classes introduced by Ekman and Friesen [39, 40].

### 2.4.2 Sign-based approaches

With sign vehicle-based approaches, facial motion and deformation are coded into visual classes. Facial actions are hereby abstracted and described by their location and intensity. Hence, a complete description framework would ideally contain all possible perceptible changes that may occur on a face. This is the goal of facial action coding system (FACS), which was developed by Ekman and Friesen [40] and has been considered as a foundation for describing facial expressions. It is appearance-based and thus does not convey any information about e.g. mental activities associated with expressions. FACS uses 44 action units (AUs) for the description off facial actions with regard to their location as well as their intensity,

the latter either with three or 0ve levels of magnitude. Individual expressions may be modeled by single action units or action unit combinations. Similar coding schemes are EMFACS [41], MAX [42] and AFFEX [43]. However, they are only directed towards emotions. Finally, the MPEG-4-SNHC [44] is a standard that encompasses analysis, coding [45] and animation off aces (talking heads) [46]. Instead of describing facial actions only with the aid of purely descriptive AUs, scores of sign-based approaches may be interpreted by employing facial expression dictionaries. Friesen and Ekman introduced such a dictionary for the FACS framework [47]. Ekman et al. [48] presented also a database called facial action coding system affect interpretation database (FACSAID), which allows to translate emotion related FACS scores into affective meanings. Emotion interpretations were provided by several experts, but only agreed affects were included in the database.

## 2.5   Facial Expression Classification

According to the psychological and neurophysiological studies, there are six basic emotions-happiness, sadness, fear, disgust, surprise, and anger as shown in Fig. 2.2. Each basic emotion is associated with one unique facial expression.

Feature classification is performed in the last stage of an automatic facial expression analysis system. This can be achieved by either attempting facial expression recognition using sign-based facial action coding schemes or interpretation in combination with judgment or sign/dictionary-based frameworks.

1. Hidden Markov models (HMM) are commonly used in the field of speech recognition, but are also useful for facial expression analysis as they allow to model the dynamics of facial actions. Several HMM-based classification approaches can be found in the literature [50, 33] and were mostly employed in

(a) happiness          (b) sadness          (c) fear

(d) disgust          (e) surprise          (f) anger
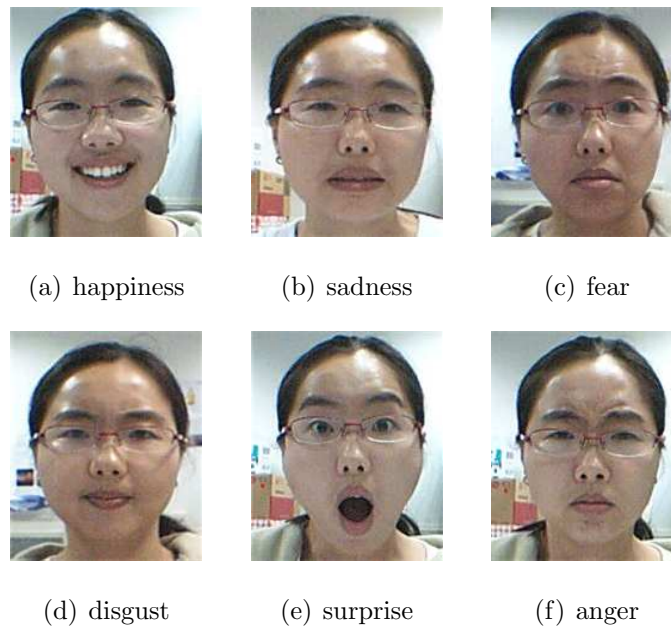
Figure 2.2: Six universal facial expressions [49].

conjunction with image motion extraction methods. Recurrent neural networks constitute an alternative to HMMs and were also used for the task of facial expression classification [51, 34]. Another way of taking temporal evolution of facial expression into account are so-called spatio-temporal motion-energy templates. Here, facial motion is represented in terms of 2D motion fields. The Euclidean distance between two templates can then be used to estimate the prevalent facial expression [14].

2. Neural networks were often used for facial expression classification [52, 20, 24, 53, 54]. They were either applied directly on face images [21] or combined with facial features extraction and representation methods such as PCA independent component analysis (ICA) or Gabor wavelet filters [22, 21]. The former are unsupervised statistical analysis methods that allow for a considerable dimensionality reduction, which both simplifies and enhances subsequent classification. These methods have been employed both in a holistic

manner [20, 55] or locally, using mosaic-like patches extracted from small facial regions [52, 22, 55]. Dailey and Cottrell [22] applied both local PCA and Gabor jets for the task of facial expression recognition and obtained quantitatively indistinguishable results for both representations. Unfortunately, neural networks are difficult to train if used for the classification of not only basic emotions, but unconstrained facial expressions. A problem is the great number of possible facial action combinations, about 7000 AU combinations have been identified within the FACS framework [38]. An alternative to classically trained neural networks constitute compiled, rule-based neural networks that were employed e.g. in [35].

In [56], the features used for NN can be either the geometric positions of a set of fiducial points on a face or a set of multiscale and multiorientation Gabor wavelet coefficients extracted from the facial image at the fiducial points. The recognition is performed by a two layer perceptron NN. The system developed is robust to face location changes and scale variations. Feature extraction and facial expression classification were performed using neuron groups, having as input a feature map and properly adjusting the weights of the neurons for correct classification. A method that performs facial expression recognition is presented in [57]. Face detection is performed using a Convolutional NN, while the classification is performed using a rule-based algorithm. Optical flow is used for facial region tracking and facial feature extraction in [58]. The facial features are inserted in a Radial Basis Function (RBF) NN architecture that performs classification. The Discrete Cosine Transform (DCT) is used in [59], over the entire face image as a feature detector. The classification is performed using a one-hidden layer feedforward NN.

The HMM can model uncertainties and time series, but it lacks the ability to represent induced and nontransitive dependencies. So NN is often employed in

most existing facial expression recognition systems based on (FACS).

# 2.6 State-of-the-art facial expression recognition systems

In this section, we have a closer look at a few representative facial expression analysis systems. First, we discuss deformation and motion-based feature extraction systems. Then we introduce hybrid facial expression analysis systems, which employ several image analysis methods that complete each other and thus allow for a better overall performance. Multi-modal frameworks on the other hand integrate other non-verbal communication channels for improving facial expression interpretation results.

## 2.6.1 Deformation extraction-based systems

Padgett et al. [60] presented an automatic facial expression interpretation system that was capable ofidentif ying six basic emotions. Facial data was extracted from $32 \times 32$ pixel blocks that were placed on the eyes as well as the mouth and projected onto the top 15 PCA eigenvectors of 900 random patches, which were extracted from training images. For classification, the normalized projections were fed into an ensemble of 11 neural networks. Their output was summed and normalized again by dividing the average outputs for each possible emotion across all networks by their respective deviation over the entire training set. The largest score for a particular input was considered to be the emotion found by the ensemble of networks. Altogether 97 images of six emotions from 6 males and 6 females were analyzed and

a 86% generalization performance was measured on novel face images. Lyons et al. Experiments were carried out on subsets of totally six different posed expressions and neutral faces of 9 Japanese female undergraduates. A generalization rate of 92% was obtained for the recognition of new expressions of known subjects and 75% for the recognition of facial expressions of novel expressers.

### 2.6.2    Motion extraction-based systems

Black and Yacoob [61] analyzed facial expressions with parameterized models for the mouth, the eyes and the eye brows and represented image flow with low-order polynomials. A concise description of facial motion was achieved with the aid of a small number of parameters from which they derived mid- and high-level description of facial actions. The latter considered also temporal consistency of the mid-level predicates in order to minimize the e7ects of noise and inaccuracies with regard to the motion and deformation of the models. Hence, each facial expression was modeled by registering the intensities of the mid-level parameters within temporal segments (beginning, apex, ending). Extensive experiments were carried out on 40 subjects in the laboratory with a 95% correct recognition rate and also with television and movie sequences resulting in a 80% correct recognition rate. The employed dynamic face model allowed not only to extract muscle actuations of observed facial expressions, but it was also possible to produce noise corrected 2D motion 0elds via the control-theoretic approach. The latter where then classified with motion energy templates in order to extract facial actions. Experiments were carried out on 52 frontal view image sequences with a correct recognition rate of 98% for both the muscle and the 2D motion energy models.

### 2.6.3 Hybrid systems

Hybrid facial expression analysis systems combine several facial expression analysis methods. This is most beneficial, if the individual estimators produce very di7erent error patterns. Bartlett et al. [55] proposed a system that integrates holistic difference-images motion extraction coupled with PCA, feature measurements along predefined intensity profiles for the estimation of wrinkles and holistic dense optical flow for whole-face motion extraction. These three methods were compared with regard to their contribution to the facial expressions recognition task. Bartlett et al. estimated that without feature measurement, there would have been a 40% decrease of the improvement gained by all methods combined. Faces were normalized by alignment through scaling, rotation and warping of aspect ratios. However, eye and mouth centers were located manually in the neutral face frame, each test sequence had to start with. Facial expression recognition was achieved with the aid of a feed-forward neural network, made up of 10 hidden and six output units. The input of the neural network consisted of 50 PCA component projections, five feature density measurements and six optical flow-based template matches. A winner takes it all (WTA) judgment approach was chosen to select the 0nal AU candidates. Initially, Bartlett et al.s hybrid facial expression analysis system was able to classify six upper FACS action units on a database containing 20 subjects, correctly recognizing 92% of the AU activations, but no AU intensities. Later it was extended to allow also for the classification of lower FACS action units and achieved a 96% accuracy for 12 lower and upper face actions [20, 55].

# 2.7 Emotion Recognition in Human-robot Interaction

## 2.7.1 Social interactive robot

In recent years, the robotics community has seen a gradual increase in social robots, that is, robots that exist primarily to interact with people. Therefore, many kinds of socially interactive robot operating as partners, peers or assistants, were invented. Different from traditional industrial robots, socially interactive robots need to exhibit a certain degree of adaptability and flexibility to drive the interaction with a wide range of humans. Socially interactive robots can have different shapes and functions, ranging from robots whose sole purpose and only task is to engage people in social interactions to robots that are engineered to adhere to social norms in order to fulfill a range of tasks in human-inhabited environments [62, 63].

Socially interactive robots are important for domains in which robots must exhibit peer-to-peer interaction skills, either because such skills are required for solving specific tasks, or because the primary function of the robot is to interact socially with people[64, 65].

The emotion exchanges and interaction is one of the most important and necessary characteristics of the social robotics, and also called the affective sciences. Affective science is the scientific study of emotion. An increasing interest in emotion can be seen in the behavioral, biological and social sciences. Research over the last two decades suggests that many phenomena, ranging from individual cognitive processing to social and collective behavior, cannot be understood without taking into account affective determinants (i.e. motives, attitudes, moods, and emotions).

The major challenge for this interdisciplinary domain is to integrate research focusing on the same phenomenon, emotion and similar affective processes, starting from different perspectives, theoretical backgrounds, and levels of analysis.

For a service robot to be more human friendly, an affective system is an essential part of the human-robot interaction (HRI), because emotions affect rational decision-making, perception, learning, and other cognitive functions of a human. According to the somatic marker hypothesis, the marker records emotional reaction to a situation [66]. We learn the markers throughout our lives and use them for our decision-making. Therefore, it is quite necessary for a believable robot to have an affective system such that it can synthesize and express emotions.

In recent years, affective techniques has increasingly been used in interface and robot design, primarily because of the recognition that people tend to treat computers as they treat other people [67]. Moreover, many studies have been performed to integrate emotions into products including electronic games, toys, and software agents[65].

For a robot to be emotionally intelligent it should clearly have a two-fold capability-the ability to display its own emotions just like human beings (usually by using facial expressions and speech[68]) and the ability to understand human emotions and motivations (also referred to as affective states).

## 2.7.2 Facial emotion expression as human being

Through facial expressions, robots can display their own emotion just like human beings. The expressive behavior of robotic faces is generally not life-like. This reflects limitations of mechatronic design and control. For example, transitions between expressions tend to be abrupt, occurring suddenly and rapidly, which rarely occurs in nature. The primary facial components used are mouth (lips), cheeks, eyes, eyebrows and forehead. Most robot faces express emotion in accordance with

Ekman and Frieser's FACS system [47, 40, 69].

There have been several attempts to build emotional robots such as Sony's Aibo [70], MIT's Kismet [71], and KAIST's AMI [72]. In Kismet, its affective system has a three dimensional affect space of valence, stance, and arousal and the appraisal of external stimuli is mapped to the space. Similarly, Aibo has its own affect space of seven emotions based on Takanishi's model [73] and generates appropriate emotional reactions to a situation. However, the affect space allows the robots to have only one emotion at a time, because the affect space has a competitive relationship among emotions. For example, Aibo always expresses only one affective state from among its seven emotions: happy, sadness, fear, disgust, surprise, angry and hungry.

Since the temporal lobe and the prefrontal cortex have undergone considerable development, human beings have several emotions simultaneously and express them in various ways. Furthermore, according to the studies of human social interactions, people feel more comfortable with a human-like agent. In [74], the authors propose a dynamic robot affective system inspired from both neuroscience and cognitive science such that it can have various emotional states at the same time and express those combined emotions just like humans do.

Instead of using mechanical actuation, another approach to facial expression is to rely on computer graphics and animation techniques. Valerie, for example, has a 3D rendered face of a woman based on Delsarte's code of facial expressions [75]. Because Valerie's face is graphically rendered, many degrees of freedom are available for generating expressions.

## 2.8   Challenges

It is important to note that the goal of tracking the dynamic information is primarily to estimate the changes of either skin surface on each facial muscle or motion energy converted from the muscular activations.

In this thesis, we are interested in how to apply dynamics of the facial muscles to perform the recognition of facial expressions, and build a dynamic physically-based expression recognition system. A human being can have several emotions and express them in various ways. The motion characteristics and elastic properties of real facial muscle have been ignored in "facial motion" tracking. In our work the skin model is constructed by using the nonlinear spring frames which can simulate the elastic dynamics of real facial skin. The facial expressions are synthesized by facial skin nodes driven by the muscle contraction [76]. When muscles contract, by solving the dynamic equation for feature skin node in the facial surface, we can observe the affective transformation on facial expressions.

## 2.9 System description

Our facial expression recognition research is conducted based on the following assumptions:

**Assumption 1.** *Using only vision camera, one can only detect and recognize the shown emotion that may or may not be the personal true emotions. It is assumed that the subject shows emotions through facial expressions as a mean to express emotion.*

**Assumption 2.** *Theories of psychology claim that there is a small set of basic expressions [40], even if it is not universally accepted. A recent cross-cultural study confirms that some emotions have a universal facial expression across the cultures*

*and the set proposed by Ekman [77] is a very good choice. Six basic emotions-happiness, sadness, fear, disgust, surprise, and anger are considered in our research. Each basic emotion is assumed associated with one unique facial expression for each person.*

**Assumption 3.** *There is only one face contained in the captured image. The face takes up a significant area in the image. The image resolution should be sufficiently large to facilitate feature extraction and tracking.*
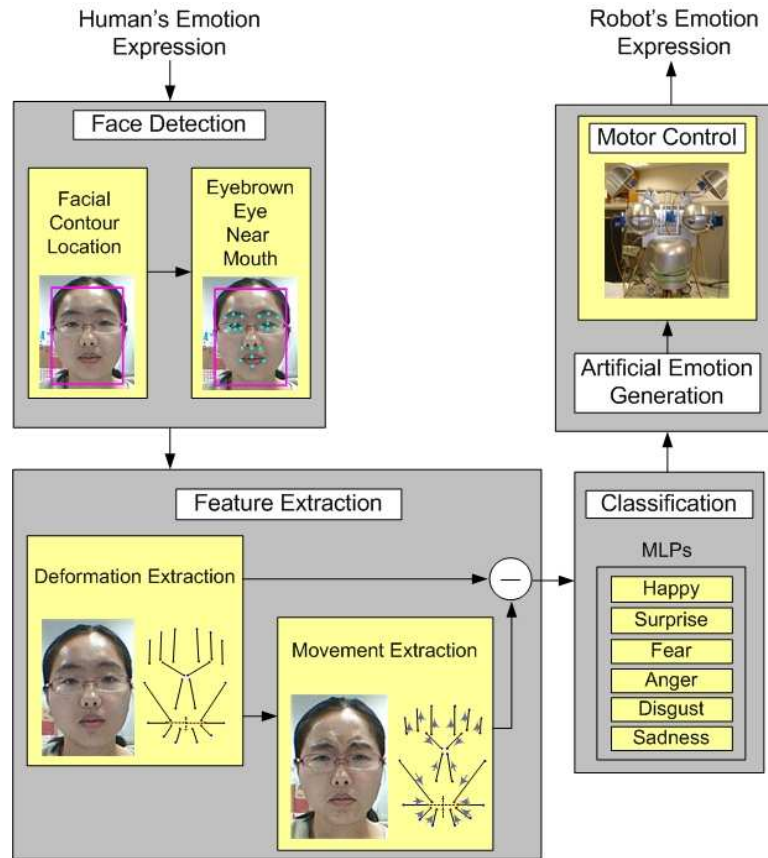


Figure 2.3: Robot imitates human facial expression.

The system framework is shown in Fig. 2.3. First the face detection module segments the face regions of a video sequence or an image and locates the positions of the eyebrows, eyes, nose and mouth. The positions can be represented by some

driven points with special mathematic properties (i.e., the minima). The module of feature extraction is used to track the driven points during a facial expression, and compute their sequential displacements compared to their corresponding fixed points. In the system a facial muscle is assumed to consist of a pair of key points, namely driven point and fixed point. The fixed points, which are derived from the facial mass-spring model, can not be moved during a facial expression. Given the outputs of feature extraction and a predefined set of facial expressions, the classification module classifies a video or an image into the corresponding class of facial expressions (i.e., happiness, fear, etc). Finally, the module of artificial emotion generation can control a social robot to imitate the facial expression in response of the user's expression.

The objective of the facial recognition is for human emotion understanding and intelligent human computer interface. The system is based on both the deformation and motion information. Fig. 2.1 shows the framework of our recognition system. The composition of our system can be distinguished in four main parts. It starts with the facial image acquisition and ends with facial expression animation.

# Face Detection and Feature Extraction

Human face detection is the first task performed in a face recognition system; consequently, to ensure good results in the recognition phase, face detection is a crucial procedure. In the last ten years, face and facial expression recognition have attracted much attention, though they truly have been studied for more than 20 years by psychophysicists, neuroscientists and engineers. Many research demonstrations and commercial applications have been developed from these efforts. The first step of any face processing system is to locate all faces that are present in a given image. However, face detection from a single image is a challenging task because of the high degree of spatial variability in scale, location and pose (rotated, frontal, profile). Facial expression, occlusion and lighting conditions also change the overall appearance of faces, as described in reference [78].

In reference [78], within a definition of face detection, the author writes: "Given an arbitrary image, the goal of face detection is to determine whether or not there are any faces in the image and, if present, return the image location and extent of each face".

Analysis of facial expressions requires a number of pre-processing steps which attempt to locate the face, to extract characteristic regions such as eyes, eyebrows, mouth and nose, to track the movement of facial features using anatomic information about the face.

## 3.1 Face Detection and Location using Skin Information

Skin has a quite characteristic range of colors, which indicates that the face region can be detected by classifying pixels on their color. There are different ways of representing the same color in a computer, each with a different color space. Each color space has its own existing background and application areas.

### 3.1.1 Gaussian Mixed Model

We know that although the images are from different ethnicities, the skin distribution is relatively clustered in a small particular area [17]. We denote a class conditional probability as $P(x|\omega)$ which is the probability of likelihood of skin color $x$ for each pixel of an image given its class $\omega$. This gives an intensity normalized color vector $x$ with two components. The definition of $x$ is given in Eq. (3.1).

$$x = [r, b]^T \tag{3.1}$$

where

$$r = \frac{R}{R + G + B}, b = \frac{B}{R + G + B} \tag{3.2}$$

Thus, we project the 3D [R,G,B] model to a 2D [r,b] model. On this 2D plane, the skin color area is comparatively more centralized which could be described by a Gauss distribution. $P(x|\omega)$ can be treated as a Gauss distribution, and the

equations of mean($\mu$) and covariance($C$) are given:

$$\mu = E(x) \tag{3.3}$$

$$C = E(x - M)(x - M)^T \tag{3.4}$$

Finally, we calculate the probability that each pixel belongs to the skin tone through the Gaussian density function as shown in Eq. (3.5). Then we use Gaussian distribution to describe this kind of distribution

$$P(x \mid \omega) \propto \exp[-0.5(x - \mu)^T C^{-1}(x - \mu)] \tag{3.5}$$

Through the distance between two pixels and the center we can get the information on how similar it is to skin and get a distribution histogram similar to the original image. The probability should be between 0 and 1, because we normalize the three components (R, G, B) of each pixel's color at the beginning. The probability of each pixel is multiplied by 255 in order to create a gray-level image $I(x, y)$. This image is also called a likelihood image.

## 3.1.2 Threshold & Compute the Similarity

After obtaining the likelihood of skin $I(x, y)$, a binary image $B(x, y)$ can be obtained by thresholding each pixel's $I(x, y)$ with a threshold $T$ according to

$$B(x, y) = \begin{cases} 1, & \text{if} \quad I(x, y) \geq T \\ 0, & \text{if} \quad I(x, y) < T \end{cases} \tag{3.6}$$

There is no definite criterion to determine a threshold. If the threshold value is too big, the false rate will increase. On the other hand, if the threshold is too small, the missed rate will increase. We hope the missed rate will be the lowest, so we define the threshold value as 0.5. That is, when the skin probability of a certain pixel is larger or equal to 0.5, we will regard the pixel as skin. In Fig. 3.1(b), the

(a) The original face image          (b) The binary image
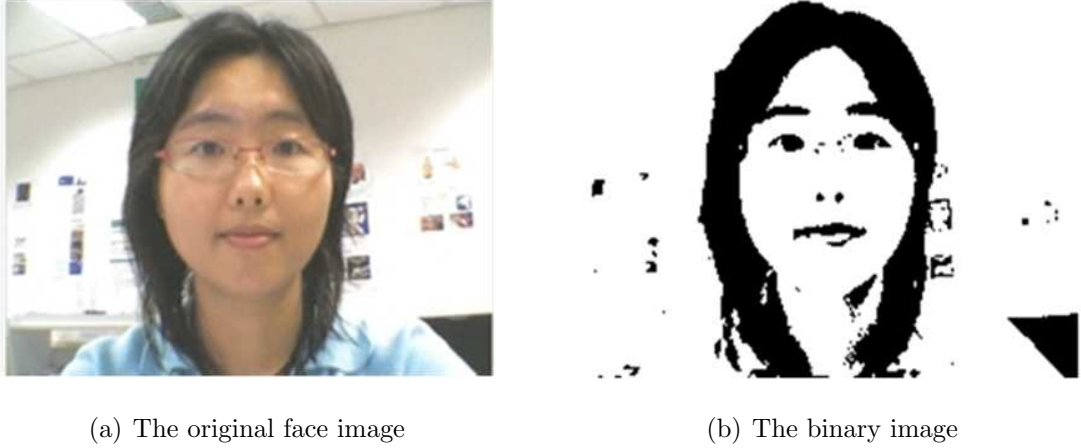
Figure 3.1: Face detection using vertical and horizontal histogram method

binary image $B(x, y)$ is derived from the $I(x, y)$ according to the rule defined in Eq. (3.6). As observed from the experiments, if the background color is similar to skin, there will be more candidate regions, and the follow-up verifying time will increase.

### 3.1.3   Histogram Projection Method

We have used integral projections of the histogram map of the face image for facial area location. The vertical and horizontal projection vectors in the image rectangle $[x1, x2] \times [y1, y2]$ are defined as:

$$V(x) = \sum_{y=y_1}^{y=y_2} B(x, y) \tag{3.7}$$

$$H(x) = \sum_{x=x_1}^{x=x_2} B(x, y) \tag{3.8}$$

The face area is located by applying sequentially the analysis of the vertical histogram and then the horizontal histogram. The peaks of the vertical histogram of the head box correspond with the border between the hair and the forehead, the

eyes, the nostrils, the mouth and the boundary between the chin and the neck.

The horizontal line going through the eyes goes through the local maximum of the second peak. The $x$ axis of the vertical line going between the eyes and through the nose is chosen as the absolute minimum of the contrast differences found along the horizontal line going through the eyes. By performing the analysis of the vertical and the horizontal histogram, the eyes' area is reduced so that it contains just the local maximums of the histograms. The same procedure is applied to define the box that bounds the right eye. The initial box bounding the mouth is set around the horizontal line going through the mouth, under the horizontal line going through the nostrils and above the horizontal line representing the border between the chin and the neck. By analyzing the vertical and the horizontal histogram of an initial box containing the face, facial feature can be tracked.
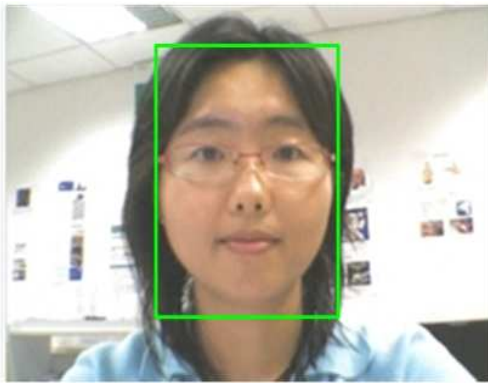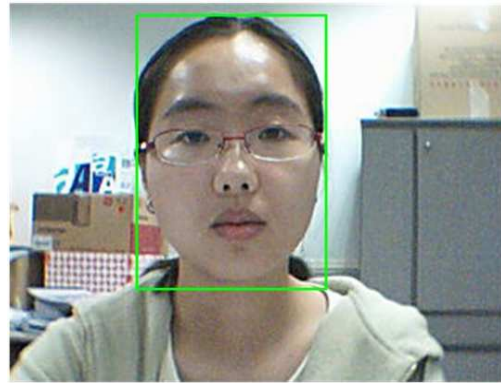


(a) Test image 1                         (b) Test image 2

Figure 3.2: The detected rectangle face boundary.

As can be seen from Fig. 3.2, faces can be successfully detected in different surroundings in these images where each detected face is shown with an enclosing window.

# 3.2 Facial Features Extraction

A facial expression involves simultaneous changes of facial features on multiple facial regions. Facial expression states vary over time in an image sequence and so do the facial visual cues. Facial feature extraction include locating the position and shape of the eyebrows, eyes, eyelids, mouth, wrinkles, and extracting features related to them in a still image of human face. For a particular facial activity, there is a subset of facial features that is the most informative and maximally reduces the ambiguity of classification. Therefore we actively and purposefully select 21 facial visual cues to achieve a desirable result in a timely and efficient manner while reducing the ambiguity of classification to a minimum. In our system, features are extracted using deformable templates with details given below.

## 3.2.1 Eyebrow Detection

The segmentation algorithm cannot give bounding box for the eyebrow exclusively. Brunelli suggests use of template matching for extracting the eye, but we use another approach as described below. Eyebrow is segmented from eye using the fact that the eye occurs below eyebrow and its edges form closed contours, obtained by applying Laplacian of Gaussian operator at zero threshold. These contours are filled and the resulting image containing masks of eyebrow and eye. From the two largest filled regions, the region with higher centroid is chosen to be the mask of eyebrow.

## 3.2.2 Eyes Detection

The positions of eyes are determined by searching for minima in the topographic grey level relief. The contour of the eyes can be precisely found. Since the real images are always affected by the lighting and noises, it is not robust and often

require expert supervision using the general local detection method such as corner detection [79]. The Snake algorithm is much more robust, but rely much on the image itself and there may be too many details in the result [80]. We can make full use of the priority knowledge of human face which describes the eyes as piecewise polynomial. A more precise contour can be obtained by making use of the deformable template.

The eye's contour model can be composed by four second order polynomials which are given below:

$$
\begin{cases}
y = h_1(1 - \frac{x^2}{w_1^2}) & -w_1 \le x \le 0 \\
y = h_1(1 - \frac{x^2}{w_2^2}) & 0 < x \le -w_2 \\
y = h_2(\frac{(x+w_1-w_3)^2}{w_3^2} - 1) & -w_1 \le x \le w_3 - w_1 \\
y = h_2(\frac{(x+w_1-w_3)^2}{(w_1+w_2-w_3)^2} - 1) & 0 < x \le -w_2
\end{cases}
\tag{3.9}
$$

where $(x_0, y_0)$ is the center of the eye, $h_1$ and $h_2$ are the heights of the upper half eye and the lower half eye, respectively.
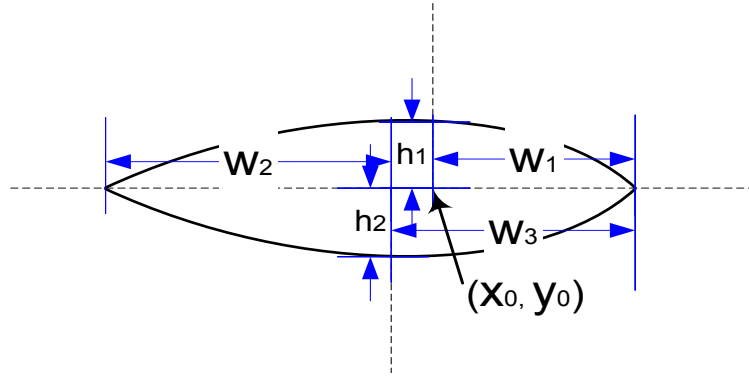


Figure 3.3: The outline model of the left eye.

Because the eyes's color are not accordant and the edge information is abundant, we can do edge detection first with a closed operation followed. The inner part

of the eye becomes high-luminance while the outer part of the eye becomes low-luminance. The evaluation function we choose is:

$$\min C = \oint_{\partial} D^+ I(\mathbf{x})d\mathbf{x} - \oint_{\partial} D^- I(\mathbf{x})d\mathbf{x} \tag{3.10}$$

where $D$ represent the eye's area, $\partial D^+$ denotes the outer part and $\partial D^-$ denotes the inner part of the eye.

### 3.2.3 Nose Detection

After the eyes' position is fixed, it will be much easier to locate the nose position. The nose is at the center area of the face rectangle. We can search this area for the light color region. Thus the two nostrils can be approximated by finding the dark area. Then the nose can be located above the two nostrils at the brightest point.

### 3.2.4 Mouth Detection

Similar to the eye's model, the lips can be modeled by two pieces of fourth order polynomials which are given below:

$$\begin{cases} y = h_1(1 - \frac{x^2}{w^2}) + q_1(\frac{x^2}{w^2} - \frac{x^4}{w^4}) & -w \leq x \leq 0 \\ y = h_2(\frac{x^2}{w^2} - 1) + q_2(\frac{x^2}{w^2} - \frac{x^4}{w^4}) & 0 \leq x \leq w \end{cases} \tag{3.11}$$

where $(x_0, y_0)$ is the lip center position, $h_1$ and $h_2$ are the heights of the upper half and the lower half of the lip respectively.
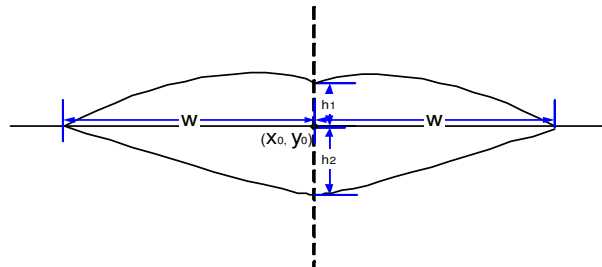
Figure 3.4: The outline model of the mouth.

The mouth's evaluation function is much easier to confirm since the color of the mouth is uniform. The mouth could be easily separated by the different color of mouth and skin. The position of mouth can be determined by searching for minima in the topographic grey level relief. The formation of the evaluation function is similar to Eq. (3.11).

## 3.2.5 Illusion & Occlusion

The wear of glasses, scarves and beards would change the facial appearance which make it difficult for face detection and feature extraction. Some previous work has addressed the problem of partial occlusion [81]. The method they proposed could detect a face wearing sunglasses or scarf but is conducted under restrained conditions. The people with glasses can be somehow detected but it may fail sometimes. Fig. 3.5 shows the face detection and feature extraction results with glasses. In this paper, we did not consider the occlusion problem such as scarf or purposive occlusion. Such occlusion may cover some of the feature points, and the face recognition can't be conducted subsequently.
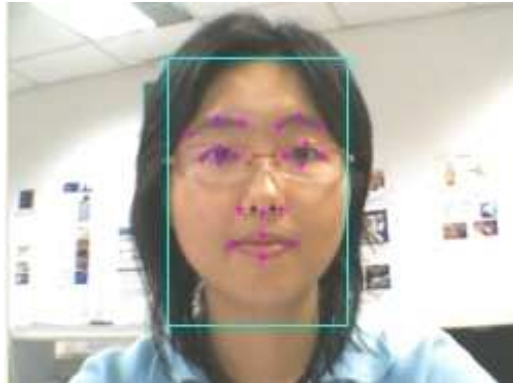
Figure 3.5: The feature extraction results with glasses.

## 3.3 Summary

In this Chapter, the face detection and facial features extraction methods are discussed. Face detection can fix a range of interests, decrease the searching range and initial approximation area for the feature extraction. Vertical and horizontal projection methods are conducted to automatically detect and locate face area. And then facial features are extracted by using deformable templates to get precise positions.

# Chapter 4

# Non-linear Mass-spring Model for Facial Expression

The muscles in our face allow us to express emotions without speaking. To make an expression, we move the facial muscles that lie beneath the skin. Unlike other skeletal muscles, which are attached to bones, the facial muscles are attached to other muscles, or to the skin. So even a tiny contraction in one such muscle can pull the skin and change your expression [82].

Yu zhang et al. proposed a physically-based dynamic facial model based on anatomical knowledge for facial expression animation. The facial model incorporates a physically-based approximation to facial skin and a set of anatomically-motivated facial muscles. The skin model is established by using a mass-spring system with nonlinear springs, and they are used to simulate the elastic dynamics of a real facial skin. Facial muscle models are developed to emulate facial muscle contraction [29]. In this Chapter, we investigate the facial muscles's tension by using linear and non-linear mass-spring models.

# 4.1 Introduction to Facial Muscles

## 4.1.1 Facial Muscles I

The Fig. 4.1 shows the facial muscles in a human face. There are nine groups of muscles in the face that control facial expression. Two groups, that cover the eyelid and orbital area, control blinking, tear duct control and movement of the eyeball. Near the nose, there are several small muscles that interconnect with other muscles in the face, enabling nostrils to flair or compress, and the upper lip to lift. A muscle runs vertically along the forehead, raising the eyebrows and helping the face to frown. The "kissing muscle" (known to anatomists as the orbicularis oris) closes the mouth and puckers the lips when it contracts. As an expressive muscle, four relatively distinct movements can be produced by orbicularis oris, a pressing together, a tightening and thinning, a rolling inwards between the teeth, and a thrusting outwards. Other muscles control the corners of the mouth: Risorius acts to stretch the mouth laterally, retracting the corners of the mouth, and has been thought (erroneously) to produce "grinning" or "smiling"; Zygomatic major lifts the corner of the mouth obliquely upwards and laterally and is a muscle that produces a characteristic "smiling expression" (Other muscles produce different "smiles"); Triangularis This muscle causes the corners of the mouth to turn down and form the lips into an inverted U, an action stereotyped as indicating grief. It produces a frown in the mouth [83].

All these muscles are connected by the facial nerve. The facial nerve contains about 10,000 individual nerve fibers and works like a telephone cable. It carries electrical impulses to a specific facial muscle, and this signal is what enables us to laugh, cry, smile, or frown [82].

The actions of above facial muscles are described as follows:

1. The frontalis muscle runs vertically on the forehead, originating in tissues
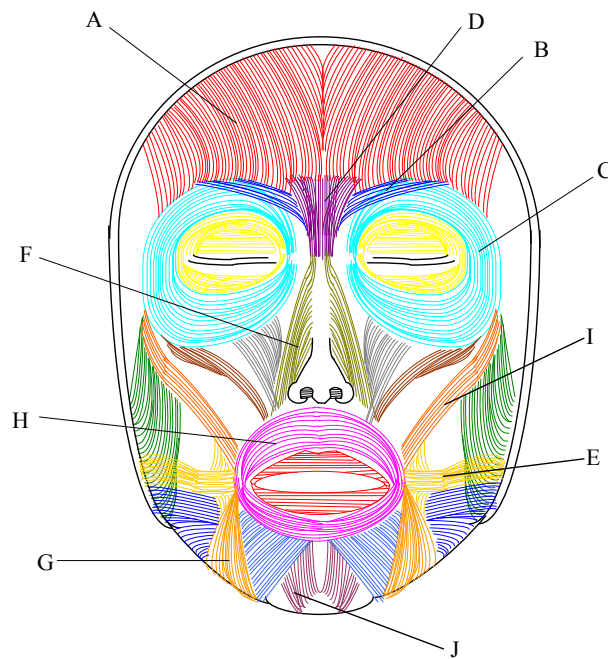
Figure 4.1: The primary muscles of facial expression include: (A) Frontalis (B) Corrugator (C) Orbicularis oculi (D) Procerus (E) Risorius (F) Nasalis (G) Triangularis (H) Orbicularis oris (I) Zygomatic minor (J)Mentalis

of the scalp (galea aponeurotica) above the hairline and inserting into the skin in the forehead and near the eyebrows. (It is considered the front part of the Epicranius muscle or Occipito-frontalis which covers the scalp from the forehead to the back of the head.) Contraction of the entire frontalis draws the eyebrows and skin of the forehead upwards and forms horizontal wrinkles running across the forehead. It is composed of inner (medial) and outer (lateral) parts, which can function relatively independently.

Frontalis is innervated by temporal branches of the facial nerve (VII) and is supplied with blood by the superficial temporal artery.

The inner frontalis is the medial part of the frontalis muscle. Its contraction raises the medial part of the brow and eyebrows, forming slanted wrinkles in the forehead and creating a slant up towards the center in the eyebrows.

The outer frontalis is the lateral part of the frontalis muscle. Its contraction raises the lateral (outer) part of the brow and eyebrows, forming wrinkles in the lateral part of the forehead and an arched shape to the eyebrows.

2. The corrugator muscle originates at the inner orbit of the eye near the root of the nose and inserts into the skin of the forehead above the center of each eyebrow. It pulls the eyebrows and skin from the center of each eyebrow to its inner corner medially and down, forming vertical wrinkles in the glabella area and horizontal wrinkles at the bridge of the nose. It most often acts simultaneously with two nearby smaller muscles, the depressor supercillii and the procerus. It is one of the most important of expressive muscles. Some suggest this is the muscle of grief and suffering (research suggests much more diverse roles). It produces a frown in the eyebrows and forehead.

3. Orbicularis oculi is a sphincter muscle around the eye and acts, in general, to narrow the eye opening and close the orbit of the eye. This muscle has

important functions in protecting and moistening the eye as well as in expressive displays. These muscles constrict skin around the eye, reduce the eye opening, and close the eye. It has three parts, an outer or orbital part, an inner or palpebral part in the eyelids, and a small lacrimal part near the tear duct. The outer part originates in the medial part of the orbit and runs around the eye via the upper eye cover fold and lid and returns in the lower eyelid to the palpebral ligament; the palpebral part originates in the palpebral ligament and runs above and below the eye to the lateral angle of the eye. These two muscles form concentric circles around the eye. Action of the palpebral part is often involuntary, as in the blink reflex.

4. The Procerus (also known as the depressor glabellae or pyramidalis nasi) muscle originates in the fascia of the nasal bone and upper nasal cartilage, runs through the area of the root of the nose, and fans upward to insert in the skin in the center of the forehead between the eyebrows. It acts to pull the skin of the center of the forehead down, forming transverse wrinkles in the glabella region and bridge of the nose. This horizontal wrinkle at the root of the nose is sometimes referred to as the "champion pucker" because this muscle often contracts in effortful activities. It usually acts together with corrugator and/or orbicularis oculi and/or the nasal part of levator labii superioris. It is very difficult to contract deliberately without involving these other muscles.

5. Risorius originates in the fascia of the masseter below the zygomatic arch and inserts in the skin near the corner of the mouth. It acts to stretch the mouth laterally, retracting the corners of the mouth, and has been thought (erroneously) to produce "grinning" or "smiling." It has a connection with the platysma in that it often contracts with it.

6. The Nasalis muscle has two main parts, the transverse or compressor part (also known as compressor naris), which constricts the nostril, and the alar or dilator part (also known as dilator naris), which flares the nostril. The compressor part of nasalis originates in the upper jaw near the canine tooth and inserts into nasal cartilage on the bridge of the nose, each side mixing with the other (thus transverse). When it contracts, it tends to draw the nostril wings towards the septum. The dilator part originates in the upper jaw and cartilage of the nose and inserts in skin of the nostril. When it contracts, it pulls the nostril wings away from the septum. (Depressor septii is considered by some to be a part of nasalis.)

7. Triangularis, a name based on its shape, (also known as Depressor anguli oris) originates in the mandible and platysma and inserts in the skin and orbicular muscle at corner of the mouth. It is a muscle whose evolutionary connection to the platysma is evident, being continuous with it and extending to the mouth. This muscle causes the corners of the mouth to turn down and form the lips into an inverted U, an action stereotyped as indicating grief. It produces a frown in the mouth.

8. Orbicularis oris is the sphincter muscle around the mouth, forming much of the tissue of the lips. It has extensive connections to muscles that converge on the mouth. This muscle acts to shape and control the size of the mouth opening and is important for creating the lip positions and movements during speech. Several different strands can be distinguished that allow it to form the lips into versatile shapes. As an expressive muscle, four relatively distinct movements can be produced by orbicularis oris, a pressing together, a tightening and thinning, a rolling inwards between the teeth, and a thrusting outwards.

9. Zygomatic major originates in the cheek bone (zygomatic arch) and inserts in muscles (o. oris, depressor, etc.) near the corner of the mouth. This muscle lifts the corner of the mouth obliquely upwards and laterally and is a muscle that produces a characteristic "smiling expression." (Other muscles produce different "smiles.") Some research suggests that the difference between a genuine smile and a perfunctory (or lying) smile is that when a person really feels happy, Zygomatic major contracts together with orbicularis oculi. Look at the videos below and see what you think (both expressions here are deliberate).

10. Mentalis is so named because it is associated with thinking or concentration, although the justification for this view is lacking. It also has been said to express doubt. It originates in the part of the mandible below the front teeth and inserts into the skin of the chin, and acts to push the chin boss upwards, wrinkling it and curving the lips upward in an inverted U.

## 4.1.2 Facial Muscles II

The facial muscles are mostly attached to both the skull and the facial tissue. One end of the facial muscle attached to skull is generally considered the origin while the other end is the insertion. Normally, the origin is the fixed point, and the insertion is where the facial muscle performs its action. In a human face, a wide types of muscles exist: rectangular, triangular, sheet, linear, sphincter [84]. Three main types of facial muscles are incorporated in our face model. They are linear, sphincter and sheet muscles. Thus the nine groups of facial muscles in section 4.1.1 can be categorized as follows.

Table 4.1: Facial Muscle Classification

| Linear muscle | Corrugator, Risorius, Nasalis, Triangularis, Zygomatic minor |
|---|---|
| Sphincter muscle | Orbicularis oculi, Orbicularis oris |
| Sheet muscle | Frontalis, Procerus, Mentalis |
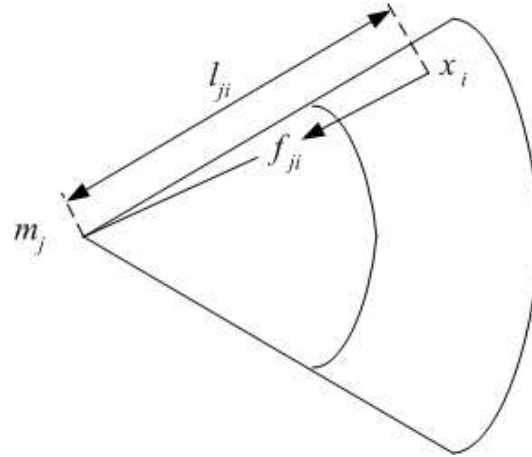


Figure 4.2: Linear muscle

**Linear Muscle**

Linear muscle consists of a bundle of fibers that share a common emergence point in bone and pulls in an angular direction. One of the examples is the zygomaticus major which attaches to and raises the corner of the mouth. Fig. 4.2 illustrates the linear muscle with the following definitions [84]:

$x_i$: arbitrary facial skin point

$m_j$: attachment point of linear muscle $j$ at the skull

$x_{ji}$: the distance between muscle attachment point $m_j$ and skin point $x_i$

On contraction, facial regions close to the skin insertion point of a muscle are affected. The effect of facial muscle contraction is to pull the surface from the area of the muscle insertion point to the muscle attachment point.
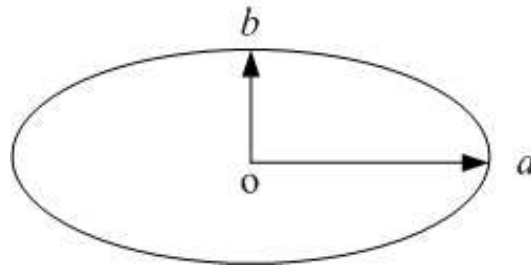
Figure 4.3: Sphincter muscle

**Sphincter Muscle**

Unlike the linear muscle,the sphincter muscle attaches to skin both at the origin and at the insertion, and contracts abound a virtual center. An example is the orbicularis oris, which circles the mouth and can pout the lips. because sphincter muscles do not behave in a regular fashion, it can be simplified to a parametric ellipsoid as shown in Fig. 4.3. The definition of the parameters list are:

$O$: epicenter of sphincter muscle influence area

$a$: the semimajor axis of sphincter muscle influence area

$b$: the semiminor axis of sphincter muscle influence area

**Sheet Muscle**

Sheet muscle consists of strands of fibers which lie in flat bundles. The obvious example of this kind of muscle is the frontalis major, which lies on the forehead and is primarily involved with the raising of the eyebrows. A sheet muscle neither emanates from a point source, nor contracts to a localized node. In fact, the sheet muscle is a series of almost-parallel fivers spread over an rectangle area, muscle model is illustrated in the Fig. 4.4

$x_i$: arbitrary facial skin point

$m_j$: point of sheet muscle attachment line
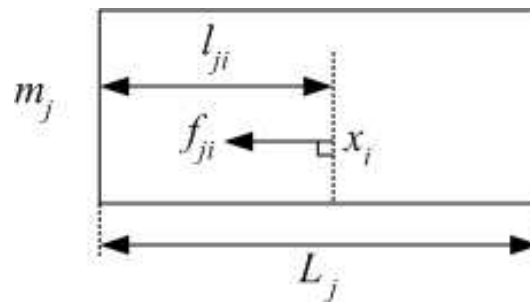
Figure 4.4: Sheet muscle

$L_j$: the length of the rectangle zone influenced by sheet muscle

$l_{ji}$: the distance between skin point xi and sheet muscle attachment line

## 4.2   Facial Motion and Key Points

For developing a representation of facial motion, we have to find a proper method to represent the movement of facial muscles. We employed the Simunek's method which is for visualization and animation of human face. This approach models the facial motions based on the deformations of muscles and uses key points to analyze the movement of the lips [76]. Using key points introduced by this method, we analyze the movement of the facial muscles. All facial muscles are implemented as vectors. Two points of the vector determine places, where the muscle is attached. The first point is mobile and we call them driven points. The second point is immovable and we call them fixed points. The movement of the muscles is implemented as extending or reducing of a distance between points of the vector. This reduction or extension performed by movement of control point. We have to determine limits of vector length by anatomy of a human face. They are depicted on following picture.

In Fig. 4.5, we mark the driven and fixed points of the muscles by using two

(a) (b)
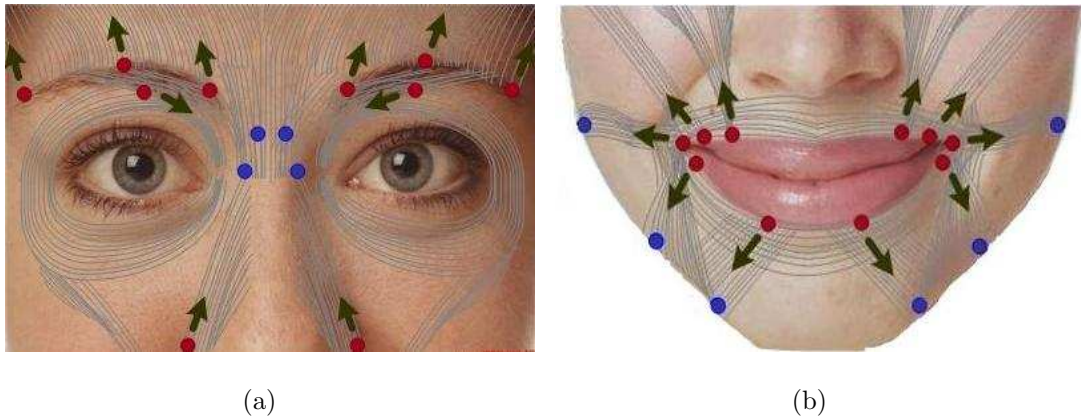
Figure 4.5: Key points

different colors: red key points denote driven points, the blue ones denote fixed points. Facial muscles are plotted by grayer lines and its driven and fixed points are also connected by them.

## 4.3 The Linear Mass-Spring Face Model

To physically simulate the deformation of the skin on the human face, we use the mechanical law of mass-spring model. Networks of masses, connected by spring, attempt to simulate the behavior of deformable bodies using a primitive model for the transmission of energy. The motion of a particle in the system is defined by its physical nature and by the position of other particles. The facial surface is composed by a set of particles with uniform mass density $m$. Their behavior is determined by their interaction with the related muscles. In a correspondence with the geometric structure of the face model, each key point of the face corresponds to a particle in the physical model. To simulate elastic effects of facial skin tissue, we connect each face driven key point with its fixed point by massless spring of natural length non equal to zero.

Suppose an driven skin mass point $\mathbf{x}_i$ is connected with its fixed points $\mathbf{x}_j$ by the spring $j$. The internal spring forces applied on $\mathbf{x}_i$ is the resultant of the tensions of the springs linking $\mathbf{x}_i$ to its fixed point:

$$f(x_i, x_j) = k_{ij} \frac{(|x_i - x_j| - d_{ij})}{|x_i - x_j|}(x_i - x_j) \tag{4.1}$$

where

$d_{ij}$ is the natural length of the spring linking $\mathbf{x}_i$ and $\mathbf{x}_j$.

$k_{ij}$ is the spring stiffness of the spring linking $\mathbf{x}_i$ and $\mathbf{x}_j$.

$$\begin{cases} k_{ij} = k_L & \varepsilon_j \leq \varepsilon^c \\ k_{ij} = k_H & \varepsilon_j > \varepsilon^c \end{cases} \tag{4.2}$$

The spring forces are computed by multiplying the elongation from the rest length $d_{ij}$ of the spring with its spring stiffness $k_{ij}$. The low-strain stiffness $k_L$ is smaller than the high-strain stiffness $k_H$. Like real skin tissue, the biphasic spring is readily extendible at low strains, but exerts rapidly increasing restoring stresses after exceeding a strain threshold $\varepsilon^c$.

## 4.4 Nonlinear Mass-Spring Model (NLMS)

In order to faithfully simulate the deformation of the facial skin tissue, it is crucial to investigate the biomechanical nature of soft tissue deformation under applied loads. Experimental data have been collected in Biomechanics about human tissue elasticity [85]. The study shows that tissues do not have a linear response: the curve representing the stretch (strain) of a tissue as a function of the applied force (stress)is typically a J-shaped curve; as the tissue gets closer to tearing, the increase in stretching becomes smaller per additional unit of exerted force. Moreover, the tissue response exhibits hysteresis: the curves for increasing and decreasing force

are different. Each branch of a specific cyclic process can be described by a non-linear pseudo-elastic function. Since the difference is insignificant, we approximate the non-linear relationship by a biphasic curve illustrated in Fig.4.6
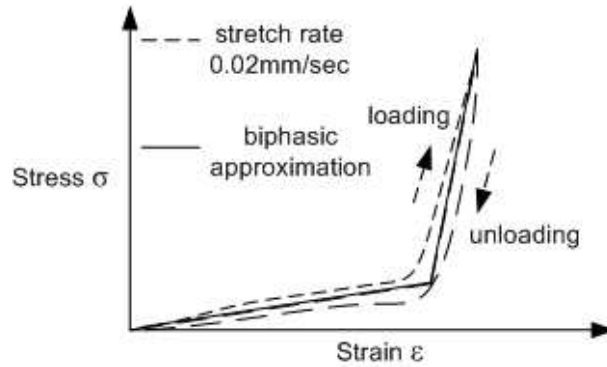


Figure 4.6: Stress-strain relationship of facial tissue

Mass-spring model is typically utilized to formulate the facial muscle contraction. The facial muscle is treated as a linear spring and the elastic stiffness is constant. Though this assumption simplifies somewhat the equation of motion at each node, it is undesirable for accurate simulation of the real tissue that has a nonlinear stress-strain relationship. It is natural to investigate the problem of the elastic stiffness calculation for nonlinearity factor varying with muscle deformation. In the existing facial expression approaches based on mass-spring model, the analysis about facial deformations mainly focuses on the displacement of facial feature or potential energy. In this section, the mass-spring model is firstly discussed for nonlinear stress-strain relationship with variable elastic stiffness.

In order to simulate nonlinear deformation of the muscle spring, we need a nonlinear function to describe the stress-strain relationship. The works as demonstrated in [86] provide us the mechanical law of soft-tissue points. Using this method, we calculate the elastic stiffness and elastic force for each functional muscle. Suppose an arbitrary driven point $x_i$ is connected to its corresponding fixed points $x_j$ by a

structure spring with rest length $d_{ij}$. Let $\Delta x_{ij} = x_i - x_j$, we introduce a function $K(x_i, x_j)$ to modulate a constant elastic stiffness $k_0$:

$$K(x_i, x_j) = (1 + (|\Delta x_{ij}| - d_{ij})^2)^\alpha k_0 \qquad (4.3)$$

and the elastic force generated by an spring is:

$$f(x_i, x_j) = K(x_i, x_j) \frac{(|\Delta x_{ij}| - d_{ij})}{|\Delta x_{ij}|} \Delta x_{ij} \qquad (4.4)$$

In equation (4.3), $\alpha$ is the nonlinearity factor controlling the modulation. In the later sections, we use $f_{ij}$ to denote $f(x_i, x_j)$.

By assigning different values to $\alpha$, function (4.3) can be chosen to model linear or nonlinear stress-strain relationship. Fig. 4.1 illustrates the stress-strain relationship for different values of $\alpha$. According to [9], we took the value of $\alpha$ as 1.0 and $k_0$ as 1.0.
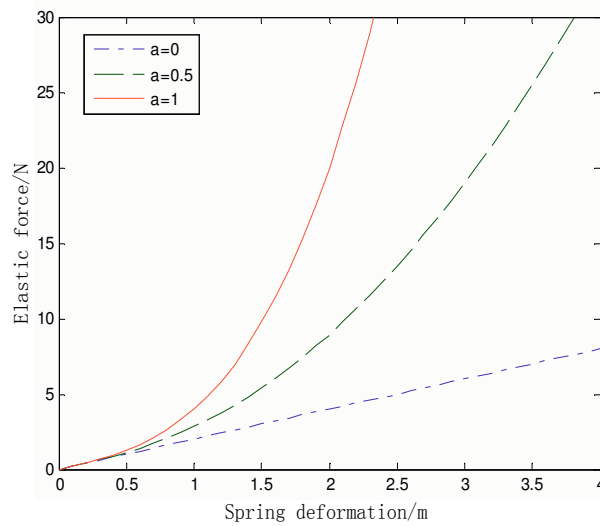


Figure 4.7: The stress-strain relationship of structure spring with different values of $\alpha$, $k_0 = 1.0$

## 4.5   Modeling Facial Muscles based on NLMS

We use a muscle mapping approach for facial muscle construction. By using OpenCV, we first save a bitmap from the color buffer. It records the RGB values of the facial surface. We then specify a set of key points on this bitmap to identify the ideal locations of the facial muscles that should be designed on it (see Fig. 4.8). Based on the Facial Action Coding System (FACS, we select 22 major functional facial muscles to simulate facial expressions. For a linear or sheet muscle, the positions of the fixed and driven points of its central muscle fiber completely define the location of the muscle. We mark the attachment and insertion points of the muscles by using two different colors. In Fig. 4.8, red key points are muscle driven points, the blue ones are muscle fixed points. For each muscle, its fixed and driven points are connected by a spring. The driven points are controlled by the related mass-springs fixed in fixed points. The positions of the key points are marked once on the reflectance image and the resulting image is named facial muscle image.
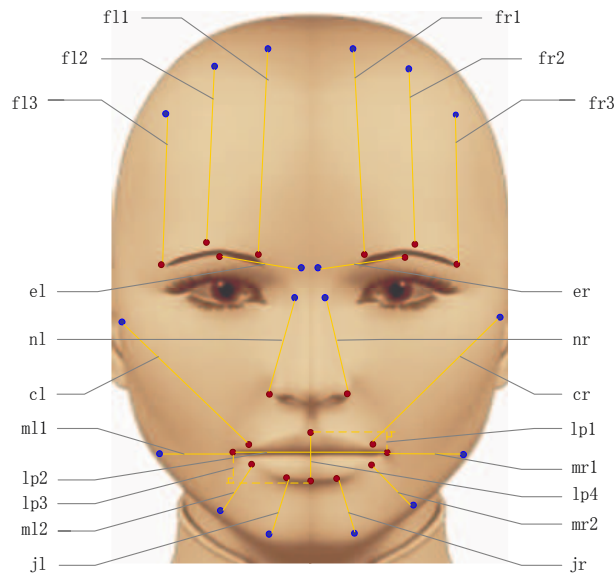


Figure 4.8: The facial mass-spring model

Once the marks are all made, the texture coordinates of each facial mesh vertex in the facial muscle image are calculated based on an orthographic projection and the facial muscle image is mapped automatically to the 2D face. Fig. 4.9 shows the examples of facial expression images and their corresponding muscles' moving direction in the face regions. The deformation maps exhibit different patterns corresponding to different facial expressions. In order to give an explicit and quantitative description, we use a nonlinear mass-spring model to describe the physical property of the deformation map.
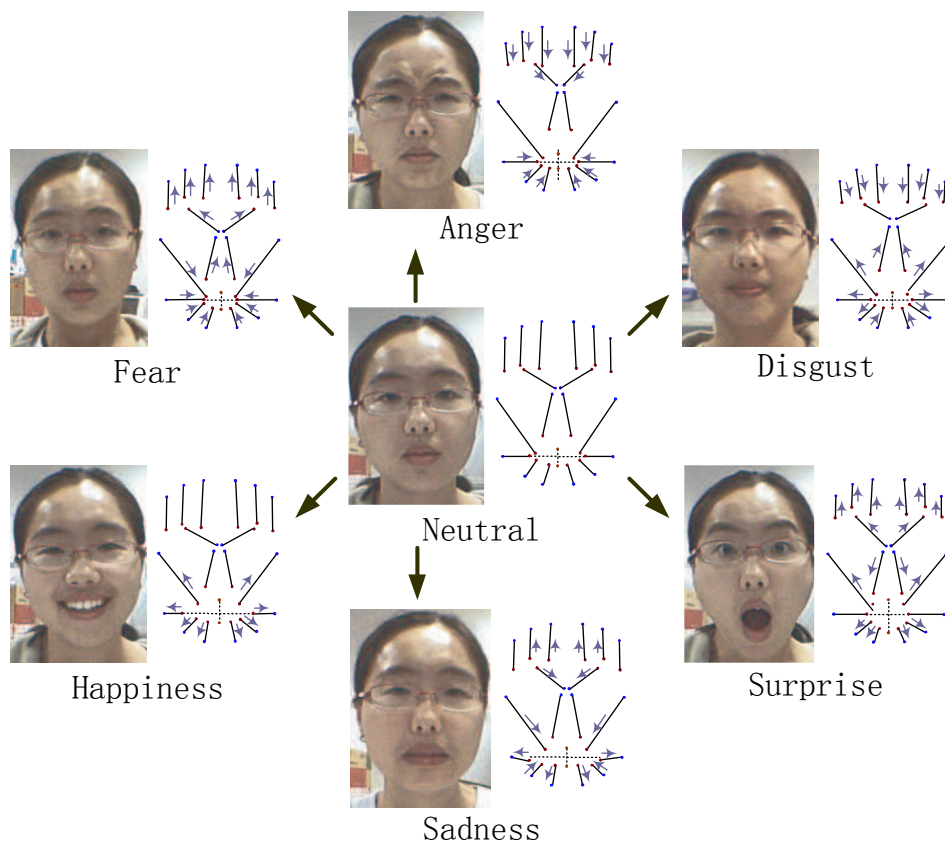


Figure 4.9: Facial expression images and the corresponding deformation maps in face regions.

# 4.6   Experiments and Discussions

In this section, we will study the facial muscles' tension for different facial expressions, and then extract the novel visual features based on such characteristics for facial expression classification. It is possible that we can encode both the magnitude and the direction of motion by using elastic force of facial muscles. The psychological experiments as shown in [87] have suggested that facial expressions are more accurately recognized from a temporal behaviors from a single static image. The temporal information often reveals the underlying emotional states. Therefore, our work concentrates on modeling the temporal behaviors of facial expressions from their dynamic appearances in an image sequence.

## 4.6.1   Classification Results Comparing with Linear Model

We employed 20 men and 20 women to make the facial expressions in our experiments. Each person was asked to make only one facial expression every time, and totally each person has to make all the six facial expressions. In each experiment, we measured the facial muscle mass-spring force of every person's expression, so totally we obtained 40 samples of such a mass-spring force for each facial expression. Thus in the Figure 4.9 we show the mean values of these samples for each facial expression under the linear and non-linear mass-spring models.

As shown in Fig 4.10, we compared linear and non-linear mass-spring face model for each muscle's tension of different facial expressions. The mean value calculated by linear model rang from $-30$ to $30$ and there are no distinct distribution for different emotion. In contrast, the mean value calculated by our technique rang from $-800$ to $1000$. It is worth saying that the nonlinear module leads to more wide distribution, which is directly related to efficiently differentiate the value of muscle's tension at different expression [9]. For instance, when the face express happy,

(a) Mouth1-Nonlinear  (b) Mouth1-Linear  (c) Cheek-Nonlinear  (d) Cheek-Linear

(e) Forehead1-Nonlinear  (f) Forehead1-Linear  (g) Lip2-Nonlinear  (h) Lip2-Linear
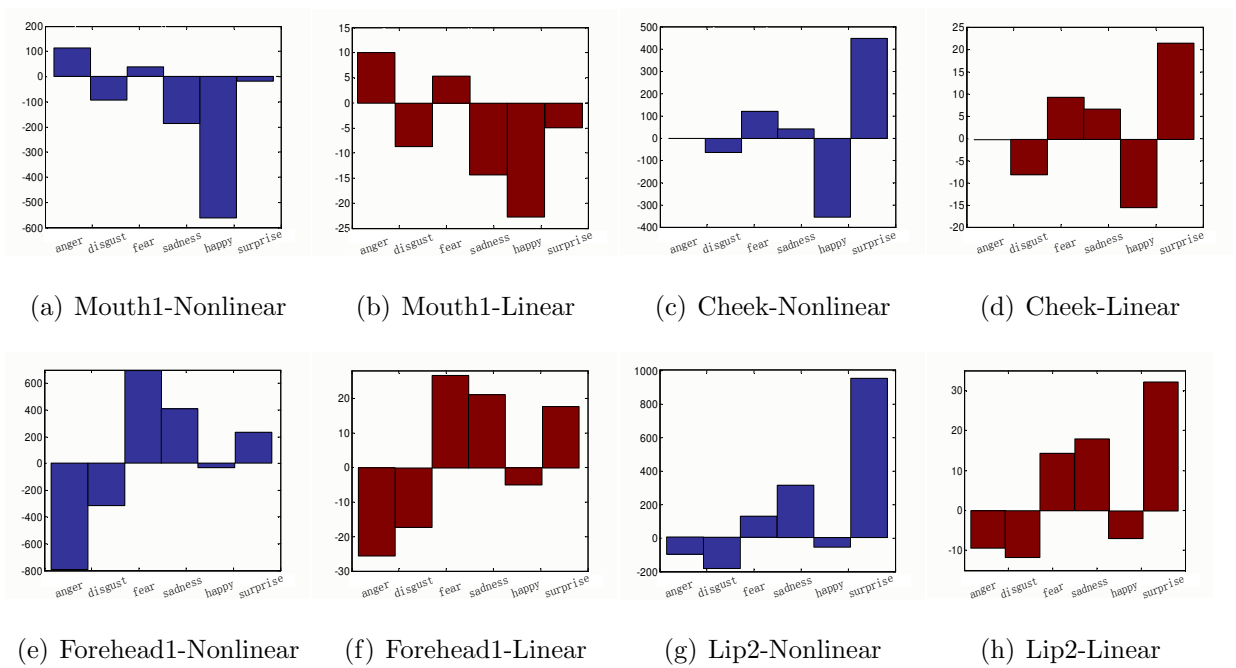
Figure 4.10: The performance of the facial muscle tracking method using
nonlinear model and linear model respectively
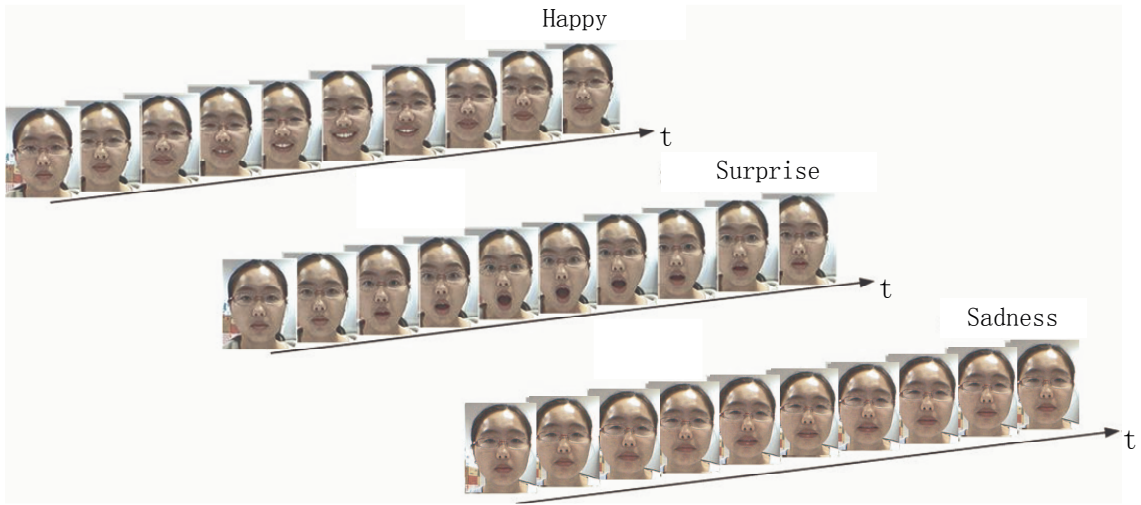
Figure 4.11: Three videos of tracking a set of the deformations in face sequence.

sadness, surprise, disgust, fear and anger, the mean value for muscle 'forehead1' could reach $-50$, $400$, $260$, $-300$, $700$ and $-800$ respectively.

## 4.6.2    Examples based on integration

Fig.4.11 shows three processes for happy, surprise and sadness. In [69], temporal changes in neuromuscular facial activity last of a second to several minutes. Therefore we empirically determined a 10-second of temporal duration based on a video frame rate of 24 frames. All the sequences start from the neutral state to the emotional state. In terms of the image sequences of Fig. 4.11, Fig. 4.12 shows the temporal curves of corresponding elastic forces. As shown in Fig. 4.12, there are three distinct phrases: starting, apex and ending. At the neutral state, all the facial features locate at their equilibrium positions and the elastic forces are equal to zero. When one facial expression reaches its apex state, the magnitude of elastic force reaches the largest value. When the expression is approaching to the ending state, the magnitude of elastic force is decreasing accordingly.
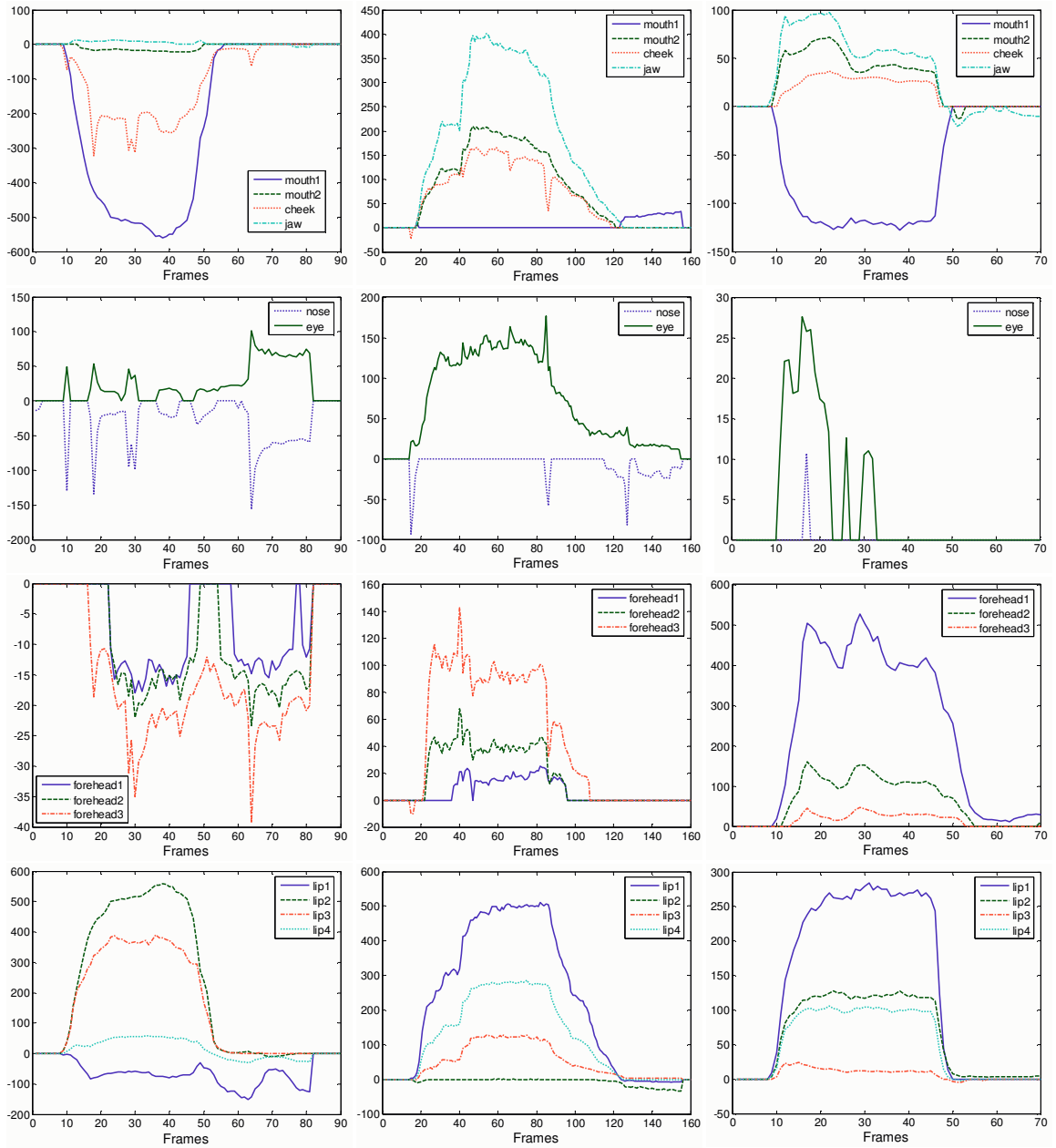
(a) Happy          (b) Sadness          (c) Surprise

Figure 4.12: Results of tracking associated with three video sequences show in
Fig. 4.11

We observed that the procedures of these three states are different for three facial expressions. Different facial expressions have their unique temporal patterns at these three states. Therefore we can make use of such magnitudes of muscle mass-spring forces to classify the facial expressions.

### 4.6.3   Examples based on facial action units

The recovered muscle motions are represented in term of magnitudes of some pre-defined motion of various facial features. Each feature motion corresponds to a simple deformation on the face. In order to objectively capture the richness and complexity of facial motions, behavioral scientists have found it necessary to develop objective coding standards. The facial action coding system (FACS) is the most commonly used and compressive coding system in the behavioral sciences. The system was again trained on Cohn and Kanade's DFAT-504 data set which contains FACS scores by two certified FACS coders in addition to the basic emotion labels. The FACS was developed by Ekman and Friesen [69] for describing facial expressions by action units (AUs). Of 44 FACS AUs that they defined, 30 AUs are anatomically related to the contractions of specific facial muscles: 12 are for upper face, and 18 are for lower face.

We refer to these motions vectors as AUs. Each AUs is indeed the combination of related muscles' deformations. We group muscles of AUs as primary muscles and auxiliary muscles. By the primary muscle, those muscle or muscle combinations can be clearly classified as or are strongly pertinent to one AU without ambiguities. In contrast, the auxiliary muscle or muscle combinations can be only additively combined with primary muscle to provide supplementary support to the AUs. Consequently, an AU contain primary muscle and auxiliary muscle. For example, six forehead muscles can be directly associated with AU1(Inner Brow

Raiser), AU2 (Outer Brow Raiser) and AU4 (Brow Corrugator), while it is ambiguous to associate eye muscle with these AUs. When forehead muscles and eye muscle deform simultaneously, the classification of this muscle combination to one of above AUs(AU1, AU2 and AU4) then becomes certain. Hence, forehead muscles are a primary muscle combination of AU1, AU2 and AU4, while eye muscle is an auxiliary muscle of AU1, AU2 and AU4. Table 4.2 and Table 4.3 give a summary of primary muscle or muscle combination and auxiliary muscle or muscle combination associated with some AUs. The AUs are used as the basic features for the classification scheme described in the next sections.

Table 4.2: The Association of Upper Face AUs to Muscle Deformation

| AU code | AU | Primary Cues | Auxiliary Visual Cues |
|---------|------|--------------|----------------------|
| 1 | Inner brow raise | forehead 1, 2, 3 | eye |
| 2 | Outer brow raise | forehead 1, 2, 3 | eye |
| 4 | Brow corrugator | eye | forehead 1, 2, 3 |
| 5 | Upper lid raise | eye | forehead 1, 2, 3 |
| 6 | Cheek raise | cheek | nose, eye |
| 7 | Lid Tightener | eye | forehead 1, 2, 3, nose |

Fig. 4.13 shows six animated processes for AU1, AU2, AU7, AU19, AU15 and AU27. The primary muscle is shown by solid curve and the auxiliary muscle is shown by broken curve. By combining primary muscles from different AUs, we have some observations: 1) The value of muscle's deformation across different AUs, e.g., muscle 'Lip1', when its deformation value reaching 270, generates a primary cues combination for AU20 shown as Fig.4.13 (e); when its deformation value reaching 860, generates a primary cues combination for AU27 as illustrated in Fig.4.11 (f) and 2) primary muscles' combinations belong to different AUs, e.g., when 'Lip 2' and 'lip 3' are positive, 'Lip 1' and 'Mouth 1' are negative, the four primary muscles

Table 4.3: The Association of Lower Face AUs to Muscle Deformation

| AU code | AU | Primary Cues | Auxiliary Visual Cues |
| --- | --- | --- | --- |
| 9 | Nose wrinkle | nose | cheek, eye, forehead 1, 2 |
| 10 | Upper lip raiser | lip 1, 3, 4, mouth 1, 2 | cheek, jaw, lip 2 |
| 12 | Lip corner puller | lip 1, 2, 3, mouth 1, 2 | cheek, jaw, lip 4 |
| 15 | Lip corner depressor | mouth 1, 2, jaw, lip 1, 3 | cheek, lip 2, 4 |
| 17 | Chin raise | mouth 2, jaw, lip 1, 3 | mouth1, cheek, lip 2, 4 |
| 20 | Lip stretcher | lip 2, 4 mouth 1, 2 | cheek, jaw, lip 2, 4 |
| 23 | Lip tighter | lip 2, 4 mouth 1, 2 | cheek, jaw, lip 2, 4 |
| 25 | Lips part | lip 1, 3, 2, 4 mouth 1, 2 | cheek, jaw |
| 27 | Mouth stretch | lip 1, 3, 2, 4 mouth 1, 2 | cheek, jaw |

generate a primary cue combination for AU1 as shown in Fig. 4.13 (d); when all four lip muscles are positive, 'lip 2' is less than 'lip 3, 4' and 'lip 3, 4' is less than 'lip 1', the four primary muscles generates a primary cues combination for AU27. These relations and uncertainties are systematically represented by a probabilistic framework presented in next chapter.

## 4.7   Summary

This chapter presents a facial expression representation system based on mass-spring system. The facial muscle dynamics model is physically-based and constructed from anatomical perspective, which is modeled by a nonlinear spring frame which can simulate the elastic dynamics of real facial skin. Based on the Lagrangian dynamics, facial tissue is deformed as the muscle force applying on it. Experimental results show the real-time face deformation process as well as realistic expression representation. Using our facial model, we can generate flexible and

(a) AU1

(b) AU5

(c) AU6

(d) AU12

(e) AU20

(f) AU27
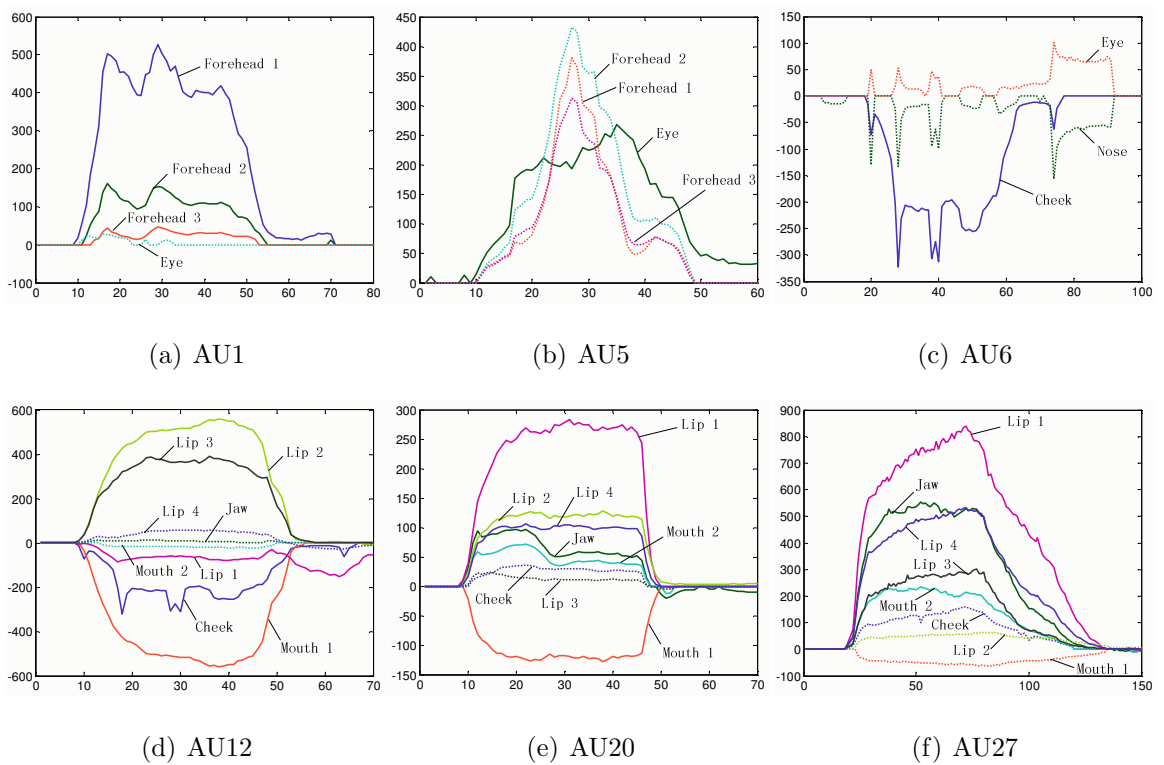
Figure 4.13: Facial muscle tracking curves showing detection AUs

realistic expressions.The biggest advantage of our expression modeling system is that it can analyze the relationship between the facial skin deformation and the in, side state, which is determined by facial muscle parameters. This enables us to predict deformation of the facial shape by detailed quantitative analysis of the relationship between facial muscles and facial skin deformation.

# Chapter 5

# Facial Expression Classification

Most research work on automated expression analysis perform an emotional classification. Once the face has been perceived and facial features have been extracted, the next step of an automated expression analysis system is to recognize the facial expression conveyed by the face. A set of categories of facial expression is defined by Ekman referred as the six basic emotions [40].

To classify the facial expressions automatically is still difficult due to some reasons. Firstly, there is no uniquely defined description either in terms of facial actions or in terms of some other universally defined facial codes. Secondly, it should be feasible to classify the multiple facial expressions. There are two common methods describing all visually distinguishable facial movements [40]. The first one is based on the integrated facial muscle motion, every available facial motion vectors, which are extracted from facial expressive model, are inputed into one classifier, then the output are the six basic emotions. The other one is based on AUs. This method needs to build two classifiers. Firstly, AUs are decided according to the combination of related muscles' deformations. Secondly, using the results from first classifier, the basic emotion is decided.

The neural network of multi-layer perceptrons (MLPs) is employed for static facial

expression classification.

## 5.1 Classifier - Multi-layer perceptrons

MLPs networks are general-purpose, flexible, nonlinear models consisting of a number of units at multiple layers. The complexity of the MLPs network can be changed by varying the number of layers and the number of units in each layer [88]. Given the hidden units and data, it has been shown that MLPs can approximate virtually any function to any desired accuracy [89]. MLPs are powerful tools when we has few prior knowledge about the relationship between input vectors and their corresponding outputs . Therefore, we use MLPs neural networ to classify different facial expression.
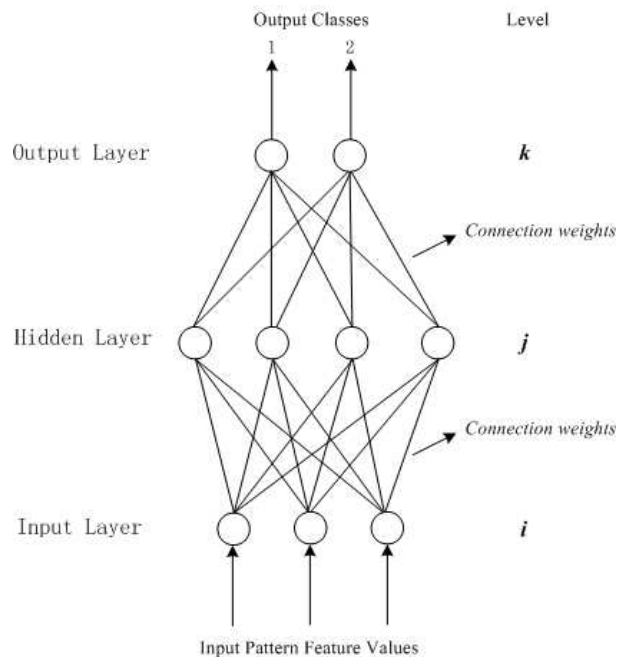


Figure 5.1: Architecture of multi-layer perceptron.

The neural network of multi-layer perceptrons consists of a network of processing

elements or nodes arranged in layers. Typically it requires three or more layers of processing nodes: an input layer which accepts the input variables (e.g. satellite channel values, GIS data etc.) used in the classification procedure, one or more hidden layers, and an output layer with one node per class (Fig. 5.1). The principle of the network is that when data from an input pattern is presented at the input layer the network nodes perform calculations in the successive layers until an output value is computed at each of the output nodes. This output signal should indicate which is the appropriate class for the input data i.e. we expect to have a high output value on the correct class node and a low output value on all the rest.

Each processing node in one layer is usually connected to the another node in the higher and lower layer. The connections carry weights which encapsulate the behavior of the network and are adjusted during training. The operation of the network consists of two stages. The "forward pass" and the "backward pass" or "back-propagation". In the "forward pass" an input pattern vector is presented to the network and the output of the input layer nodes is precisely the components of the input pattern. For successive layers the input to each node is then the sum of the scalar products of the incoming vector components with their respective weights. That is the input to a node $j$ is given by

$$input_j = \sum_i \omega_{ji} out_i \qquad (5.1)$$

where $\omega_{ji}$ is the weight connecting node $i$ to node $j$ and $out_i$ is the output from node $i$.

The output of a node $j$ is

$$output_j = f(input_j) \qquad (5.2)$$

which is then sent to all nodes in the following layer. This continues through all

the layers of the network until the output layer is reached and the output vector is computed. The nodeat input layer do not perform any of the above calculations. They simply take the corresponding value from the input pattern vector.

The function f denotes the activation function of each node. A sigmoid activation function is frequently used,

$$f(x) = \frac{1}{1 + exp(-x)} \qquad (5.3)$$

where $x = input_j$. This ensures that the node acts like a thresholding device.

The multi-layer feed-forward neural network is trained by supervised learning using the iterative back-propagation algorithm. In the learning phase a set of input patterns, called as the training set, are presented as feature vectors into the input layer , together with their corresponding desired output pattern which usually represents the classification results for the input patterns. Beginning with small random weights, for each input pattern the network is required to adjust the weights attached to the connections so that the difference between the network's output and the desired output for that input pattern is decreased. Based on this difference the error terms or $\delta$ terms for each node in the output layer are computed. The weights between the output layer and the layer below (hidden layer)are then adjusted by the generalised delta rule[90]

$$\omega_{kj}(t+1) = \omega_{kj}(t) + \eta(\delta_k out_k) \qquad (5.4)$$

where $\omega_{kj}(t+1)$ and $\omega_{kj}(t)$ are the weights connecting nodes $k$ and $j$ at iteration $(t+1)$ and $t$ respectively, $\eta$ is a learning rate parameter. Then the $\delta$ terms for the hidden layer nodes are calculated and the weights connecting the hidden layer with the layer below (another hidden layer or the input layer) are updated. This procedure is repeated until the last layer of weights has been adjusted.

The $\delta$ term in Eq. (5.4) above is the rate of change of error with respect to the input to node $k$, and is given by

$$\delta_k = (d_k - out_k)f'(input_k) \tag{5.5}$$

for nodes in the output layer, and

$$\delta_j = f'(input_k)\sum_k \delta_k \omega_{ki} \tag{5.6}$$

for nodes in the hidden layers, where $d_k$ is the desired output for a node $k$.

The back-propagation algorithm is a gradient descent optimization procedure which minimizes the mean square error between the network's output and the desired output for all input patterns $P$

$$E = \frac{1}{2P}\sum_p \sum_k (d_k - out_k)^2 \tag{5.7}$$

The training set is used to train the network iteratively until the set of weights is converged or the values of error function are reduced to an acceptable level. Fig. 5.2 shows the training procedure of the multi-layer feed-forward neural network. To measure the generalization ability of the multi-layer feed-forward neural network it is common to have a set of data to train the network and a separate set to assess the performance of the network during or after the training is complete. Once the neural network has been trained, the trained weights will be used in the classification phase. During classification, image data are fed into the network which performs the classification by assigning a class label to a pixel or segment in terms of the probability values computed at the output layer. Typically the output node is assigned by a class label which has the highest probability value.
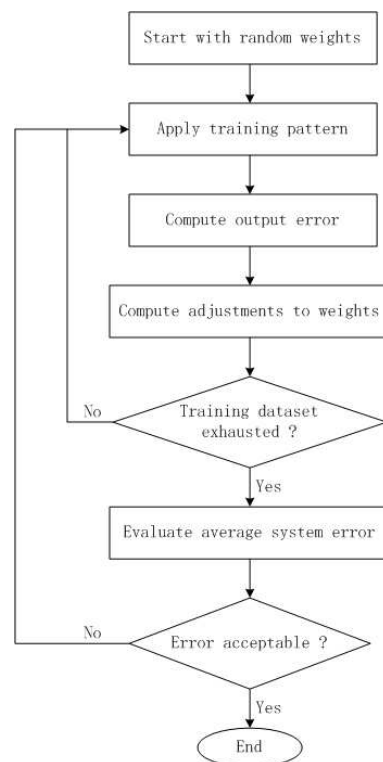
Figure 5.2: Training procedure for multi-layer perceptron network.

## 5.2   Integration-based approaches

In the system, facial expression recognition is formulated as a classification problem. The input for the classification module is a 22 dimension vector, and each element denotes the magnitude which has the largest absolute value during a facial expression. To classify the input vectors, we employ MLPs as the classifier, since it is able to construct arbitrary decision boundaries.

Generally speaking, the number of inputs to the network is determined by the number of functional muscles. Similarly, the number of outputs is equal to the number of emotion classes. The number of hidden nodes is a free parameter and its value depends on the complexity of the classification problem. We build the MLPs model as shown in Fig. 5.3. Fig. 5.4 shows the temporal dependencies by linking the node of in Fig. 5.3.
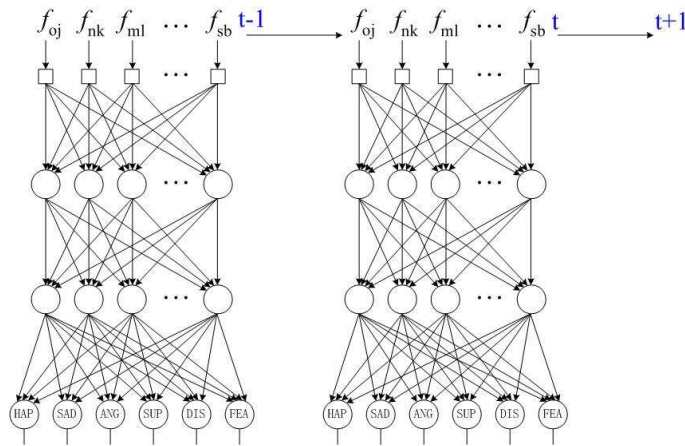


Figure 5.3: The MLPs model of six basic emotional expressions. Note: HAP − Happiness. SAD − Sadness. ANG − Anger. SUP − Surprise. DIS − Disgust. FEA − Fear. Other notations in the figure follow the same convention above.

The top level of layer in the model contains facial muscles information variables. All the nodes in this layer are observable.

The hidden layer is analogous to linguistic description of the relations between

hidden nodes and facial expressions. Each expression, which is actually an attribute node in the classification layer.

The classification layer consists of a class (hypothesis) variable including six states: happy, sadness, disgust, surprise, anger, and fear, respectively, and a set of attribute variables denoted as HAP, ANG, SAD, DIS, SUP, and FEA corresponding to the six facial expressions. The goal of this level of abstraction is to find the probability of class state ci, which represents the chance of class state ci given facial observations. When this probability is maximal, it has the largest chance that the observed facial expression belongs to the state of class variable ci.
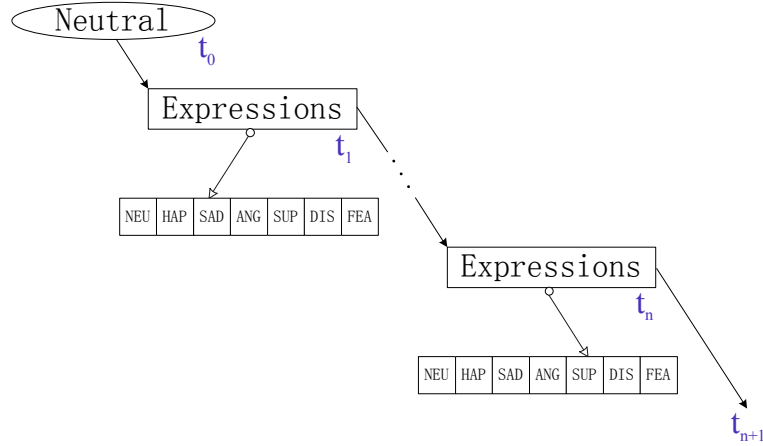


Figure 5.4: The temporal links of MLPs for modeling facial expression (two time slices are shown). Node notations are given in Fig. 5.3.

When used as pattern classifiers, MLPs networks represent the probabilities of the training data. We adopt the logistic activation function for each neuron.

$$y_j = \frac{1}{1 + \exp(-v_j)} \tag{5.8}$$

where $v_j$ is the induced local field (weighted sum of all synaptic inputs plus the bias) of neuron $j$, $y_j$ is the output of the neuron $j$.

During recognition, the feature vectors derived from the feature generation procedure form a vector sequence $F = \{\vec{f}_{fr1}, \vec{f}_{fl1}, \vec{f}_{fr2}, \vec{f}_{fl2}, \vec{f}_{fr3}, \vec{f}_{fl3}, \vec{f}_{er}, \vec{f}_{el}, \vec{f}_{nr}, \vec{f}_{nl},$ $\vec{f}_{cr1}, \vec{f}_{cl1}, \vec{f}_{cr2}, \vec{f}_{cl2}, \vec{f}_{mr1}, \vec{f}_{ml1}, \vec{f}_{mr2}, \vec{f}_{ml2}, \vec{f}_{jr}, \vec{f}_{jl}\}$ The network produces six outputs $y_{out,k}$, ($k = Happy, Sadness, Anger, Sunrise, Disgust, Fear$) The outputs are then normalized by a softmax function as follows

$$z_k = \frac{e^{\tilde{y}_{out,k}}}{\sum_{r=1}^{6} e^{\tilde{y}_{out,r}}}, k = Happy, Sadness, Anger, Suprise, Disgust, Fear \qquad (5.9)$$

where $tildey_{out,k} = \frac{y_{out,k}}{P(C_k)}$ represents the scaled outputs and $P(C_k)$ is the prior probability of class $C_k$. For MLPs, no output normalization is necessary because the outputs are always bounded between 0.0 and 1.0. Therefore, for MLPs, we used the scaled output for classification

$$z_k = \frac{y_{out,k}}{p(C_k)}, k = Happy, Sadness, Anger, Suprise, Disgust, Fear \qquad (5.10)$$

For networks with six outputs, happy, fear, sadness, surprise, disgust and anger, a typical class labeling rule is

$$E_k = \arg\max_k \{z_k\} \qquad (5.11)$$

where $E_k$ is the scaled output of the MLPs, $E_k \in [0, 1]$ is a decision threshold. Then, the decision criterion can be written as:

$$\text{If} z_l(\vec{x}) > \zeta, \vec{x} \in \text{the emotional class that correspond to } l \qquad (5.12)$$

A decision is made for each input vector, and the error rate is the proportion of incorrect labeling decisions to the total number of decisions.

In this study, we investigated MLPs with two hidden layers where the numbers of nodes in the hidden layers. Fig. 5.4 shows the temporal dependencies by linking

the node of in Fig. 5.3 .

## 5.3 Action units-based approaches

We build the MLPs model as shown in Fig. 5.5, which consists of two classifiers. In the context of expression classification, the number of inputs and outputs are same as previous. The number of hidden nodes is a free parameter and its value depends on the complexity of the classification problem. Table 5.1 contains the Facial Action Units (AUs) associated with facial expressions.

Table 5.1: The Association of Six Expressions to AUs

| Emotional Category | AUs |
|---|---|
| Happy | AU6, AU12 |
| Sadness | AU1, AU15, AU17, AU4, AU7 |
| Disgust | AU9, AU10, AU17, AU25 |
| Surprise | AU5, AU27, AU1, AU2 |
| Anger | AU4, AU7, AU9, AU17, AU23 |
| Fear | AU1, AU5, AU7, AU4, AU20 |

The top level of classifier in the model also contains facial muscles information and its output is the results of action units. The visual observations are the facial feature measurements as summarized in Table 4.2 and Table 4.3.

Then in the second classifier, the classification results of action units from the first classifier are the inputs, and the outputs are the facial expression results. The relation between AUs and facial expressions is based on Table 5.1. Each expression
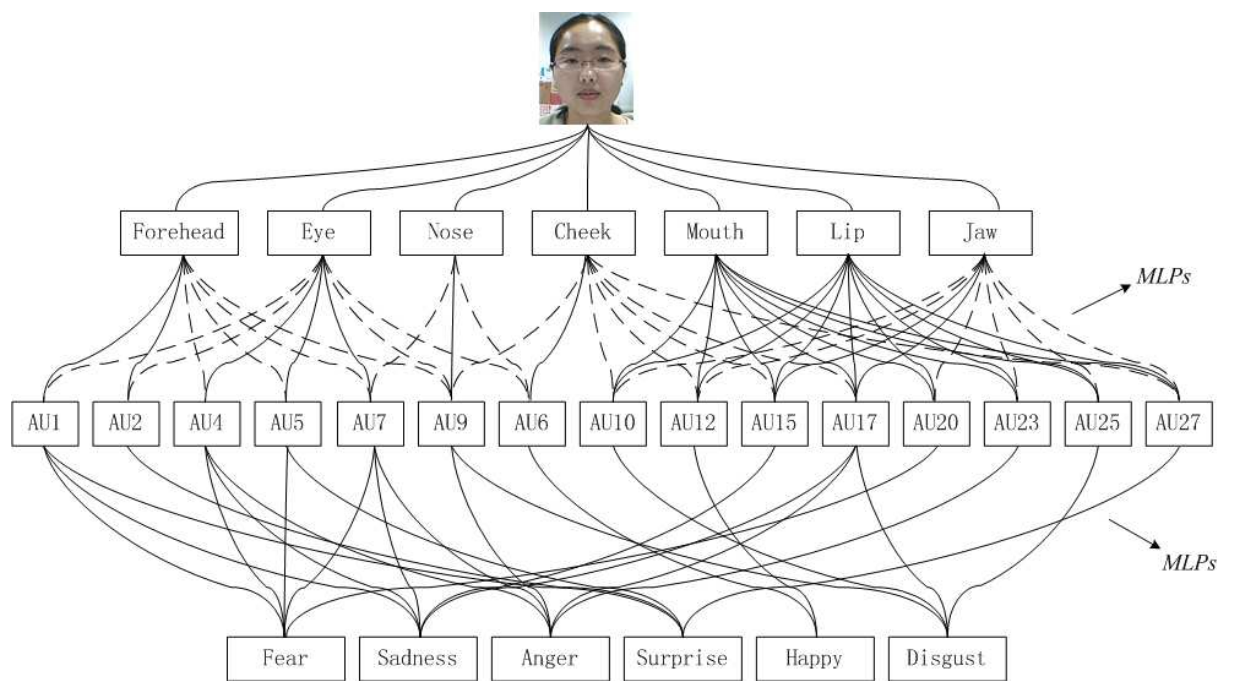
Figure 5.5: The concept links of the facial expression for interpreting an input face image.

category, which is actually an attribute node in the classification layer.

The classification layer consists of a class (hypothesis) variable including six states: happy, sadness, disgust, surprise, anger, and fear, respectively, and a set of attribute variables denoted as HAP, ANG, SAD, DIS, SUP, and FEA corresponding to the six facial expressions. The goal of this level of abstraction is to find the probability of class state ci, which represents the chance of class state given facial observations. When this probability is maximal, it has the largest chance that the observed facial expression belongs to the state of class variable.

## 5.4 Experiments and Discussions

In the system, the resolution of the acquired images is $320 \times 240$ pixel. The system is developed by using Microsoft Visual Studio . NET 2005. OpenCV [91] is employed to implement the module of face detection and key point extraction. To evaluate the system for facial expression recognition, we generate a total of 600 videos for six facial expressions (100 videos for each facial expression), namely happy, sad, fear, disgust, anger and surprise. In this work, one video corresponds to one facial expression and consists of an image sequence. All the facial videos are automatically captured from one person, since we do not touch the problem of face recognition. Then, all the data are divided into two groups randomly, 480 for training and 120 for testing. Thus we have 80 training data and 20 testing data for each facial expression class.

### 5.4.1 Facial expressions classification based on integration-based approaches

We create a short image sequence involving multiple expressions as shown in Fig. 5.6 (a). Each expression sequence began from a neutral face.For each sequence, we observe 100 frames. It can be seen visually that the temporal evolution of the expressions varies over time, exhibiting the spontaneous behavior. Fig. 5.6 (b) provides the analysis result by our facial expression model. The result naturally profiles the momentary emotional intensity and the dynamic behavior of facial expression that the magnitude of facial expression gradually evolves over time, as shown in Fig. 5.6 (a). Such a dynamic aspect of facial expression modeling can more realistically reflect the evolution of a spontaneous expression starting from a neutral state to the apex and then gradually releasing. Since there are interpersonal variations with respect to the amplitudes of facial actions, it is often difficult to determine the absolute emotional intensity of a given subject through machine extraction. In this approach, the belief of the current hypothesis of emotional expression is inferred relying on the combined information of current visual cues through causal dependencies in the current time slice, as well as the preceding evidences through temporal dependencies. Hence, as we can observe from the results, the relative change of the emotional magnitude can be well modeled at each stage of the emotional development; this is exactly what we want to achieve.

The accuracy of our facial expression model is also evaluated, as shown in Fig. 5.6. Here, we take this image set as an sequence showing that a subject poses different expressions starting from neutral states. Notice that, for this real-time sequence, we manually identify the pupil positions and our facial feature detection algorithm then detects and tracks the remaining features.
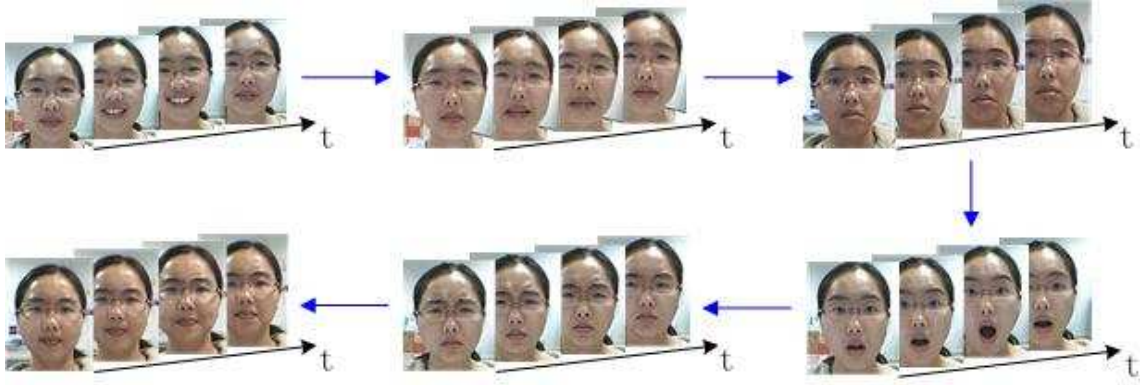
(a)
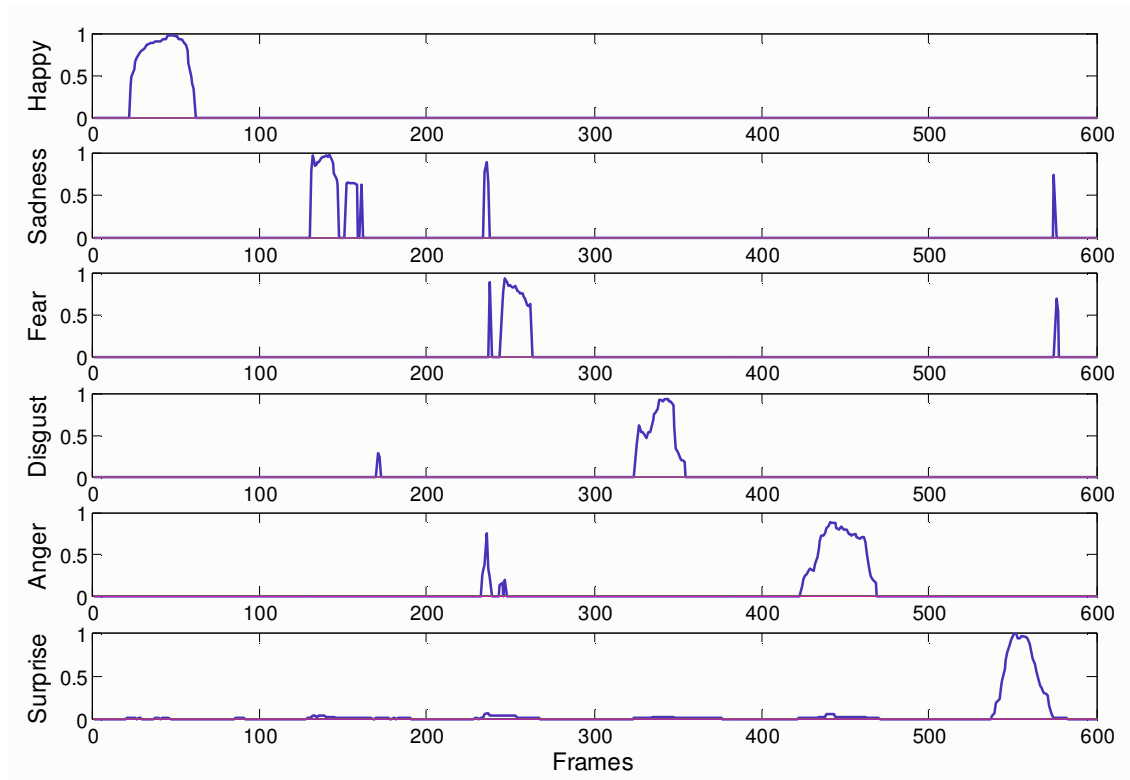


(b)

Figure 5.6: Real-time emotion code traces from a test video sequence: (a) Frames form the sequence; (b) Continuous outputs of each of the six expression detectors

Table 5.2: Emotion Classification Results Using Nonlinear Mode

| *Emotion* | Happiness | Sadness | Fear | Disgust | Anger | Surprise |
|---|---|---|---|---|---|---|
| Happiness | **0.842** | 0 | 0.126 | 0.032 | 0 | 0 |
| Sadness | 0.009 | **0.733** | 0.153 | 0.070 | 0.035 | 0 |
| Fear | 0.054 | 0.063 | **0.706** | 0.023 | 0 | 0.154 |
| Disgust | 0 | 0.173 | 0.076 | **0.616** | 0.135 | 0 |
| Anger | 0 | 0 | 0.005 | 0.133 | **0.862** | 0 |
| Surprise | 0 | 0 | 0.088 | 0 | 0 | **0.912** |

Table 5.3: Emotion Classification Results Using Linear Model

| *Emotion* | Happiness | Sadness | Fear | Disgust | Anger | Surprise |
|---|---|---|---|---|---|---|
| Happiness | **0.632** | 0.083 | 0.219 | 0.052 | 0.004 | 0.010 |
| Sadness | 0.038 | **0.570** | 0.186 | 0.113 | 0.093 | 0 |
| Fear | 0.051 | 0.141 | **0.498** | 0.020 | 0.013 | 0.277 |
| Disgust | 0 | 0.014 | 0.132 | **0.561** | 0.287 | 0.006 |
| Anger | 0 | 0.011 | 0.036 | 0.251 | **0.702** | 0 |
| Surprise | 0.023 | 0.039 | 0.122 | 0.021 | 0.004 | **0.791** |

To evaluate the accuracy of facial expression recognition, all the results are tabulated in Table 5.2. We set $\alpha = 1$ for the nonlinear model and $\alpha = 0$ for the linear model. As shown in Table 5.2 and 5.3, the system based on nonlinear mass-spring model achieved better performance for all the facial expressions than the linear model. In particular, nonlinear model achieved the significant improvements for happy, sad, fear and surprise compared with linear model. This indicates two folds: 1) Nonlinear mass-spring model is more reasonable for describing the movements of facial muscles compared with the linear model. 2) Our proposed novel features based on elastic forces derived from nonlinear spring model are effective for facial expression recognition.

## 5.4.2 Facial expressions classification based on action units-based approaches

Using the MPLs classifier introduced in 5.3 , we classify the action unites. Table 5.1 demonstrates the classification algorithms on the 2D embedding of the original data. The original data set are of $320 \times 240$ dimension, and the goal is to classify the action unites. To visualize the problem we restrict ourselves to the two features(2D embedding) that contain the most information about the class. The distribution of the data is illustrated in Table. 5.4 and 5.5.

Using action units-based approach, we also evaluate the accuracy of facial expression recognition, as shown in table 5.6 and 5.7. We set $\alpha = 1$ for the nonlinear model and $\alpha = 0$ for the linear model. As shown in table 5.6 and 5.7, the system based on nonlinear mass-spring model achieved better performance for all the facial expressions than the linear model. Compared with table 5.2 and 5.3, action units-abased approach combined with non-linear model achieve the best performance.

Table 5.4: Upper Face AUs Classification Results Using Nonlinear Model

| *AUs* | AU1 | AU2 | AU4 | AU5 | AU6 | AU7 |
|---|---|---|---|---|---|---|
| AU1 | **0.883** | 0.053 | 0 | 0.064 | 0 | 0 |
| AU2 | 0.112 | **0.781** | 0 | 0 | 0 | 0.107 |
| AU4 | 0.101 | 0.112 | **0.787** | 0 | 0 | 0 |
| AU5 | 0.085 | 0 | 0 | **0.786** | 0.065 | 0.064 |
| AU6 | 0 | 0.087 | 0 | 0 | **0.825** | 0.088 |
| AU7 | 0 | 0 | 0 | 0.115 | 0.096 | **0.789** |

Table 5.5: Upper Face AUs Classification Results Using Nonlinear Model

| *AUs* | AU9 | AU10 | AU12 | AU15 | AU17 | AU20 | AU23 | AU25 | AU27 |
|---|---|---|---|---|---|---|---|---|---|
| AU9 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AU10 | 0 | **0.880** | 0.087 | 0.033 | 0 | 0 | 0 | 0 | 0 |
| AU12 | 0 | 0.095 | **0.776** | 0.129 | 0 | 0 | 0 | 0 | 0.005 |
| AU15 | 0 | 0.056 | 0.070 | **0.772** | 0 | 0 | 0 | 0 | 0.102 |
| AU17 | 0 | 0 | 0.002 | 0.008 | **0.753** | 0.042 | 0.073 | 0.031 | 0.091 |
| AU20 | 0 | 0 | 0.082 | 0 | 0 | **0.747** | 0.060 | 0.093 | 0.018 |
| AU23 | 0 | 0 | 0 | 0 | 0.081 | 0.072 | **0.756** | 0 | 0.101 |
| AU25 | 0 | 0.002 | 0.003 | 0.048 | 0.054 | 0.028 | 0.015 | **0.833** | 0.017 |
| AU27 | 0 | 0.015 | 0.027 | 0.002 | 0.005 | 0.005 | 0.038 | 0.063 | **0.845** |

Table 5.6: Emotion Classification Results Using Nonlinear Mode

| *Emotion* | Happiness | Sadness | Fear | Disgust | Anger | Surprise |
|---|---|---|---|---|---|---|
| Happiness | **0.904** | 0 | 0.091 | 0.005 | 0 | 0 |
| Sadness | 0 | **0.821** | 0.127 | 0.033 | 0.019 | 0 |
| Fear | 0.010 | 0.036 | **0.860** | 0.007 | 0 | 0.087 |
| Disgust | 0 | 0.131 | 0.022 | **0.749** | 0.098 | 0 |
| Anger | 0 | 0 | 0.005 | 0.094 | **0.901** | 0 |
| Surprise | 0 | 0 | 0.065 | 0 | 0 | **0.935** |

Table 5.7: Emotion Classification Results Using Linear Model

| *Emotion* | Happiness | Sadness | Fear | Disgust | Anger | Surprise |
|---|---|---|---|---|---|---|
| Happiness | **0.715** | 0.062 | 0.183 | 0.034 | 0 | 0.006 |
| Sadness | 0.029 | **0.663** | 0.168 | 0.099 | 0.041 | 0 |
| Fear | 0.049 | 0.135 | **0.534** | 0.017 | 0.009 | 0.256 |
| Disgust | 0 | 0.009 | 0.106 | **0.631** | 0.251 | 0.003 |
| Anger | 0 | 0.008 | 0.025 | 0.216 | **0.751** | 0 |
| Surprise | 0.019 | 0.025 | 0.116 | 0.018 | 0 | **0.822** |

## 5.5   Summary

In this chapter, we present how to classify the facial expressions. We formulate the dynamic visual information fusion based on the Multi-layer Perceptrons(MLPs) for real-time facial expression recognition in video sequences and propose an efficient recognition scheme based on the detection of keyframes in videos.Both integration-based approach and action units-based approach are discussed.

# Facial Expression Imitation System in Human Robot Interaction

Facial expression recognition and imitation is an effective way for a social robot to understand human emotions and communicate with human beings, which plays a major role in human interaction and nonverbal communication. In order to build the effective communications between human and robots, an easy approach is to build up an expressive robotic face which can imitate human emotions.

## 6.1 Interactive Robot Expression Imitation System

As shown in Fig. 6.1, we build an interactive robot expression animation system which has the advantage of being especially designed for human robot interaction The experimental setup is depicted in Fig. 6.2. The input to the system is a video stream capturing the user's face.
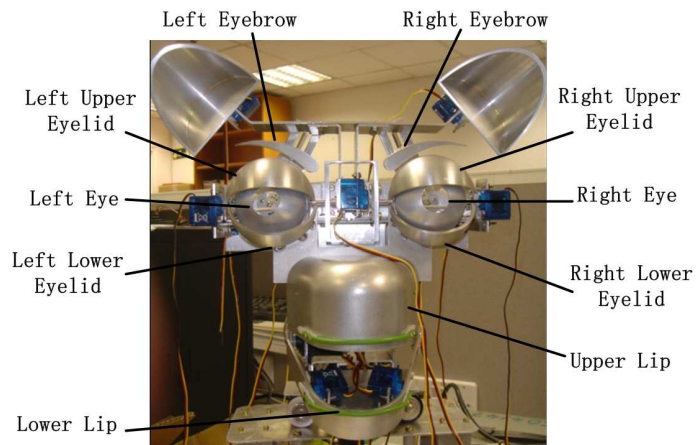
Figure 6.1: The robot head.

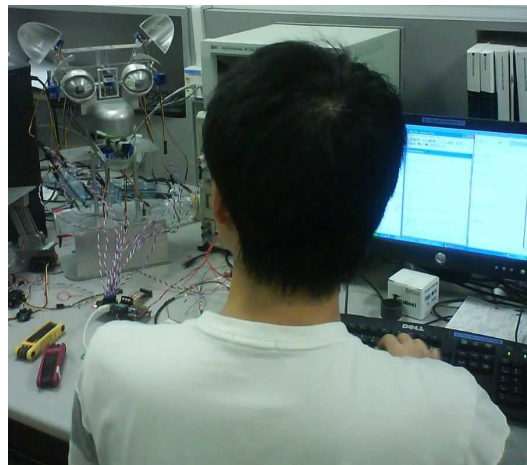

Figure 6.2: The experimental setup.

## 6.1.1    Expressive robotic face

The robot head consists of 16 Degrees of Freedom (DOF) to imitate the facial expressions. The development of the expressive robotic face is further sub-divided into:

- The mechanical design of the robotic face, including the various components of the robotic face, and the joint and motor placement which can be used to produce different facial expressions.

- The software control of the servo motors. The motors are controlled through the New Micros ServoPod, which provides the PWM signals to the 16 servo motors. Therefore we can use IsoMax, which is the New Micros operating system language, to implement the action units [92] or imitate the facial expressions by controlling those servo motors. For example, Mouth stretch can be imitated by controlling two servo motors of upper lip and lower lip.

A methodology for facial motion clone is developed, that is to copy a whole set of morph targets from a 2D real face image to an expressive robotic face. The inputs include two face images, one is in neutral position and the other is in a position containing some motion that to be animated, e.g. in a happy expression. The target face model exists at the neutral state. The goal is to obtain the target face model with the expression copied from the source face. Based on the feature tracking method we described before, the tester's facial features vector at the neutral state is subtracted from that at the expression. Therefore, the displacement and velocity information are extracted. They are multiplied by the a weight vector to reach the desired animation effects, e.g. exaggerated expression. The weight vector can be predefined according to the desired animation effects. Subsequently, the weighted vector is added on the face plane of the robot head in its neutral state. The robot head is able to show its emotions through an array of features situated in the frontal

part of the head. These are depicted in Fig. 6.3, and are shown in correspondence with the six universal expressions.



Figure 6.3: The robotic face is able to show its emotions through facial features situated in the frontal part of the head. The figure illustrates the features' configuration for each universal expression.

## 6.1.2 Generation of artificial facial expression

The facial expression generation is based on Ekman's six basic emotions(happiness, surprise, sadness, disgust, fear, anger) [40, 69]. In the system, the robot can imitate six human facial expressions plus the neutral state of no expressions.

In the system, the robot head are triggered to imitate human facial expressions by the emotion generator engine, and can generate vivid imitations according to the tester's facial expressions. For instance, our robot can imitate the happiness once it detects a facial expression of happiness. In this application, the robot is just used to imitate the human facial expression. Generally speaking, the response of the robot occurs slightly later than the apex of the human expression. In order to display simultaneously the correspondences between human and robot expressions

in the video, we put them side by side. In this case, we analyzed the contents of the video and commands with the facial expression code sent to the robot. Fig. 6.4 illustrates nine detected keyframes from the frame video. These are shown in correspondence with the robots response. The middle column shows the recognized expression. The right column shows a snapshot of the robot head when it interacts with the detected and recognized expression.

## 6.2 Summary

In this chapter, we describe the mechanism of our robot on imitating the facial expressions. The expressive robotic face includes a total of 16 Degrees of Freedom (DOF), whereby various emotions can be expressed in a way that an untrained human can understand and appreciate. From its concept design, the robotic face's affective states are triggered by the emotion generator engine. It's facial features can give a vivid animation according to the tester's expression. This occurs as a response to its internal state representation, captured through multimodal interaction (vision, audio, and touch). Experimental results show that our robot can imitate the human facial expressions effectively.

Figure 6.4: Left column: Some detected keyframes associated with the video. Middle column: The recognized expression. Right column: The corresponding robot's response.

# Chapter 7

# Conclusion and Future Work

## 7.1 Conclusions

This thesis investigates the problem of how to recognize and imitate the six kinds of human facial expressions. Recognizing the facial expressions has been a challenging problem due to the high degree of freedom of facial motions. In our work, two methods for integration-based approach and action units-based approach recognition are presented. Our methods can successfully recognize the static, track and identify dynamic on-line facial expressions of real-time video from one web camera. The face area is automatically detected and located by making using of face detection and skin hair color information. Our system utilizes a subset of Feature Points (FPs) for describing the facial expressions. 21 facial features are extracted from the captured video and tracked by optical flow algorithm.

In the system, nonlinear mass-spring model was employed to simulate twenty two facial muscles' deformations during facial expressions, and then the elastic forces of the facial muscles' deformation were taken as the novel features to be grouped

into a vector. Then such vectors were input into the module of facial expression recognition. The experimental results showed that our proposed nonlinear facial mass-spring model coupled with the MLPs classifier is effective to recognize the facial expressions compared with the linear mass-spring model.

We also incorporate facial expression motion energy to describe the facial muscle's tension during the expressions for person-independent tracking. It is composed by the expression potential energy and kinetic energy. The potential energy is used as the description of the facial muscle's tension during the expression. Kinetic energy is the energy which a feature point possesses as a result of facial motion. For each facial expression pattern, the energy pattern is unique and it is utilized for the further classification. Combined with the rule based method, the recognition accuracy can be improved for real-time person-independent facial expression recognition.

At the back end of the system, a social robot is designed to imitate the facial expressions. Experimental results of facial expression generation demonstrated that our robot can imitate six types of facial expressions effectively.

## 7.2 Future Work

There are a number of directions which could be done for future work.

1. Until now, there is no publication to explain how to estimate the model parameter $\alpha$ and $k_0$, investigate is still a problem in our future work. In our system, we currently can not could not evaluate the expression quality of the proposed robot head, so one possible solution is to investigate user's responses to the imitated facial expressions of the proposed robot.

2. In practice, six facial expressions are not enough to reflect human emotions.

For example, hot anger and cold are two different anger expressions. Thus we will define more facial expressions and improve our proposed system to accurately recognize and imitate more facial expressions in the future.

3. One direction to advance our current work is to combine the human speech and make both virtual and real robotic talking head for human emotion understanding and intelligent human computer interface, and explore virtual human companion for learning and information seeking.

# Bibliography

[1] C. C. Liu, P. Rani, and N. Sarkar, "Human-robot interaction using affective cues," *The 15th IEEE International Symposiun on Robot and Human Interactive Communication*, pp. 285–290, September 2006.

[2] S. S. Ge, "Social Robotics: Integrating Advances in Engineering and Computer Science," in *Proceedings of Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology International Conference*, (Chiang Rai, Thailand), pp. xvii–xxvi, May 9-12 2007.

[3] S. S. Ge, C. Wang, and C. C. Hang, "A facial expression imitation system in human robot interaction," in *will appear in The 17th International Symposium on Robot and Human Interactive Communication*, 2008.

[4] S. S. Ge, Y. Yang, T. H. Lee, and C. Wang, "Facial expression recognition and tracking based on distributed locally linear embedding and expression motion energy," *will appear in Journal of Intelligent Service Robotics -Special Issue*, 2008.

[5] L. Brethes, F. Lerasle, and P. Danes, "Data fusion for visual tracking dedicated to human-robot interaction," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, (Barcelona, Spain), pp. 2075–2080, April 2005.

[6] A. Jaimes and N. Sebe, "Multimodal humanccomputer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 116–134, 2005.

[7] T. Cootes, D. Cooper, C. Taylor, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, pp. 38–59, 1995.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *European Conf. on Computer Vision (ECCV)*, vol. 2, 1998.

[9] Y. Zhang, E. C. Pracash, and E. Sung, "A new physical model with multilayer architecture for facial expression animation using dynamic adaptive mesh," *IEEE Transactions on Visualization and Computer Graphics*, vol. 10, pp. 339–352, May/June 2004.

[10] B.Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[11] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 743–756, 1997.

[12] H. Hong, H. Neven, and C. V. Malsburg, "Online facial expression recognition based on personalized galleries," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition (FG'98)*, (Nara, Japan), pp. 354–359, April 1998.

[13] J. Steffens, E. Elagin, and H. Neven, "Personspotter-fast and robust system for human detection, tracking and recognition," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition (FG'98)*, (Nara, Japan), pp. 516–521, April 1998.

[14] I. Essa and A. Pentland, "Coding, analysis, interpretation and recognition of facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 757–763, 1997.

[15] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *IEEE Conference of Computer Vision and Pattern Recognition*, pp. 84–91, 1994.

[16] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998.

[17] S. McKenna, S. Gong, and Y. Raja, "Modelling facial colour and identity with gaussian mixtures," *Parttern Recognition*, vol. 31, pp. 1883–1892, December 1998.

[18] J. Daugman, "Complete discrete 2d gabor transform by neural networks for image analysis and compression," vol. 36, pp. 1169–1179, 1988.

[19] D. Pollen and S. Ronner, "Phase relationship between adjacent simple cells in the visual cortex," vol. 212, pp. 1409–1411, 1981.

[20] M. Bartlett, *Face Image Analysis by Unsupervised Learning and Redundancy Reduction*. PhD thesis, University of California, San Diego, 1998.

[21] W. A. Fellenz, J. G. Taylor, N. Tsapatsoulis, and S. Kollias, "Comparing template-base, feature-based and supervised classification of facial expressions

from static images," in *Proceedings of Circuits, Systems, Communications and Cmputers (CSCC'99)*, pp. 5331–5336, 1999.

[22] M. N. Dailey and G. W. Cottrell, "Pca gabor for expression recognition," Tech. Rep. CS1999-0629, 26, 1999.

[23] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, December.

[24] Z.Zhang, M. Schuster, and S. Akamatsu, "Comprison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *IEEE Proceeding of the Second International Conference on Automatic Face and Gesture Recognition (FG' 98)*, (Nara, Japan), pp. 454–459, April 1998.

[25] I. A. Essa and A. Pentland, "Facial expression recognition using a dynamic model and motion energy," in *Int. Conf. on Computer Vision (ICCV)*, pp. 360–367, 1995.

[26] K. Karpouzis, G. Votsis, and G. Moschovitis, "Emotion recognition using feature extraction and 3-d models," in *Proceedings of IMACS Internatioanl Multiconference on Circuits and Systems Communications and Computers (CSCC'99)*, (Athens, Greece), pp. 5371–5376, 1999.

[27] W. J. Hardcastle, *Physiology of Speech Production.* New York, NY: Academic Press, 1976.

[28] K. Mase, "Recognition of facial expression from optical flow," *Institute of electronics information and communication engineers Trans.*, vol. E74, pp. 3474–3483, 1991.

[29] Y. Zhang, E. Sung, and E. C. Prakash, "A physically-based model for real-time facial expression animation," in *Third International Conference on 3-D Digital Imaging and Modeling, 2001. Proceedings*, (Quebec City, Que., Canada), pp. 399–406, May 2001.

[30] J. Lien, *Automatic recognition of facial expression using hidden Markov models and estimation of expression intensity*. PhD thesis, The Robotics Institute, CMU, April 1998.

[31] Y.-L. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 97 – 115, February 2001.

[32] M. Wang, Y. Iwai, and M. Yachida, "Expression recognition from time-sequential facial images by use of expression change model," in *IEEE Proceedings of the Second International Conference on Automatic Face and Gesture Recognition (FG'98)*, (Nara, Japan), pp. 324–329, April 1998.

[33] T. Otsuka and J. Ohya, "Extracting facial motion parameters by tracking feature points," in *Proceedings of First International Conference on Advanced Multimedia Content Processing*, pp. 442–453, November 1998.

[34] M. Rosemblum, Y. Yacoob, and L. Davis, "Human expression recognition from motion using a radial basis function network architecture," *IEEE Transactions on Neural Networks*, vol. 7, no. 5, pp. 1121–1138, 1996.

[35] S. Kaiser and T. Wehrle, "Automated coding of facial behavior in human-computer interactions with facs," *Journal of Noverval Behavior*, vol. 16, no. 2.

[36] D. Messinger, A. Fogel, and K. L. Dickson, "What's in a smile," *Developmental Psychology*, vol. 35, no. 3, pp. 701–708, 1999.

[37] G. E. Schwartz, P. L. Fair, P. Salt, M. R. Mandel, and G. L. Klerman, "Facial expression and imagery in depression: An electromyographic study," *Psychosomatic Medicine*, vol. 38, pp. 337–347, 1976.

[38] P. Ekman, *Methods for Measuring Facial Actions. In K. R. Scherer and P. Ekman, editors.* Cambridge University: Handbook of Methods in Nonverbal Behaviour Research, 1982.

[39] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

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

[41] W. V. Friesen and P. Ekman, *Emotional Facial Action Coding System.* Unpublished manual, 1984.

[42] C. Izard, *Teh Maximally Descriminative Facial Movement Coding System (MAX).* PhD thesis, Instructional Resource Center, University of Delaware, Newark, Delaware, 1979.

[43] C. E. Izard, L. M. Dougherty, and E. A. Hembree, *A System for Indentifying Affect Expressions by Holistic Judgments*, 1983. Unpublished manuscript.

[44] R. Koenen, *Mpeg-4 Project Overview.* International Organisation for Standartistion, ISO/IECJTC1/SC29/WG11, La Baule, Octorber, 2000.

[45] N. Tsapatsoulis, K. Karpouzis, and G. Stamou, *A Fuzzy System for Emotion Classification based on the MPEG-4 Facial Definition Parameter.* European Association for Signal Processing (EUSIPCO), 2000.

[46] M. Hoch, G. Fleischmann, and B. Girod, "Modeling and animation of facial expressions based on bsplines," *The Visual Computer*, pp. 87–95, November 1994.

[47] W. V. Friesen and P. Ekman, *Dictionary - Interpretation of FACS Scoring*, 1987. Unpublished manuscript.

[48] P. Ekman, E. Rosenberg, and J. C. Hager, *Facial Action Coding System Affect Interpretation Database (FACSAID)*, July, 1998. http://nirc.com/Expression/FACSAID/facsaid.html.

[49] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. 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.

[50] J. Cohn, A. Zlochower, J.-J. J. Lien, and T. Kanade, "Automated face analysis by feature point tracking has high concurrent validity with manual facs coding," *Psychophysiology*, vol. 36, pp. 35 – 43, 1999.

[51] H. Kobayashi and F. Hara, "Dynamic recognition of basic facial expressions by discrete-time recurrent neural network," in *Proceedings of the International Joint Conference on Neural Network*, pp. 155–158, 1993.

[52] C. Padgett and G. Cottrell, *Representing face images for classifying emotions*, vol. 9. Cambridge, MA: MIT Press, 1997.

[53] J. Zhao and G. Kearney, "Classifying facial emotions by backpropagation neural networks with fuzzy inputs," *Proceedings of the International Conference on Neural Information*, vol. 1, pp. 454–457, 1996.

[54] M. Yoneyama, Y. Iwano, A. Ohtake, and K. Shirai, "Facial expression recognition using discrete hopfield neural networks," in *Proceedings of the International Conference on Image Processing (ICIP)*, vol. 3, pp. 117–120, 1997.

[55] G. Donato, S. Bartlett, C. J. Hager, P. Ekman, and J. T. Sejnowski, "Calssifying facial actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, 1999.

[56] S. Y. Kang, K. H. Young, and R.-H. Park, "Hybrid approaches to frontal view face recognition using the hidden markov model and neural network.," *Pattern Recognition*, vol. 31, pp. 283–293, Mar. 1998.

[57] I. Craw, D. Tock, and A. Bennett, "Finding face features," in *European Conf. on Computer Vision (ECCV)*, pp. 92–96, 1992.

[58] W. Keith, "A muscle model for animating three-dimensional facial expression," *Computer Graphics*, vol. 21, July 1987.

[59] K. Scott, D. Kagels, S. Watson, H. Rom, J. Wright, M. Lee, and K. Hussey, "Synthesis of speaker facial movement to match selected speech sequences," in *In Proc. 5th Australian Conf. on Speech Science and Technology*, 1994.

[60] C. Padgett, G. Cottrell, and B. Adolps, "Categorical perception in facial emotion classification," in *Proc. Cognitive Science Conf.*, vol. 18, pp. 249–253, 1996.

[61] M. J. Black and Y. Yacoob, "Recognizing facial expressions in image sequences using local parameterized models of image motion," *Computer Vision*, vol. 25, no. 1, pp. 23–48, 1997.

[62] F. Guan, L. Y. Li, S. S. Ge, and A. P. Loh, "Robust Human Detection and Identification by Using Stereo and Thermal Images in Human Robot Interaction," *International Journal of Information Acquisition*, vol. 4, no. 2, pp. 1–22, 2007.

[63] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun, "Towards robotic assistants in nursing homes: Challenges and results,," *Robotics and Autonomous Systems*, vol. 42, pp. 271–281, 2003.

[64] B. Scassellati, *Foundations for a theory of mind for a humanoid robot.* PhD thesis, Department of Electronics Engineering and Computer Science, MIT Press, Cambridge, MA, 2001.

[65] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, pp. 143–166, 2003.

[66] L. Canamero, "Emotional and intelligent ii: The tangled knot of social cognition," Tech. Rep. FS-01-02, AAAI Press, 2001.

[67] J. Cassell and et al, *Embodied Conversational Agents.* PhD thesis, MIT Press, Cambridge, MA, 1999.

[68] C. Breazeal and L. Aryananda, "Recognition of affective communicative intent in robot-directed speech," *Autonomous Robots*, vol. 12, pp. 83–104, 2002.

[69] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System.* Salt lake City, USA: A Human Face, 2002.

[70] R. C. Arkin, M. Fujita, T. Takagi, and R. Hasegawa, "An ethological and emotional basis for human-robot interaction," *Robotics and Autonomous System*, vol. 42, pp. 191–201, 2003.

[71] C. Breazeal, *Sociable machines: Expressive social exchange between humans and robots.* PhD thesis, Department of Electronics Engineering and Computer Science, MIT Press, Cambridge, MA, 2000.

[72] H. W. Jung, Y. H. Seo, M. S. Ryoo, and H. S. Yang, "Affective communication system with multimodality for a humanoid robot, ami," in *International Conference on Humanoid Robots*, pp. 690–706, 2004.

[73] A. Takanishi, "An anthropomorphic robot head having autonomous facial expression function for natural communication with human," in *9th International Symposium of Robotics Research*, pp. 197–204, 1999.

[74] G. Park, S. Lee, W. Y. Kwon, and J. B. Kim, "Neurocognitive affective system for an emotive robot," in *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, (Beijing, China), pp. 2595–2600, October 2006.

[75] R. Gockley, J. Forlizzi, and R. Simmons, "Modeling affect in socially interactive robots," in *The 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06)*, (Hatfield, UK), pp. 558–563, September 2006.

[76] F. I. Parke, "Parameterized models for facial animation," *IEEE Computer Graphics and Applications*, vol. 2, pp. 61–68, Nov. 1982.

[77] P. Ekman and R. J. Davidson, *The Nature of Emotion Fundamental Questions.* New York: Oxford Univ. Press, 1994.

[78] M. H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 34–58, Jan. 2002.

[79] C. Harris and M. Stephens, "A combined edge and corner detector," in *Proc. of the 4th Alvey Vision Conference*, pp. 147–151, 1988.

[80] D. Williams and M. Shah, "Edge characterization using normalized edge detector," *Computer Vision, Graphics and Image Processing*, vol. 55, pp. 311–318, July 1993.

[81] K. Hotta, "A robust face detection under partial occlusion," in *Proc. of Int. Conf. on Image Processing*, pp. 597–600, 2004.

[82] W. J. Lipham, *Cosmetic and Clinical Applications of Botulinum Toxin*. 6900 Grove Road Thorofare USA: SLACK Incorporated, 2004.

[83] J. C. Hager, P. Ekman, J. T. Cacioppo, and R. E. Petty, *The Inner and Outer Meanings of Facial Expressions*. New York, USA: The Guilford Press, 1983.

[84] P. L. Williams, R. Warwick, M. Dyson, and L. H. Bannister, *Greys Anatomy*. London: Churchill Livingstone, 1989.

[85] Y. Zhang, E. C. Prakash, and E. Sung, "A physically-based model with adaptive refinement for facialanimation," in *Computer Animation, 2001. The Fourteenth Conference on Computer Animation. Proceedings*, (Seoul, South Korea), pp. 28–39, November 2001.

[86] Y. M. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 699–714, May 2005.

[87] J. Bassili, "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.

[88] I. Kanellopoulos, *Use of Neural Networks for Improving Satellite Image Processing Techniques for Land Cover/Land Use Classification.* European Commission, Joint Research Centre. http://ams.egeo.sai.jrc.it/eurostat/Lot16-SUPCOM95/final-report.html.

[89] J. E. Dayhoff, *Neural Network Architectures.* New York: Van Nostrand Reinhold, 1990.

[90] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," *Distributed Processing. Explorations in the Microstructures of Cognition*, vol. 1, pp. 318–362, 1988.

[91] Intel Corporation, *OpenCV Reference Manual*, 2001. `http://www.intel.com/technology/computing/opencv/index.htm`.

[92] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 699–714, 2005.