MIT Media Laboratory, Perceptual Computing Techincal Report #303

Analysis, Interpretation and Synthesis of Facial Expressions

by

Irfan Aziz Essa

B.S., Illinois Institute of Technology (1988)
S.M., Massachusetts Institute of Technology (1990)

Submitted to the Program in Media Arts and Sciences,
MIT Media Laboratory,
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1995

Author ....
Program in Media Arts and Sciences
September 20, 1994

Certified by ....
Alex P. Pentland
Professor of Computers, Communication and Design Technology
Program in Media Arts and Sciences, MIT
Thesis Supervisor

Accepted by ....
Stephen A. Benton
Chairman, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Analysis, Interpretation and Synthesis of Facial Expressions

by

Irfan Aziz Essa

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on September 20, 1994, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis describes a computer vision system for observing the "action units" of a face using video sequences as input. The visual observation (sensing) is achieved by using an *optimal estimation* optical flow method coupled with a geometric and a physical (muscle) model describing the *facial structure*. This modeling results in a time-varying spatial patterning of facial shape and a parametric representation of the independent muscle action groups responsible for the observed facial motions. These muscle action patterns are then used for analysis, interpretation, recognition, and synthesis of facial expressions. Thus, by interpreting facial motions within a physics-based optimal estimation framework, a new control model of facial movement is developed. The newly extracted action units (which we name "FACS+") are both physics and geometry-based, and extend the well known FACS parameters for facial expressions by adding temporal information and non-local spatial patterning of facial motion.

Thesis Supervisor: Alex P. Pentland
Title: Professor of Computers, Communication and Design Technology
Program in Media Arts and Sciences, MIT

Doctoral Committee

Thesis Advisor ....

Alex P. Pentland
Professor of Computers, Communication and Design Technology
Program in Media Arts and Sciences, MIT

Thesis Reader ....

Norman Badler
Professor of Computer and Information Science
University of Pennsylvania

Thesis Reader ....

Alan Yuille
Professor of Applied Sciences
Harvard University

Acknowledgments

This thesis and the research that it describes could not be accomplished had it not been for the guidance and aid of multitude of individuals. Many of you have had a significant influence on me during my time at MIT, in a variety of ways, both academic and personal. To all of you (and you all know who you are) I express my sincere gratitude, and I hope that I can repay you in some small ways as I am able. I would like to single out the following people who had a major impact on me in the last few years.

First, and perhaps foremost, I thank Sandy Pentland, my advisor and my mentor, for providing invaluable guidance, advice, support and criticism since my first arrival at MIT six years back. It was because of him that my graduate studies were so enjoyable and so intellectually rewarding. He has provided a good balance of freedom and interest, while teaching me not only how to do research, but how to write papers and give talks. I am extremely impressed by his ability to get an intuitive feel of the most difficult of ideas. I hope some of his talents have rubbed off on me.

I would also like to thank the other members of my thesis committee. I am glad I was able to persuade both Norm Badler and Alan Yuille to be on my committee. I am grateful for their thorough scrutiny of my work. Norm has also been a great source of information and his ideas and suggestions have really helped me in my research. Alan has also kept himself available to me despite his travels and has been encouraging in the importance of the work I undertook.

In addition my committee, I have to thank some additional professors around MIT. John Williams has been a good friend and a source of encouragement and support. He encouraged me to get more and more involved with the work at the MIT Media Lab, leading to my eventual transfer to the Media Lab. Special thanks to Marc Raibert for convincing me that I should look for a research direction that I really enjoy and feel comfortable with.

My years at MIT and the MIT Media Lab have been priceless. I consider myself lucky to be a part of this institute and the lab. The people I have met over the years have made the experience unique and one that I will remember for ever. I am especially proud of the fact that I am a member of the Vision and Modeling group (vismod) and that I got an opportunity to spend such quality time with all the vismoders. Special thanks go to Ted Adelson, Aaron Bobick and Roz Picard who (with Sandy) have made vismod a top-notch research group over the years. Also thanks to Ted, Aaron and Roz for their suggestions and ideas, not only relating to research, but in a multitude of things.

Special thanks to Trevor Darrell, with whom I worked on developing a real-time facial tracking system. I enjoyed working with you and I hope we continue doing so. Thanks to Eero Simoncelli for lending me his ideas (and code) to start thinking about how optical flow can be used to observe facial motion, and thanks to John Wang, whose code I initially used to get a feel of what I wanted to do. Also thanks to folks at U. Penn., namely Catherine Pelachaud and Steve Platt for sharing with me their facial model and to Stevie Pieper for sharing his expertise (and data) for facial modeling. Thanks also to Stan Sclaroff for many enlightening discussions.

Almost all the members of the Vision and Modeling group have somehow or the other helped me during the course of this research for which I am really grateful. I have really enjoyed working with you all. Special thanks go to Lee and Claudio have been were very kind to help me videotape subjects and in digitizing. Martin Friedmann has been a great source of help with computing and Brad Horowitz‘s soul still lives in some of my initial code. Special thanks to my office mates over the years, Matthew, Lee, Alex and Baback for bearing with me. Special thanks also to some other members of vismod, namely Thad, Andy, Stephen, Dave, Sourabh, Yair, Ali, Fang, Ken, Tom, and all the rest of you for everything.

Special thanks go the administrative staff of Vismod. Laureen has always been the greatest. Laurie has helped in many ways. Judy has helped out in so many ways that I cannot count. It suffices to say that she has been a friend and a supporter. I am also grateful for her invaluable help in making most of my written work (including this thesis) more readable. Thanks also to Linda Peterson and Santina Tonelli for reminding me of all the academic deadlines and other related discussions. Thanks also to Greg Tucker and Ben Lowengard for their help.

Also, thanks to all the folks who agreed to make facial expressions on tape for me. They will be preserved for posterity, especially all the all their funny expressions.

A hearty thanks to the most important people in my life, my family. If it were not for the caring, love, affection and continuous support of my mom and dad, my sisters and brother, none of this would be at all possible. Special thanks to Uncle Ali, who is no longer with us for convincing me to go for a Ph .D. A Special thanks also to my friends Amber, Shahab, Khursheed, Saeed and Kishwar.

Special thanks to a very special person who carries much meaning in my life. A person who has been there for me when I needed her and borne all my mood swings due the frustrations that come with being a researcher and Ph. D. Candidate A person that has been a supporter, a friend and a constant source of encouragement. Yes, I do mean my dear wife, Shani. She is my love and my best freind, and if she was not there telling me I could do it, this thesis might not have become a reality.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

It was only by the bumpy plane of pinkish tissue, surrounded and tufted with hair, in which the eyes were situated, that this creature wished to be judged by, or through which it was exposed.

Doris Lessing, *The Four Gated City*, p 519 [49]

1.1 Face perception and face processing

The communicative power of the face makes it a focus of attention during social interaction. Facial expressions and the related changes in facial patterns inform us of the emotional state of people around us and help to regulate both social interaction and spoken conversation. This expressive nature of faces is illustrated in Figure 1-1 showing a person expressing *anger*, *disgust*, *fear*, *happiness*, and *sadness*. To fully understand the subtlety and informativeness of the face, considering the complexity of the movements involved, one must study face perception and the related information processing.

For this reason *face perception* and *face processing* have become major topics of research by cognitive scientists, sociologists and most recently by researchers in computer vision and computer graphics. The field of *perceptual computing*, which brings together computer and cognitive scientists, has special interests in this kind of work and forms the context of research described here.

Clearly, the automation of human face processing by a machine/computer will be a significant step towards developing an effective human-machine interface. We must consider the ways in which a system with the ability to understand facial gestures (*analysis*), and the means of automating this interpretation and/or production (*synthesis*) might enhance human-computer interaction (HCI). It is this analysis and synthesis, relating computer vision and computer graphics, that forms the premise of this dissertation and is described in detail in this document.

Why is this analysis and synthesis important? Figure 1-1 shows a person expressing *anger*, *disgust*, *fear*, *happiness*, and *sadness*. While we may all agree with the descriptions of expressions shown in these photographs, no specific and elaborate attempt has been made at an automatic analysis to achieve this type of expression categorization.

This thesis presents a method that combines dynamic visual estimation with active modeling and control of the facial structure to form an efficient framework for improved analysis, interpretation and synthesis of facial expressions. This method involves observing facial expressions in both temporal and spatial domains, then using dynamic control to estimate and correct the observation signal, and then finally by using this signal for both analysis and synthesis. This method provides a tool for extracting an extended *Facial Action Coding System (FACS)* model (FACS+) using a physics-based model of both skin and muscle, driven by optical flow. This method is capable of very detailed analysis in both time and space, with improved accuracy providing the required information to observe coarticulation of expressions resulting in improved modeling of facial motion. The results of the analysis/modeling process can then be used for interpretation and categorization of facial expressions and for very detailed synthesis of facial motion.

Thesis Outline

In this thesis, first some of the applications of machine-based facial interpretation, analysis, recognition and synthesis are presented. Then the goals and the contributions of the research described in this thesis are elucidated. Then follows a discussion of the relevant literature and background (Chapter 2). This discussion crosses the domains of psychology and cognitive sciences, computer graphics, computer vision and human-machine interaction. Chapters 3, 4, and 5 describe the quantitative and qualitative details of visual sensing, facial modeling and dynamic estimation. This is followed by analysis of the data acquired using our experiments (Chapter 6) and details of the experiments and results (Chapter 7). Finally, the last two chapters discuss the limitations, suggestions for future work, and the conclusions of this research.

1.2 Facial Expressions and Machines

Of all the nonverbal behaviors – body movements, posture, gaze, voice, *etc.,* – the face is probably the most accessible “window” into the mechanisms which govern our emotional and social lives. The current technological developments provide us with the means to develop automated systems for monitoring facial expressions and animating synthetic facial models. Face processing by machines could revolutionize fields as diverse as medicine, law, communications, and education [29]. This progress would make it feasible to automate many aspects of face processing that humans take for granted (face recognition, expression and emotion recognition, lip reading, *etc.*), and to develop new technological aids (robotics, man-machine systems, medical, teleconferencing, *etc.*).

Realistic animation of faces would serve a major role in bridging the gap between man and machine. Computers with animated faces could be used in classrooms to teach children. Machines that would know how to express emotions would be instrumental in a establishing a completely new paradigm for man machine interaction. Machines that can recognize expressions will be able to relate to the emotion and feeling of the user. A machine that can both model and recognize expressions, will be one step closer to having a virtual persona.

Animation and synthesis of facial expressions also has applications out of the realm of human-machine systems. It can be used to generate 3-D synthetic actors that would have expressions and emotions to reflect the context of the story and the environment they reside in. Such animations (sometimes exaggerated to reflect the story-telling context) are aimed at establishing an emotional relationship with the audience.

Facial motion analysis can be applied to applications of reading lips. It can be used to compliment speech recognition. On its own it would be a great resource for the hearing impaired.

An important application, perhaps one that is addressed effectively in recent years is a video-phone or a teleconferencing application. It is argued that due the limits on the bandwidth of transmissions, an efficient form of model-based coding of facial expression information and its transmission it to another location is required. Additionally, concepts of telepresence and virtual offices would become possible as one could sit in one continent, carry on a meeting with different people in different continents and still be able to observe each and every facial gesture.

In basic research on brain, facial expressions can identify when specific mental processes are occurring. Computers can become useful tools for such studies. Such processing can also be used towards understanding emotion and the related facial expressions.

Facial expressions also hold promise for applied medical research, specially in cases of analyzing the psychological state of the patient. Detailed facial modeling can be used to visualize faces for biomedical applications. Several researchers have used 3-D biomechanical models for pre/post surgery simulations and surgical path planning.

1.3 Thesis Overview

This thesis research consists of the following major elements:

Visual Sensing: Using computer vision techniques to determine parameters of expressions by estimating the pattern changes, evolving over time, of a face in a sequence of images. This involves observing image sequences and calculating the temporal change (*i.e.,* computing motion).

Modeling and Graphics: Defining a computer graphics model to describe a face as a geometric shape to which the above change in facial pattern can be applied.

Physically-based (Anatomical) Modeling: Extending this geometric model to be a physically-based model of the face, with dynamic properties. This includes adding muscle information and other biomechanical constraints to the geometric model. Extra details are added in regions where flow computation

Figure 1-1: *Various Expressions, considered to be the universal expressions.*

has a higher variance of error. A multi-grid mesh is developed to account for low frequency and high frequency motions separately.

Dynamic Estimation and Control: Devising a dynamic estimation and control loop to correct the estimates of the feature change, based on the constraints of the dynamic system and the error covariance of the temporal changes in images. This will correct the behavior of the dynamic model and allow estimation and computation of muscle activation (*i.e.,* facial control input) from visual observations.

Analysis and Identification: Establishing the relationship between visual input and the physical model in terms of a set of orthogonal, but time-varying *basis* parameters and using these basis parameters to determine a set of control parameters for specific muscle group actuations. This basis decomposition will be accomplished by application of principle component analysis and similar statistical analysis techniques. Using the dynamics of the underlying biological model gives credence to the time variations in these parameters. This newer set of parameters forms the extended FACS model, (named FACS+). This FACS+ model is then used for analysis, recognition and synthesis of facial expressions.

Synthesis and Real-time Simulation: The determined control parameters of muscle actuations for facial expression provide a set of proper "*control knobs*" for synthesis. These are then used for real-time facial tracking and animation.

This thesis describes a method that uses an optical flow method to extract information about facial motion by observing the changes in facial patterns over time at varying levels of detail. These image observations are related to a dynamic and physical facial model to acquire *model-based* observations. Using these observations within an active control loop, the dynamics of the facial structure are estimated and updated. This analysis provides an anatomically plausible model for different expressions. Hence the main contribution of this thesis is to achieve an automated understanding (*analysis*) of facial gestures, from which the detailed interpretation (FACS+) of facial gestures is acquired. Another contribution of this thesis is that it describes a system in which a dynamic model of the face is actively updated for vision observations within a dynamic modeling, estimation and control framework.

Chapter 2

Background

Certain physical acts are peculiarly effective, especially the facial expressions involved in social communication; they affect the sender as much as the recipient.

Marvin Minsky, *The Society of Mind* (p 44) [60]

2.1 Introduction

The human face provides information that regulates a variety of aspects of our social life. Facial expressions and gestures inform us of the emotional state of our companions. Verbal and non-verbal communication is aided by our perception of facial motion; visual speech effectively compliments verbal speech. Additionally, the face serves to identify its owner. In short, faces are accessible "windows" into the mechanisms that govern our emotional and social lives.

There have many attempts to understand how meaning is derived from the complex rigid and nonrigid motions associated with the face. In this chapter we will review some of the previous work done in the field of face perception starting with a brief overview of the psychology literature. Then we discuss some of the work in facial modeling and animation. The latter part of this chapter will concentrate on face processing by computers, with special attention to the contributions of our work with reference to some of the existing literature. A good source of information on current research in facial modeling and analysis is the Mosaic WWW server set up at University of California at Santa Cruz: `http://mambo.ucsc.edu/`.

2.2 Psychological Background

The face is a multisignal, multimessage response system capable of tremendous flexibility and specificity. It is the site for sensory inputs and the communicative outputs. Faces convey information via four general classes of signals [25]:

static facial signals: permanent features of the face like the bony structure and soft tissue masses contributing to facial appearance

slow facial signals: changes in facial appearance over time, wrinkles, texture, *etc.,*

artificial signals: exogenously determined features, such as eyeglasses, and cosmetics, AND,

rapid facial signals: phasic changes on neuromuscular activity leading to visible changes in facial appearance.

All four of these classes contribute to facial recognition, however only the rapid signals convey messages (via emotions) in social context. The neuropsychology of facial expression supports the view that facial movements express emotional states, and that the two cerebral hemispheres are differently involved in control and interpretation of facial expression. Some of the initial work studying the relationship between facial expression and emotion was undertaken by Duchenne [24] and Darwin [22] in the early nineteenth century. Their work still has a strong influence on the research techniques used to examine expression perception. To date, the majority of studies on facial expressions have examined the perception of posed expressions in static photographs like those shown in Figure 1-1. Most of these studies suggest seven *"universal"* categories of expressions that can be discriminated by members of all cultures, both literate and preliterate [13, 27].

Researchers are now beginning to study facial expressions in spontaneous and dynamic settings to avoid the potential drawbacks of using static expressions and to acquire more realistic samples. The problem, of course, is how to categorize active and spontaneous facial expressions in order to extract information about the underlying emotional states. We present a brief overview of the predominant theories used to address this problem.

Measurement of Facial Motion

To categorize expressions, we need first to determine the expressions from facial movements. Ekman and Friesen [28] have produced a system for describing all visually distinguishable facial movements. The system, called the *Facial Action Coding System* or *FACS*, is based on the enumeration of all "action units" of a face that cause facial movements. As some muscles give rise to more than one action unit, the correspondence between action units and muscle units is approximate. Ekman *et al.* [28] give as an example the frontalis muscle, used for eyebrow raising action – this is separated into two action units depending on whether the inner or outer part of the muscle causes the motion. Figure 2-1 shows examples of action units involved in simple eyebrow motions. There are 46 AUs in FACS that account for changes in facial expression and 12 AUs that describe changes in head orientation and gaze. FACS coding is done by individuals trained to categorize facial motion based on anatomy of facial activity, *i.e.,* how muscles singly and in combination change the facial appearance. A FACS coder "disects" an expression, decomposing it

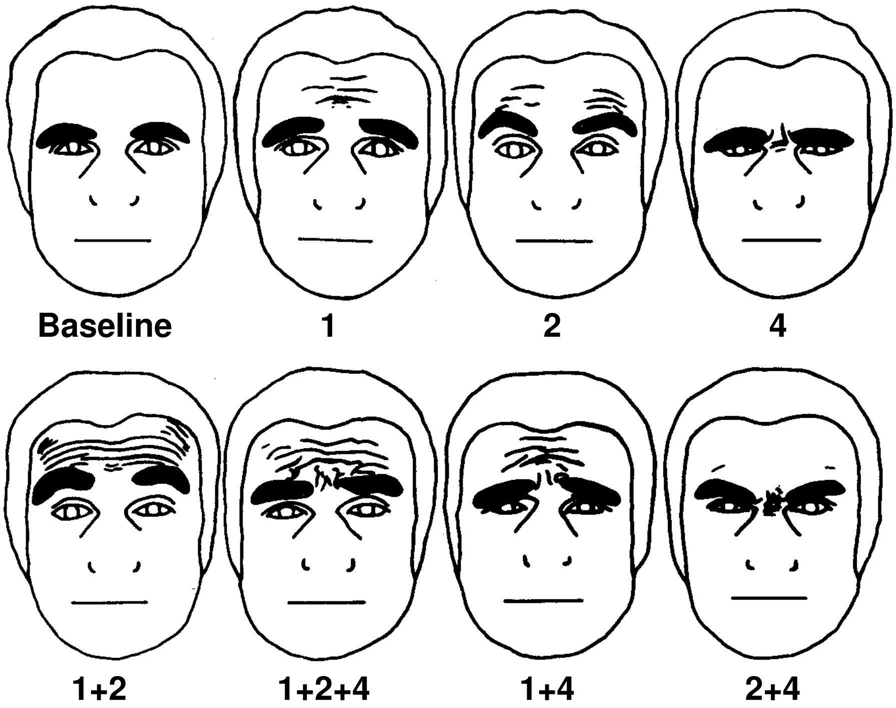


Figure 2-1: *The different action units for the brow and forehead identified by the FACS [28] model. Action units 1, 2, and 4 may occur alone (top) or in combination (bottom). Different action units are associated with different expressions. For example, action unit 1 indicates sadness, but 1 with 2 indicates surprise. This figure is obtained by tracing photographs and is included here from Bruce [13].*

into specific AUs that produced the motion. The FACS scoring units are descriptive, involving no inferences about emotions. Using a set of rules, FACS scores can be converted to emotion scores to generate a FACS' emotion dictionary.

Another method for measuring visible appearance changes in a face is *Maximally Discriminative Affect Coding System (MAX)* by Izard [42, 43]. MAX's units are formulated in terms of appearances that are relevant to the eight specific emotions, rather than in terms of individual muscles. All the facial actions that MAX specifies as relevant to particular emotions are also found in FACS emotion dictionary, hence making MAX a subset of the FACS model. This reason has led to more applications of the FACS model (especially outside of psychology) than MAX.

Emotion and Facial Motion

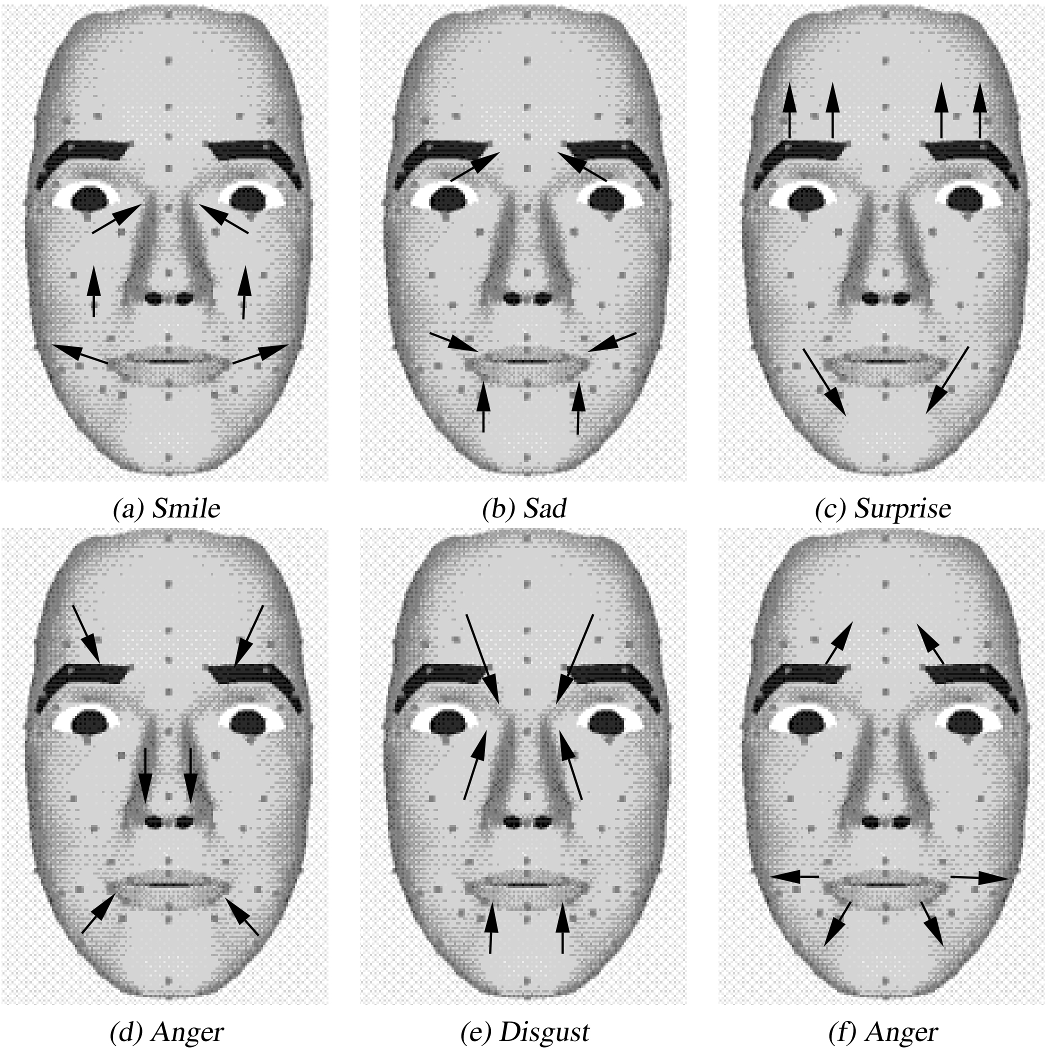
The validity of FACS as an index of facial emotion has been demonstrated in a number of studies [27]. Unfortunately, despite efforts in the development of FACS as a tool for describing facial motion, there has been little exploration of whether "action units" are the units by which we categorize expressions. Supporters of the FACS model claim (in [26, 27]) that emotions that are often confused with one another are those that share many action units. However, the literature contains little systematic investigation of comparisons between different bases for description of postures or relative positions of facial features [73, 35, 13].

Emotion recognition requires delineating the facial patterns that give rise to the judgment of different emotions. It involves the description of information in the face that leads observers to specific judgments of emotion. The studies based on the above methods of "coding" expressions are performed by presenting subjects with photographs of facial expressions and then analyzing the relationship between components of the expressions and judgments made by the observers. These judgment studies rely on static representations of facial expressions. The use of such stimuli has been heavily criticized since "judgment of facial expression hardly ever takes place on the basis of a face caught in a state similar to that provided by a photograph snapped at 20 milliseconds" [16]. The feature-based descriptions derived from static stimuli ignore several levels of facial information relevant to the judgment of emotions. One of these levels is the rate at which the emotion is expressed. Another level is related to the structural deformation of the surface of the face. Bassili [6] argues that because facial muscles are fixed in certain spatial arrangement, the deformations of the elastic surface of the face to which they give rise during facial expressions may be informative in the recognition of facial expressions.

Bassili [7] conducted experiments by covering faces of actors with black makeup and painting white spots in random order over it. Faces were divided into upper and lower regions (to corollate with FACS data for upper and lower regions) and recognition studies were conducted. This study showed that in addition to the spatial arrangement of facial features, movement of the surface of the face does serve as a source of information for facial recognition. Figure 2-2 shows 3-D model of a face with some typical facial motions marked for the six expressions. These are the same motions that Bassili used with good results in his studies.

This study suggests that the use of such "frozen" action descriptions as proposed by the FACs model are unsatisfactory for a system developed to code *movements*. The lack of temporal and detailed spatial (both local and global) information is a severe limitation of the FACS model (see [29]). Additionally, the spatial arrangement of facial features also suggests the importance of understanding the face as a mobile, bumpy surface rather then a static flat pattern.

The goal of the research presented in this thesis is to provide a method for extracting an extended FACS model (FACS+) using a physics-based model of both skin and muscle, driven by visual motion extracted from image sequences, not photographs. Our method is capable of very detailed analysis in time and in space, thus providing the information required to observe coarticulation of expressions and to obtain an improved model of facial motion.



(a) Smile (b) Sad (c) Surprise

(d) Anger (e) Disgust (f) Anger

Figure 2-2: *Motion Cues for Facial Expressions. Based on [7]*

2.3 Graphics and Animation

Computer-based modeling and animation of faces has attracted considerable interest in computer graphics for many years. The initial efforts to represent and animate faces using computers go back almost 20 years (see [64] for review). Facial modeling is interesting in computer graphics context as it generates synthetic facial models for 3-D character animation. This in itself has many applications in virtual reality, visualization, telepresence, autonomous creatures, personable interfaces and cinematic special effects.

Facial representation and facial animation share the same set of issues as other representation and animation activities; *modeling*, *motion control* and *image rendering*. Some applications for facial animation such as visualization, biomedical applications and realistic simulations require detailed physics-based models. In other applications, the only requirement is that the faces be believable within the context and the setting. There is extensive debate on which applications need physics-based details and which don't; some have suggested a mix-and-match compromise [98, 52]. The only desire for all (animators and researchers) is that the face of an animated 3-D characters have believable expressions in a given story telling context and that the audience establish an emotional relationship with the characters portrayed. Caricatures and purposeful facial exaggeration [10] are often more acceptable than attempts at realism.

Facial modeling is concerned with developing geometric descriptions or procedures for representing faces. It addresses the issues of facial conformation, realism, likeness, expressiveness and animatability. For modeling purposes, the visible surface of a face is modeled as a network of connected polygons. Some implementations have used curved surface modeling techniques [89]. Facial motion control is concerned with techniques and algorithms for specifying the motions between expressions. This control is achieved by defining a controlled motion of the polygon vertex positions over time in such a way that the rendered facial surfaces have a desired expression in each frame of the animated sequence. Facial image sequences depend on the use of high-quality rendering techniques, which are extensively researched throughout the computer graphics community.

Control of Facial Motion

The earliest, and still a widely used, scheme for implementing and controlling facial animation uses key expression bases and interpolation. Parke [62] demonstrated this key-framing approach, where two or more complete (static) poses were used and the intermediate information was calculated by simple interpolation techniques. Progressive research to reduce the complexity of this kind of synthesis resulted in parameterized representation for facial animation.

The limitations in producing a wide range of realistic expressions using a limited set of parameters led to the development of facial models based on the anatomical structure of the human face. Platt and Badler [75] have developed a partial face model in which the vertices of a face surface are interconnected elastically to form the skin, and are connected to the underlying bone structures by muscles modeled with elastic properties and contraction forces. In their system, facial expression is manipulated by applying forces to the elastically connected skin mesh through the underlying muscles. The muscle actions are patterned after the FACS model described earlier. This adds the anatomical properties of the face to the geometric FACS model. Thus the next generation of facial animation systems included *muscles* as control models.

The major limitation of FACS is that it is a completely geometric model. The muscle model overcomes this limitation by relating the geometric deformations of FACS to the muscle groups. The resulting facial models as presented by Waters *et al.* [92], are anatomically consistent muscle and tissue models. Waters uses simple spring and mass models for skin and muscle to define anatomical models such that expressions are not specific to the topology of a particular face. Viaud *et al.* [88] have extended Waters' model to represent wrinkles.

Magenenat-Thalmann *et al.* [54] developed another muscle based model in which the control parameters are *Abstract Muscle Actions (AMA)* procedures. The AMA procedures are similar to FACS, except that they are not independent. Hence, manipulation directly to the expression level are efficient, but arbitrary combinations of low level motions are not possible.

"Performance" Driven Animation

More recent work develops animation control techniques that rely on information derived from human subjects. Perhaps the first instance of of an application of such a technique was the short animated film "Tony De Petrie" by Emmett [30]. For this film, a human face was photographically digitized for a large number of expressions and mapped to a very exaggerated caricature face.

Parke [64] used a more traditional rotoscoping technique in conjunction with a parameterized facial model in the "3DV" animation. For this animation a live action performance was manually analyzed frame by frame to extract key frame parameters. These key frame parameters are then applied to a key frame animation system. A similar technique by was used by animators at Rhythm and Hues to animate a speaking cat in the Disney movie "Hocus Pocus".

DeGraf [23] demonstrated a real-time facial animation system that used a special purpose interactive device called "waldo" to achieve real-time control. The waldo is a multi-axis electromechanical device that allowed a puppeteer to control a number of facial parameters simultaneously in real time. An advanced version of the same concept, called VActor, was demonstrated by SimGraphics Corporation [38]. This system requires the user to wear a helmet with sensors that attach to the various parts of a face, allowing them to control computer generated creatures. Presently, the VActor system can also be used with an infrared motion analysis system (by Adaptive Optics Associates), which requires pasting infrared reflective dots on a users face[1]. Many other systems for such puppeteering applications have appeared in recent years. Most of them require some sort of a wearable device with sensors or use marks on faces to track motion.

[1]Personal Communication with Dean Wormell of United Technologies Adaptive Optics Associates.

Terzopoulos and Waters [85] have developed a technique for automatic estimation of face muscle contraction parameters from video sequences. These muscle parameters are then used for synthesis of facial expressions. The major limitation of this system is that it does require markings on the face, especially at the cheeks. We will discuss this system again in the section where we talk about facial tracking systems.

Perhaps the most detailed and robust method to automatically track facial motion from video performances and to generate facial animation in 3-D, based on this data is presented by Williams [98]. In his system, a number of fiducial points on the surface of a real face are automatically tracked and the locations of these points are used to control a texture map of a face on a 3-D model, generating realistic facial animation. What makes this system the most robust and detailed is that it uses a 128,000-triangle mesh instead of a few hundred triangles, resulting in increased smoothness and detail in the facial surface. Litwinowitcz *et al.* [52, 66] has presented additions to Williams' work. In one method, by tracing lines of people's faces from video and in the other by spatial mapping of data acquired by video, they warp the texture of the face to generate very realistic facial animation. However, their implementations do use a less detailed model.

Another notable performance driven animation method is that developed by Kurihara and Arai [48]. Their method uses photographs to extract information about expressions. Using this method a theatrical performance was staged in 1993 in Tokyo with one live actor and one computer face.

All of the above described methods require mapping of digitized photographs and image sequences of real faces with facial marking, on polygonal meshes. A library of expressions is then used to control animation. Our work is a great improvement over all of the above as our models have a detailed representation for facial motion in space and time and we use video sequences of real people without any kind of markings. We also present a system to do real-time tracking and animation of facial expressions for extended sequences. Each of these methods are presented in the upcoming chapters.

Speech Synchronization

Any facial animation system aimed at realistic modeling of "synthetic actors" should certainly support speech animation. This kind of animation has a major application in model-based teleconferencing and telepresence systems. Speech animation is often controlled by using a higher level of parameterization built on top of a lower level of parameterization. This second level of parameterization is usually described in terms of *speech phonemes* and the associated visible mouth shapes, or *visemes*. Only a small number of phonemes are needed to produce convincing speech animation. In English, lipreading is based on the observation of forty-five phonemes and associated visemes [90].

This type of multi-level parameterization; one level for emotion/expression, another for speech, has been used by Bergeron and Lachapelle [11], Hill *et al.* [41], Lewis and Parke [50], Magenenat-Thalmann *et al.* [54], and many others (see [64] for review). Waters has also extended his muscle model of faces to generate facial animations with sound [96]. His system takes ascii text, using a letter-to-sound system (LTS), generates synthetic speech, then returns phonemes. These phonemes are then used to compute mouth shapes, resulting in synthetic speech with a facial animation.

Coupling of these independent multi-level parameterizations with emotion is also being pursued. Pelachaud [67] has concentrated on developing methods for 3-D animation of facial expressions in which expressions of emotions are correlated with the intonation of the voice. Cohen *et al.* [19] is concentrating on coarticulation in synthetic visual speech. Cassell *et al.* [17] have developed a system that automatically generates an animated conversation with appropriate speech, intonation, facial expressions and hand gestures using a rule-based technique. Many other systems using faces with speech, especially those aimed at multi-modal human-machine interaction, have appeared in the last two years [84, 39].

The method we describe in the upcoming chapters is extremely well suited to extraction of control parameters for visual speech. We need to record people speaking, may it be just the phoneme sounds or combinations of emotions with speech (for intonation), and analyze the facial and lip motion for these sounds. This analysis can then be used to generate a new set of parameters for visual speech.

Biomedical Modeling

The detailed anatomy of the head and the face is a complex assembly of bones, cartilage, muscles, nerves, blood vessels, glands, fatty tissue, connective tissue, skin and hair. Neuroanatomy suggests that emotional expressions are the result of the movement of skin and connected tissue caused by the contraction of the 44 bilaterally symmetrical facial muscles [29]. To date, no facial model with such detailed anatomy has been implemented. However, several simplified models of bone structure, muscles, skin and tissues have been implemented for facial modeling. Most of these are aimed at the visualization of facial structure for biomedical applications. Waters has suggested the use of his muscle model, initially aimed at facial animation, as a tool for analyzing, and predicting tissue mobility on the face [93]. He uses CT data and data from other sources to generate biomedically valid models of facial shape.

Pieper has used *finite element methods* to model physical attributes of skin for his virtual facial surgery applications [72, 71]. His facial model is extremely detailed and uses material properties from cadaver studies. He has shown excellent post and pre-surgery simulations.

In our work we have developed a detailed model of skin and muscles, based on the works of Platt and Badler [75, 74], Waters [92] and Pieper [72, 71].

2.4 Face processing and analysis by machines

As discussed earlier, facial expression is a primary variable in psychological and sociological research; facial expression communicates information about personal experience, plays a critical role in the communication of interpersonal behavior, and provides a window into brain and autonomic nervous system functioning. Given the importance of facial expression, the need for an objective and automated facial analysis system is compelling.

The previous section concentrated on how we can make synthetic faces appear real and have realistic expressions and speech synthesis. In this section we discuss the relevant work on analysis of facial expressions, especially within the context of automating facial analysis. This kind of analysis obviously forms a percusor to the synthesis aspects described earlier, as better models for motion can be extracted from analysis and used in synthesis.

With the current rate of improvements in technology it appears feasible to automate many aspects of face processing that humans take for granted (face recognition, expression and emotion recognition, lip reading, *etc.*), and to develop new technological aids (robotics, man-machine systems, medical, teleconferencing, *etc.*). It is for this reason the study of facial image processing is an increasingly interdisciplinary topic.

The first instance of detailed analysis of facial expressions appears in 1862 by Duchenne de Boulogne [24]. Duchenne's fascinating photographs and insightful commentary provided generations of researchers with foundations for experimentation in the perception and communication of human facial affect. Some of Duchenne's hypotheses are still widely accepted and most of them have been corroborated by Ekman and his colleagues.

Electromyography (EMG) of facial expression is a very traditional approach, exercised before computer vision methods were being considered to automate facial pattern analysis. Facial EMG has successfully differentiated between positively and negatively valenced effects and has concurrent validity with some FACS action units [27]. Facial EMG is also more sensitive to subtle changes in facial muscles than human observers using FACS, and is without a doubt the most accurate facial motion measurement method. However, the need to attach electrodes to subjects is a significant limiting factor that rules out its use in naturalistic ("in the field") observations. Moreover, the electrodes may hamper the subjects' experience and expression of emotion. EMG does, however, have a natural application in biomedical fields. Bennet [9] has shown how EMGs can be used in a surgical operating room to measure muscle actuations of a patient under anesthesia, in order to determine the level of anesthesia (and discomfort).

In the previous section we discussed the applications of special hardware that was designed specifically for "puppeteering" applications. No one, to our knowledge, has used such hardware, or the infra-red motion tracking systems for facial analysis. Infra-red motion tracking systems are commercially available and are being used to track gait and other human body motions, especially for medical applications. These systems do have a high level of accuracy, however, our attempt to use such a system for tracking facial expressions did not meet with remarkable success and the reflective dots were obtrusive in forming natural expressions. These approaches suggest a need for non-invasive methods for analyzing facial motion and it is for this reason that vision-based sensing is becoming a promising approach.

Vision-based sensing

Computer vision deals with the problem of scene analysis; more specifically the extraction of 3-D information about scenes and objects from 2-D time-varying images obtained by video cameras. Over the years, many algorithms have been developed for determining 3-D shape, texture and motion in scenes. Facial image processing has been an active area of research for at least two decades. Most of the efforts to date have focused on face recognition and head tracking [44, 100, 87, 15]. However, recently much progress has made in estimation of 3-D rigid and nonrigid motion leading to gesture analysis as well as facial expression recognition, understanding and tracking [85, 98, 56, 20, 99].

Face Recognition

Face recognition has typically been posed as a static problem requiring the application of pattern recognition techniques to static images. Significant work has been done to see if face recognition is possible through the analysis and storage of very low level features of images like gray-levels and/or edges. Kanade *et al.* [44] showed that static facial features, like the curvature of the face, the location of facial components (eyes, nose, *etc.*), could be automatically identified and measured from an image of the face. The measurements were sufficiently reliable to permit accurate identification of photographic images of faces. Recently, researchers have successfully used traditional pattern recognition techniques like *principal component analysis* to obtain a psychologically plausible model for human face recognition. Turk and Pentland *et al.* [87, 68] present a modified implementation of this technique, using *eigenvalue analysis*, to locate and recognize faces. However, the study of face features is still pursued intensely. Yuille's [100] work of *deformable templates* uses image features to fit a deformable template to a face and the parameters of this template are then used for shape analysis. Brunelli and Poggio [15] present a comparison between template matching and feature matching for face recognition, and show that template matching results in higher recognition accuracy. Many other researchers have used techniques such as edge detection, intensity variation, *etc.,* to locate the lips, mouth, eyes, and nose on a face (see [14, 86] for review). All of this work is formulated in a static and a passive framework; no observations are made actively and the system does not evolve over time (*i.e.,* is not dynamic).

Face Tracking

We have already discussed facial tracking in the context of facial animation in the previous section. Here we will concentrate more on the analysis aspects of facial motion. We will again discuss two of the works described previously, with more emphasis on their vision-based analysis.

Williams' [98] method automatically tracks a number of fiducial points on the surface of a real face and maps the motion onto a very detailed facial model. The locations of these points control a texture map and when these points are moved on the basis of tracked motion, realistic facial expressions are generated. This is an efficient and an extremely practical system, capable of capturing much detail. It does require direct user input but that is desired, since the main goal of this is system is to *mix-and-match* between motion in image sequences and user defined motion. This system is neither active nor dynamic, and the lack of both is defended extensively by the author [2].

Terzopoulos and Waters' [85] method traces linear facial features, estimates the corresponding parameters of a three dimensional wireframe face model, and reproduces facial expression. A significant limitation of this system is that it requires facial features be highlighted with make-up, especially on the cheeks. Although, active contour models (*snakes*) are used, the system is still passive; the facial structure is passively dragged by the tracked contour features without any active control on the basis of observations. Our method is an extension of Terzopoulos's work, as our method does not look at any prescribed (and marked) regions on a face for extraction of muscle forces, and muscle actuations are computed actively from facial motion.

Another approach was introduced by Mase [56] who developed a method for tracking facial action units using optical flow. This approach is an extension of Mase and Pentland's [57] work on lip reading. The major limitation of this work is that no physical model is employed; the face motion estimation is formulated statically rather than formulated within a dynamic optimal estimation framework. This approach serves as a motivating factor for our work as it shows the applicability of optical flow as a measure of facial motion.

Yacoob [99] presents a system that extends Mase's work. Yacoob's system uses optical flow computation to determine motion in different regions of the face and then on the basis of FACS, defines a rule-based system for recognition of facial expression from dynamic image sequences.

Another system worth mentioning is developed by Reinders *et al.* [78]. This method is similar to ours as it does use *a priori* knowledge of the facial model. It then uses this model and the shape representations of facial features in a feature extraction scheme. These features are then tracked to extract rigid and nonrigid motion. A muscle model is also used to manifest facial motion for expression. This method is in many ways similar to our method, but lacks detail in both modeling and in motion extraction. It is mainly aimed at facial tracking and lacks that active dynamic control system that makes our analysis detailed and robust. It relies on FACS parameters for motion modeling and uses a neural network to "decide" which motion leads to what FACS unit.

Teleconferencing, Telepresence, and Coding

So far we have only discussed methods used to track facial motions. In the previous section we also discussed issues of facial modeling and control of facial motion. The combination of these two methods (facial modeling and control methods and facial tracking methods) can be used to develop a teleconferencing/telepresence system or a model-based image coding system. All of the systems described in the previous section can be used for this application

[2]Personal Communication with Lance Williams.

Perhaps the most interesting application of a teleconferencing system is introduced by Haibo Li, Pertti Roivainen and Robert Forchheimer [51], who describe an approach in which a control feedback loop between computer graphics and computer vision processes is used for a facial image coding system. This results in an efficient model-based coding system. This method, though dedicated to teleconferencing applications, is similar to our method. The limitation of this work is the lack of both a dynamic model and of observations of motion over large predefined areas on the face. Our method overcomes these limitations.

In addition to the above methods for teleconferencing applications, there are many other researchers working on model-based coding and tracking of facial motion [78, 97]

Another interesting application that has been extensively studied with reference to analysis of facial motion is automatic lipreading. Interested readers are recommended to review Petajan [70] and Goldschen [40].

Chapter 3

Vision-based Sensing: Visual Motion

Movement is an inherently continuous process that usually produces smooth changes in an image. [This] is a rather intrinsic property of movement with regard to its perceptual analysis, since its very continuity should help in the task of following pieces of an object around in an image to find out how they are moving.

David Marr, *Vision*, p 183 [55]

3.1 Introduction

The dynamic evolution of images over time provides enormous amounts of information about a scene. Because of the richness of visual motion as an information source, its analysis is useful for many practical applications. These range from image-processing problems like image-coding, and the enhancement of image sequences, to active machine vision problems. A significant motivation for studying motion is the abundance of evidence suggesting that biological vision systems devote considerable resources to this type of information processing. Although the processing constraints in a biological system are somewhat different than those in an artificial vision system, both must extract motion information from the same type of intensity signals. For this reason the point-by-point image motion field was chosen as input for the current study of facial expressions. This form of biologically-viable sensing method helps provides a solid ground for the active and dynamic framework of facial modeling that forms the main part of this thesis.

3.2 Optical Flow / Motion Vectors

Many computer vision techniques start by computing a point-by-point estimate of the motion field known as "optical flow". The purpose is to measure the apparent motion of the image brightness pattern from one frame to the next based on the assumption that these intensity patterns are locally preserved from frame to frame. Because of this assumption it is important to differentiate between optical flow and the true motion field. It is our goal to obtain good estimates of the *true* motion field, extracted by using an active and dynamic model of the face to determine the pattern of facial change over time. We will discuss in detail the modeling paradigm for the face in the next chapter. In this chapter, for the sake of introducing concepts, we will not rely on any symbolic representation of the model or scene (*i.e.,* the face).

Many local approaches to computing optical flow are known. All of these methods suffer from the same inherent problems; intensity singularities, non-motion brightness changes, temporal aliasing and multiple motions. One way of addressing these problems is by probabilistic modeling of the motion field using *a priori* information. Simoncelli [82] has developed a framework for integrating prior information and directional uncertainty information within a Bayesian analysis of the problem. We will adopt Simoncelli's formulations for the optical flow computation for the first stage of our estimation and analysis framework.

3.3 Probabilistic Modeling

Viewing the problem of estimating visual motion probabilistically has many advantages:

1. It produces useful extensions of the standard quadratic gradient techniques for computing optical flow, including an automatic gain control mechanism, and the incorporation of a prior distribution on the velocity field.

2. It provides (two-dimensional) confidence information, allowing later stages of processing to tailor the use of velocity estimates according to the shape of the distribution. For example, it enables the development of algorithms to combine information recursively over time, or over space to estimate higher-order motion parameters such as rigid body translations or nonrigid deformations. This is used later in Chapter 5 for computing deformations in facial structure over time.

3. It provides a framework for "sensor fusion," in which image velocity estimates must be combined with information derived from other uncertain sources. This framework is very important for this work as described in Chapter 5; the dynamics of a physical model is used to help with modeling and computing flow.

4. A distributed representation is appropriate for modeling biological motion processing (see Simoncelli [82].)

Once we have a probabilistic formulation of the problem, we introduce a *coarse-to-fine* mechanism to acquire multiple motion fields varying over scale, and to tackle temporal aliasing.

We need to introduce an uncertainty model to compute an expression for the probability of image velocity, conditional on measurements made from image sequences. This expression will characterize deviations from the standard gradient constraint equation. The original differential constraint on optical flow is:

$$\mathbf{f}_s \cdot \mathbf{v} + f_t = 0, \tag{3.1}$$

where we define the image intensity signal as a continuous function of position and time: $f(x, y, t)$, with $\mathbf{f}_s = [f_x, f_y]^T$ and f_t as the spatial and temporal partial derivatives of the image f. The $(\cdot)$ operator indicates an inner product. We define $\tilde{\mathbf{v}}$ as the measured optical flow, and $\mathbf{v}$ as the true motion. Complete agreement between the computed optical flow and true velocity field is impossible therefore we describe the difference between these using a random variable, $\mathbf{n_V}$, to characterize the error uncertainty,

$$\tilde{\mathbf{v}} = \mathbf{v} + \mathbf{n_V}. \tag{3.2}$$

Similarly, let $\tilde{f}_t$ be the actual temporal derivative, and f_t the measured derivative. Then, we write

$$f_t = \tilde{f}_t + n_{f_t},$$

with n_{f_t} as a random variable characterizing the uncertainty in measurements relative to the true derivative. And similarly,

$$\mathbf{f}_s = \tilde{\mathbf{f}}_s + \mathbf{n}_{\mathbf{f}_s},$$

is the measured spatial derivative.

Now the gradient constraint applies to the actual measurements and the optical flow vector. After some algebraic manipulations we may write:

$$\mathbf{f}_s \cdot \mathbf{v} + f_t = \mathbf{f}_s \cdot \mathbf{n_V} + \mathbf{v} \cdot \mathbf{n}_{\mathbf{f}_s} - \mathbf{n}_{\mathbf{f}_s} \cdot \mathbf{n_V} + n_{f_t}. \tag{3.3}$$

This equation gives us a probabilistic relationship between the image *motion field* and the *measurements* of spatio-temporal gradient. It accounts for errors in our derivative measurements, and for deviations of the velocity field from the optical flow, but it assumes that the underlying optical flow constraint is valid.

In order to make use of this formulation, we must characterize the random variables n, in these definitions. We will assume that we may characterize these random variables with *independent zero-mean Gaussian distributions* [61].

Given these assumptions of independent zero-mean Gaussian noise sources, the right side of Equation (3.3) is a zero-mean Gaussian random variable with variance equal to $\mathbf{f}_s^T \Lambda_{\mathbf{V}} \mathbf{f}_s + \lambda_{f_t}$, where $\Lambda_{\mathbf{V}}$ and λ_{f_t} are a covariance matrix and a variance corresponding to $\mathbf{n_V}$ and n_{f_t}, respectively. We interpret the equation as providing a conditional probability expression:

$$\mathcal{P}(f_t \mid \mathbf{v}, \mathbf{f}_s) = \exp\left\{-\tfrac{1}{2}(\mathbf{f}_s \cdot \mathbf{v} + f_t)(\mathbf{f}_s^T \Lambda_{\mathbf{V}} \mathbf{f}_s + \lambda_{f_t})^{-1}(\mathbf{f}_s \cdot \mathbf{v} + f_t)\right\}.$$

After some manipulations based on Bayes' rule and using a zero-mean Gaussian with covariance Λ_p, for the prior distribution $\mathcal{P}(\mathbf{v})$, we get a resulting distribution which should also be a Gaussian:

$$\mathcal{P}(\mathbf{v} \mid \mathbf{f}_s, f_t) = \exp\left\{-\tfrac{1}{2}(\hat{\mathbf{v}} - \mathbf{v})^T \Lambda_{\mathbf{v}}^{-1}(\hat{\mathbf{v}} - \mathbf{v})\right\}. \tag{3.4}$$

The covariance matrix, $\Lambda_{\mathbf{v}}$, and estimated flow, $\hat{\mathbf{v}}$ (which is also the mean vector for Gaussian distributions) may be derived using standard techniques (*i.e.,* completing the square in the exponent):

$$\begin{aligned}
\Lambda_{\mathbf{v}} &= \left[\mathbf{f}_s(\mathbf{f}_s^T \Lambda_{\mathbf{v}} \mathbf{f}_s + \lambda_{f_t})^{-1}\mathbf{f}_s^T + \Lambda_p^{-1}\right]^{-1}, \\
\hat{\mathbf{v}} &= -\Lambda_{\mathbf{v}} \mathbf{f}_s(\mathbf{f}_s^T \Lambda_{\mathbf{v}} \mathbf{f}_s + \lambda_{f_t})^{-1} f_t.
\end{aligned}$$

The advantage of the Gaussian form is that it is parameterized by these two quantities ($\Lambda_{\mathbf{v}}$ and $\hat{\mathbf{v}}$) that are computed in analytic form from the derivative measurements.

If we choose $\Lambda_{\mathbf{v}}$ to be a diagonal matrix, with diagonal entry $\lambda_{\mathbf{v}}$, (*i.e.,* no cross-covariances) then we can write the above equation as:

$$\begin{aligned}
\Lambda_{\mathbf{v}} &= \left[\frac{\mathbf{J}}{(\lambda_{\mathbf{v}}\|\mathbf{f}_s\|^2 + \lambda_{f_t})} + \Lambda_p^{-1}\right]^{-1}, \\
\hat{\mathbf{v}} &= -\Lambda_{\mathbf{v}} \cdot \frac{\mathbf{b}}{(\lambda_{\mathbf{v}}\|\mathbf{f}_s\|^2 + \lambda_{f_t})}.
\end{aligned} \tag{3.5}$$

where matrix $\mathbf{J}$ and vector $\mathbf{b}$ are defined as:

$$\mathbf{J} = \mathbf{f}_s \mathbf{f}_s^T = \begin{pmatrix} f_x^2 & f_x f_y \\ f_x f_y & f_y^2 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} f_x f_t \\ f_y f_t \end{pmatrix}. \tag{3.6}$$

3.4 Multiple Motions and Temporal Aliasing

Given that our imagery contains only a single global motion, it would suffice to stop our computation at the above described state. In typical scenes of facial expression and pattern change, slowly varying motion fields are rare. Although a single velocity may account for much of the motion in some regions, it is likely that subportions are moving differently and at different velocities. Consider the motion field of a smiling face. The motion field is quite diverse over the whole face. There will be significant motion in the lower part of the face, but small motion in the eye region is also quite probable. The low frequency subband estimates will be unable to capture these local variations.

In order to get better estimates of *local* velocity, higher-frequency bands must be used, with spatially smaller filters. What we would like to do is to use the coarse motion estimate to "undo" the motion, roughly stabilizing the position of the image over time. Then higher frequency filters can be used to extract local perturbations to this large-scale motion. That is, we can use higher frequency filters to estimate optical flow on the warped sequence, and this "optical flow correction" may then be composed with the previously computed optical flow to give a new optical flow estimate. This correction process may be repeated for progressively finer scales.

In image-processing situations, where the image-gathering has already occurred, we can "warp" a spatially and temporally localized region of the image content in a direction opposite to the computed motion. For our purposes, we compute the warped image sequence:

$$\mathcal{W}\{f, v\}(x, y, t + \Delta t) = f(x - v_x \Delta t, y - v_y \Delta t, t + \Delta t), \tag{3.7}$$

where (v_x, v_y) is the warp vector field corresponding to the velocity estimated from the coarser scale measurements. Note that the warping only need be done over a range of Δt that covers the temporal extent of the derivative filters that will be applied. The warping procedure may be applied recursively to higher frequency subbands. This type of multi-scale "warping" approach has been suggested and used by a number of authors [3, 77].

3.5 Coarse-to-Fine Flow with Recursive Estimation

A "coarse-to-fine" algorithm is a technique for combining information from different spatial scales and for imposing a prior smoothness constraint (see, for example, [83, 18]). This basic technique does, however, have a serious drawback; if the coarse-scale estimates are incorrect, then the fine-scale estimates will have no chance of correcting the errors.

To fix this, we must have knowledge of the error in the coarse-scale estimates. Since we are working in a probabilistic framework, and we have information describing the uncertainty of our measurements, we may use this information to properly combine the information from scale to scale. We define a *state evolution* equation with respect to scale:

$$\mathbf{v}(l+1) = E(l)\mathbf{v}(l) + \mathbf{n_v}(l); \qquad \mathbf{n}_0(l) \sim N(0, \Lambda_0). \tag{3.8}$$

where l is an index for scale (larger values of l correspond to finer scale), $E(l)$ is the linear interpolation operator used to extend a coarse scale flow field to finer resolution, and $\mathbf{n_v}$ is a random variable corresponding to the certainty of prediction of the fine-scale motion from the coarse-scale motion. We assume that the $\mathbf{n_v}(l)$ are independent, zero-mean, and normally distributed. Implicitly, we are imposing a sort of fractal model on the velocity field. This type of scale-to-scale Markov relationship has been explored in an estimation context in [82, 18, 5].

We also need to define a *measurement* equation:

$$-f_t(l) = \mathbf{f}_s(l) \cdot \mathbf{v}(l) + (n_{f_t} + \mathbf{f}_s(l) \cdot \mathbf{n}_{\mathbf{f}_s}). \tag{3.9}$$

We will assume, as before, that the random variables are zero-mean, independent and normally distributed (*i.e.,* Gaussian). Remember that this equation is initially derived from the total derivative constraint for optical flow. This equation is a bit different than the measurement equation used in most estimation contexts. Here, the linear operator relating the quantity to be estimated to the measurement, f_t, is also a measurement [82].

Given these two equations, we may write down the optimal estimator for $\mathbf{v}(l+1)$, the velocity at the fine scale, given an estimate for the velocity at the previous coarse scale, $\hat{\mathbf{v}}(l)$, and a set of fine scale (gradient) measurements. The solution is in the form of a standard Kalman filter [36], but with the time variable replaced by the *scale*, l:

$$\begin{aligned}
\hat{\mathbf{v}}(l+1) &= E(l)\hat{\mathbf{v}}(l) + \mathcal{K}(l+1)\nu(l+1), \\
\Lambda_{\mathbf{v}}(l+1) &= \Lambda'(l+1) - \mathcal{K}(l+1)\mathbf{f}_s^T(l+1)\Lambda'(l+1), \\
\mathcal{K}(l+1) &= \Lambda'(l+1)\mathbf{f}_s(l+1)\cdot\left[\mathbf{f}_s^T(l+1)\left(\Lambda'(l+1)+\lambda_{f_t}\right)\mathbf{f}_s(l+1)+\Lambda_{\mathbf{f}_s}\right]^{-1}, \\
\nu(l+1) &= -f_t(l+1) - \mathbf{f}_s^T(l+1)E(l)\hat{\mathbf{v}}(l), \\
\Lambda'(l+1) &= E(l)\Lambda_{\mathbf{v}}(l)E(l)^T + \Lambda_{\mathbf{v}}.
\end{aligned} \tag{3.10}$$

Here, $\nu(l)$ corresponds to an *innovations* process. The innovations process represents the new information contributed by the measurements at level l.

The problem with the equations given above is that due to the temporal aliasing, we cannot compute the derivative measurements at scale l without making use of the velocity estimate at scale $l-1$. In order to avoid this problem, we must write $\nu(l)$ in terms of derivatives of the warped sequence. That is, expanding around a time t_0, we write:

$$\nu(l+1) \approx -\frac{\partial}{\partial t}\mathcal{W}\{f(l+1), E(l)\hat{\mathbf{v}}(l)\}(x, y, t_0). \tag{3.11}$$

Thus, the innovations process is computed as the temporal derivative of the the image at scale $l+1$, after it has been warped with the interpolated flow field from scale l. In order to make the solution computationally feasible, we ignore the off-diagonal elements in $\Lambda'(l+1)$ (i.e., the cross-correlations between adjacent interpolated flow vectors).

The Kalman solution may be put into the alternative "update" form by use of some matrix manipulations [36] and Equation (3.6) to get:

$$\Lambda_{\mathbf{v}}(l+1) = \left[\frac{\mathbf{J}}{(\lambda_{\mathbf{v}}\|\mathbf{f}_s\|^2 + \lambda_{f_t})} + \Lambda'(l+1)^{-1}\right]^{-1}. \tag{3.12}$$

Similarly, we may rewrite the updated estimation (mean) vector as:

$$\hat{\mathbf{v}}(l+1) = E(l)\hat{\mathbf{v}}(l) + \boldsymbol{\Lambda}_{\mathbf{v}}(l+1)\frac{\mathbf{b}'}{(\lambda_{\mathbf{v}}||\mathbf{f}_s||^2 + \lambda_{f_t})}, \tag{3.13}$$

where the vector $\mathbf{b}'$ is defined by $\mathbf{b}' = \mathbf{f}_s\nu(l+1)$.

These mean and covariance expressions are the same as those of Equation (3.5) except that:

1. the prior covariance $\boldsymbol{\Lambda}_p$ has been replaced by $\boldsymbol{\Lambda}^{'}(l+1)$,
2. the vector $\mathbf{b}$ has been replaced by $\mathbf{b}'$, which is computed in the same manner but using the *warped* temporal derivative measurements, and
3. the mean $\hat{\mathbf{v}}(l+1)$ is augmented by the interpolated estimate from the previous scale.

The output of the above described process is an estimate of the optical flow, $\hat{\mathbf{v}}(t)$, from one frame f_t to the other frame f_{t+1}, computed by determining the estimates $\hat{\mathbf{v}}(l)$ over scales l and then accumulating/composing into the overall flow. Similarly, since the error covariances at all the scales ($\boldsymbol{\Lambda}_{\mathbf{v}}(l)$) is available, we can also determine the overall error covariances $\boldsymbol{\Lambda}_{\mathbf{v}}(t)$ between frames. This error covariance is the covariance matrix of errors in observation over time, and will be a useful parameter in time-varying estimation. These error covariances are employed in the dynamic estimation module as discussed in Chapter 5. The next chapter introduces the geometric models for representing structured models of faces. Chapter 7 discusses the implementation of the algorithm presented here and also the results of the optical flow computation.

Chapter 4

Modeling

To move things is all that mankind can do; ... for such the sole executant is muscle, whether in whispering a syllable or in felling a forest.

Charles Sherrington, 1924[1]

4.1 Introduction

In the previous chapter, a method for extraction of facial motion was discussed without any reference to faces or their structure. A *priori* information about the facial structure can provide additional constraints for accurate motion extraction and is essential for fulfilling our goal of determining FACS+.

This chapter concentrates on the facial modeling; from geometric models to detailed physics and biomechanics based models. An overview of FACS and muscle models is presented, followed by their applications in computer graphics. Our extensions to the FACS and muscles models, achieved by describing the facial models as finite element meshes and elastic muscles, is presented in the later part of this chapter. The next chapter (Chapter 5) will combine this modeling with the visual sensing method, described in the previous chapter.

FACS model

The Facial Action Coding System, as described earlier in Chapter 2, is used for describing facial expressions and was developed by Ekman and Friesen in 1978 [28]. This approach for facial movement is achieved by describing and encoding the most basic facial muscle actions and their effects on facial expression. It independently models all facial motion based on muscle action. These "Action Units" (AUs) are based on the visible motions of parts of the face. However, for coding of the motions related to these AUs the

[1]From *p 231* [45].

Action (Result)	Muscular Basis	Closest AUs
pull up lip corner (raise cheek)	Zygomastic Major Orbicularis Oculi	$AU6^-$ (cheek raiser)
pull lip corners backwards	Zygomastic Major Buccinator	$AU12^-$ (lip corner puller)
raise upper lip	Levator Labii	$AU10^-$ (upper lip raiser)
depress lip corners	Depressor Angli Oris	$AU15^-$ (lip corner depressor)
raise chin	Mentalis	$AU17^-$ (chin raiser)
drop jaw	Jaw action (relaxed)	$AU26^+$ (jaw drops)
part lips	Depressor Labii	$AU25^+$ (lips part)
relax upper lip	Orbicularis oris	$AU25^-$ (lips part)
make wrinkle at root of nose	Depressor Supercilli AND Levator Labii	$AU9^-$ (nose wrinkler)
raise brow OR lower brow	Ventor Frontalis AND Depressor Supercilli	$AU1^-$ (brow raiser) $AU4^+$ (brow lowerer)

Table 4.1: *Relation between muscles and FACS action units. The signs of $+$ and $-$ indicate the direction of actuation. Adapted from [56, 63].*

underlying muscle actions are also accounted for, providing a more detailed description. An association between these action units and the muscles on the face is presented in Table 4.1 while Figure 4-1 shows a schematic of facial muscles. Ekman's FACS model has 46 major Action Units and divides the face into three areas (brows, eyes and lower face).

Computer graphics and geometric models

Although not originally intended for computer graphics and geometric modeling, the FACS scheme is now widely used for geometric and kinematic manipulation of facial shape and expression control [75, 89, 65]. The FACS model simply describes the actuation of a set of muscles (AUs) for a set of facial expressions. Computer graphics and animation implementations require modeling the face as a geometric shape and the actuations as movements applied to the different parts of the shape. This resulted in describing a face in polygonal form and defining vertices at locations required for AU motions. The simplest geometric model with almost all of the AUs defined is the CANDIDE model [79]. The CANDIDE model represents all the AUs in a parametric form and since it is a very simple model (only about 100 triangles) it is widely used in the model-based image coding community [51]. Figure 4-2 shows the CANDIDE model as presented by Welsh [97].

Over the years, some very detailed models of faces based on FACS for expression modeling have appeared (see [65] for review). Perhaps the most widely used model is by Waters [92] and uses muscles

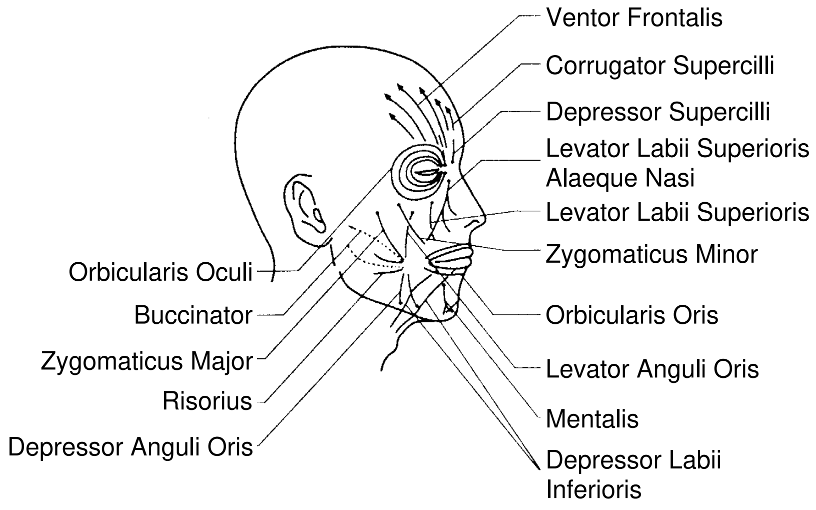


Figure 4-1: *Schematic of Facial Muscles. Adapted from [56].*

to actuate various FACS actions. This model is discussed in the next section. One of the most detailed models is by Platt and Badler [75, 74] and has parametric representation for the FACS motions and detailed information for different regions of the face.

These models of varying detail are sufficient for modeling expressions as per the FACS model, since the vertices of the face can be actuated on the basis of the FACS Action Units. However, such a system is not dynamic and does not provide support for active observation and dynamic estimation. For this reason we must also build dynamics and muscle descriptions onto our geometric model. Using Platt's model as the base model, we have extended it to incorporate a *multi-grid* finite element mesh to describe the skin surface and have developed an elastic model for the face muscles. Details of these extensions follow in Section 4.3 of this chapter.

Muscle-based Deformation

Facial parameterization as provided by the models has the inherent limitation of being restricted to a specific facial topology and requires hard-wiring of all performable actions. This limitation led Waters [92] to develop a muscle model that is controllable by a limited number of parameters and is non-specific to the facial topology, allowing a richer vocabulary for modeling facial expressions. Waters designed this muscle model by first defining a skin surface made of a spring lattice and the attaching muscles as non-linear (muscle-like) actuators. This allows facial models to be derived from scanners and mapped onto the facial model with muscles and skin defined by this topologically-invariant data, as long as muscle attachment points can be determined [94, 93, 95].

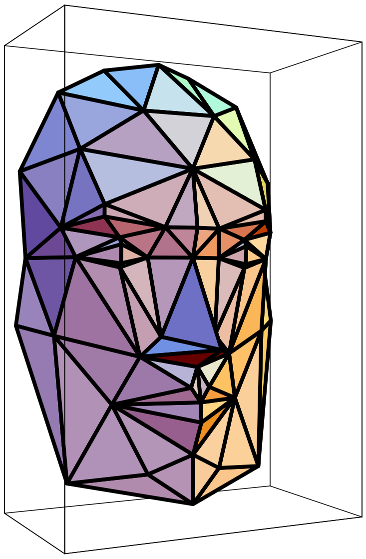


Figure 4-2: *The* CANDIDE *model: a geometric facs model [97].*

It is important to note that FACS and the related AU descriptions are purely static and passive, and therefore the association of a FACS descriptor with a dynamic muscle is inherently inconsistent. By modeling the elastic nature of facial skin and the anatomical nature of facial muscles Waters has achieved a dynamic model of the face, including FACS-like control parameters. However, a major limitation of Waters' method is the lack of inertial properties which, from a theoretical viewpoint, means that the system is a static system not a dynamic one. By implementing a procedure similar to that of Waters', we have also developed a dynamic muscle-based model of a face, however our model includes both inertial and damping properties. This addition is important for our work as we want to map visual motion measurements, including small motions, onto our facial model. The intertial and damping properties play and important role in fully utilizing the energy surface for the face, as it deforms dues to observed motion, to attain a equilibrium state.

4.2 Physically-based Modeling

A physically-based dynamic model of a face requires the application of *finite element* and *discrete element* modeling methods. These methods give our facial model an *anatomically-based* facial structure by modeling facial tissue, and muscle actuators, with a geometric model to describe force-based deformations and control parameters.

For dynamic modeling we need to integrate the system dynamics with respect to the following equation of rigid and nonrigid dynamics,

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{D}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{R}. \tag{4.1}$$

where $\mathbf{u} = [U, V, W]^T$ is the global deformation vector describing the deformation in the facial structure over time. We use the polygonal mesh used to describe the geometric shape of a face (example meshes are shown in Figures 4-2 and 4-6), as the finite element mesh with n nodes (*i.e.,* vertices) and m elements (*i.e.,* polygons), and assume that the material properties are linear and deformations are small. This allows us to define M as a $(3n \times 3n)$ mass matrix, which accounts for the inertial properties of the face, $\mathbf{K}$ as a $(3n \times 3n)$ stiffness matrix, which accounts for the internal energy of the facial structure against deformations due to its elastic nature, and $\mathbf{D}$ as a $(3n \times 3n)$ damping matrix. The Vector $\mathbf{R}$ is a $(3n \times 1)$ vector, characterizing the force actuations by the muscles. For a more detailed exposition of the finite element method and its applications to physically-based modeling to graphics and vision see [8, 32, 59].

These matrices are calculated using the principle of *virtual work,* and the direct stiffness method by integrating over the volume of m finite elements, followed by the assemblage of the m elements:

$$\mathbf{K} = \sum_m \int_V \boldsymbol{\mathcal{B}}^T \mathbf{E} \boldsymbol{\mathcal{B}} dV, \tag{4.2}$$

$$\mathbf{M} = \sum_m \int_V \mathbf{H}^T \rho \mathbf{H} dV, \tag{4.3}$$

$$\mathbf{D} = \sum_m \int_V \mathbf{H}^T \kappa \mathbf{H} dV. \tag{4.4}$$

Here $\mathbf{E}$ is a constitutive relationship matrix, ρ the mass density and κ the damping coefficient. $\mathbf{H}$ is the polynomial interpolation matrix and relates the nodal displacements $\bar{\mathbf{u}}$ to the displacements of the whole element, while $\boldsymbol{\mathcal{B}}$, the strain displacement matrix filled with partial dericvatives of the elements of $\mathbf{H}$ with respect to spatial coordinates, relates the nodal displacements to the strains across the element.

$$\begin{aligned} \mathbf{u}(x, y, z) &= \mathbf{H}(x, y, z)\bar{\mathbf{u}} \\ \boldsymbol{\epsilon}(x, y, z) &= \boldsymbol{\mathcal{B}}(x, y, z)\bar{\mathbf{u}} \end{aligned} \tag{4.5}$$

The $\boldsymbol{\mathcal{B}}$ and $\mathbf{H}$ matrices are the basis of the finite element formulations. The formulations for these matrices is described in detail in [8, 81].

FEM to model skin

A direct application of the above concepts to a polygonal model of a face surface provides a FEM model of skin. Material properties like *modulus of elasticity*, density, thickness, damping coefficient *etc.,* of skin are required. Also a decision needs to made as to what kind of finite element should be used to model skin. The different kinds of finite elements define the relationships of motion across the three dimensions of the element. This is mostly determined by the elements of the constitutive relationship matrix $\mathbf{E}$.

Pieper [72, 71] uses a 20 node three-dimensional isoparametric element to describe each finite element.

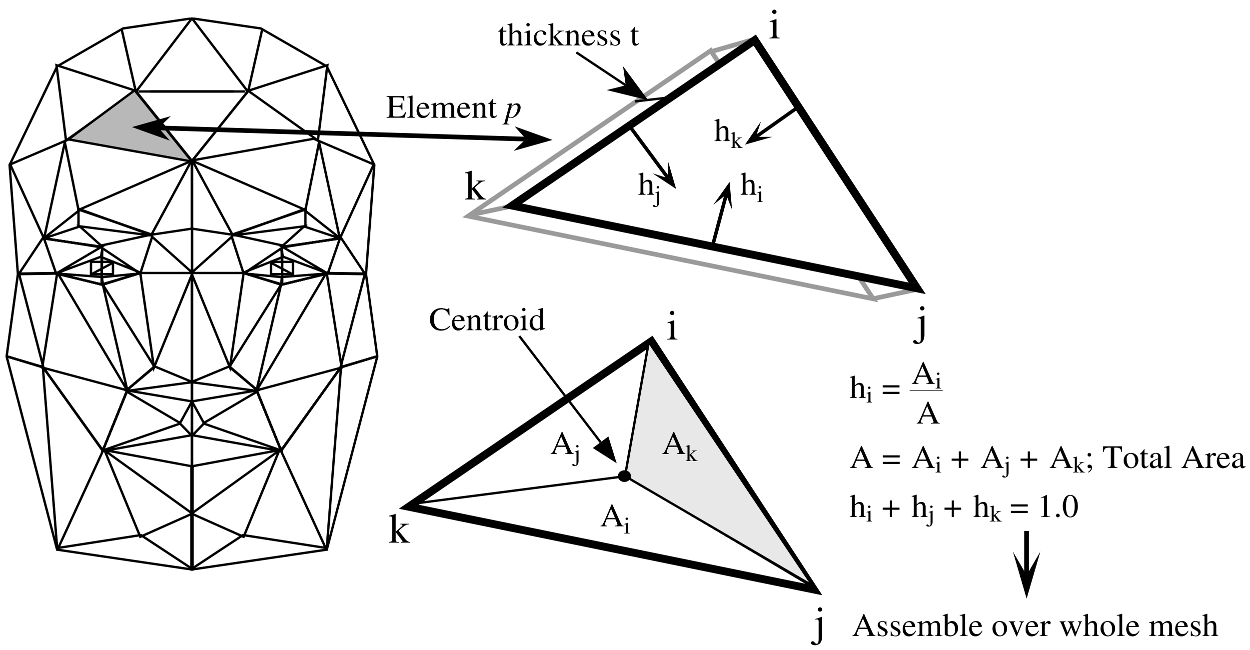


Figure 4-3: *Using the FEM mesh to determine the continuum mechanics parameters of the skin.*

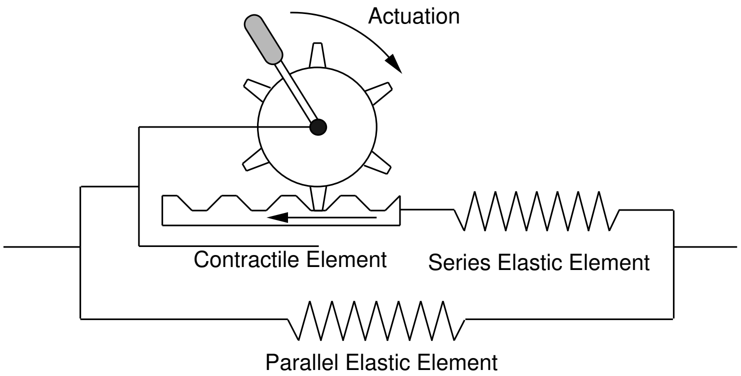


Figure 4-4: *Mechanical Muscle model. Shows a "rack and pinion" system for force actuations.*

His model, which was developed to model detailed deformation for his facial surgery simulator, has two to three layers of such finite elements to describe the skin surface. The total number of finite elements to describe the whole skin surface is usually 20. For material properties he acquired data from cadaver studies [45]. Pieper also makes the assumption that dynamic motion is not important for his simulation and hence ignores inertial properties. He uses a small time-step and his motions are constrained to be quite small per time step. His simulations needs permit him the luxury of using larger and fewer finite elements.

In order to extract a basis for facial motions by observation of facial pattern change, we require a very robust dynamic model. This dynamic model can only be achieved for our application by having a large number of finite elements. It is for this reason we decided to use *isoparametric triangular shell element*. By defining each of the triangles on the polygonal mesh as a finite element (an example is shown in Figure 4-3) we we can calculate the mass, stiffness and damping matrices for each element (using $dV = tdA$), given the material properties of skin. Then by the assemblage process of the direct stiffness method as per Equations (4.2), (4.3), and (4.4) the required matrices for the whole mesh can be determined. As the integration to compute the matrices is done prior to the assemblage of matrices, each element may have different thicknesses t and material properties, although large differences in thickness and properties of neighboring elements are not suitable for convergence [8].

For a more detailed exposition of the finite element method to model skin and muscles, interested readers are directed to [8, 81, 2, 31, 72].

Models for muscle

The next step in formulating this dynamic model of the face is the combination of the skin model with a dynamic muscle model. This requires information about the attachment points of the muscles to the face, or in the geometric case, information about the attachment to the vertices of the geometric surface/mesh. Work of Waters [93] and Pieper [72] provides us some of this information based on anatomical data. Figure 4-5 shows the attachment of a frontalis muscle to the upper brow of a face. In this way whenever the muscle is actuated the skin deforms, or in the inverse case, whenever our dynamic observations of facial motion deform the skin, we can "sense" the actuations of the rack and pinion system in Figure 4-4. Our system uses a muscle model that is a simplification of the model presented in [37]. See McMahon or Kandel *et al.* [45] for more detailed models.

4.3 Modeling a Face

So far we have discussed how the finite element method can be used to model skin surface and how we can attach elastic muscle models to it. Now we will discuss the specifics of our model. Note that facial motion is extremely complex. Even the simplest of motions, *e.g.,* a wink of an eye, creates many changes in the facial pattern. Some of this is large scale motion that covers the eyebrows, the cheek region, and the eyelid, but additionally there is small local motion just next to the eye and near the nose bridge, and wrinkles are generated near the eyes. In the previous chapter, we established that using a *coarse-to-fine* algorithm has the advantage of allowing us to capture both local (high frequency) and global (low frequency) motion. We also want our facial mesh have a multi-grid representation to account for global and local motion.

Our facial model, as stated earlier, is based on the generic face model developed by Platt and Badler [75]. After some modifications to suit our needs, this model now has 1223 vertices and 2321 triangular polygons. This polygonal model is shown in Figure 4-6. As this figure illustrates, this is a very detailed model of a

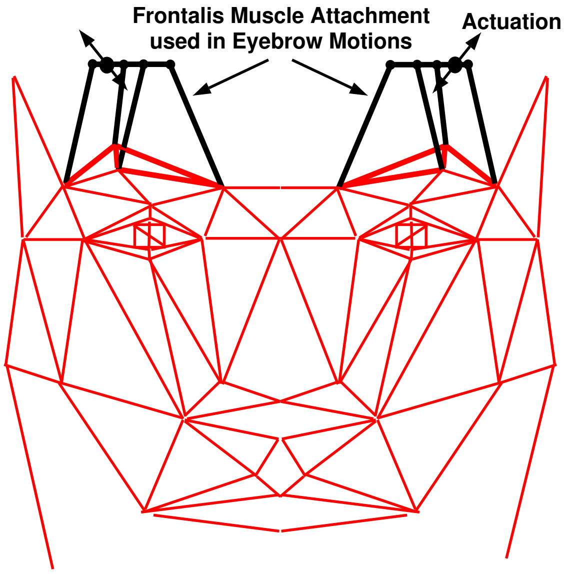


Figure 4-5: *Attachments of Frontalis muscle used in eyebrow motions.*

face. Our goal is to generate two layers of finite element meshes using this mesh, one layer that is a coarse mesh in order to account for large-scale motions and the other, a finer mesh to model smaller motions.

Coarse mesh

Figure 4-7(a) shows the point data for the generic face shown in Figure 4-6. Platt and Badler provide a detailed description of 80 different regions of the face, each of them having a set of possible motions (see Appendix A for a list). These facial regions are shown in Figure 4-7(b). The centroids of these 80 regions are shown in Figure 4-7(c). Also shown are the nodes that are fixed so as to distinguish between rigid and nonrigid motion. Using these region centroids as nodes, a coarse level finite element mesh is generated as shown in Figure 4-7(d). Using the method described in the previous section we can generate matrices in Equation (4.1) by using *triangular isoparametric shell elements* and assembling over the whole mesh. The motion estimates, as they are computed from image sequences, are applied directly to these individual regions. More details of this mapping are presented in the next chapter. This mesh has 80 nodes and 134 elements of which 10 nodes (around the eye-nose bridge and on the outer edge of the mesh) are designated as fixed nodes.

Fine mesh

The fine mesh is composed of all the triangular polygons shown in Figure 4-6. Applying the method of descretizing the face model with the triangular isoparametric elements, we define a mesh with 1223

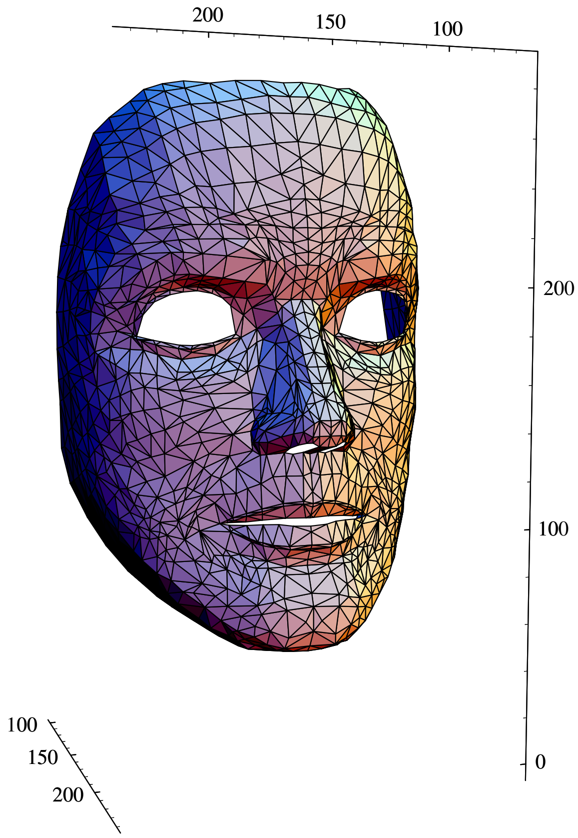
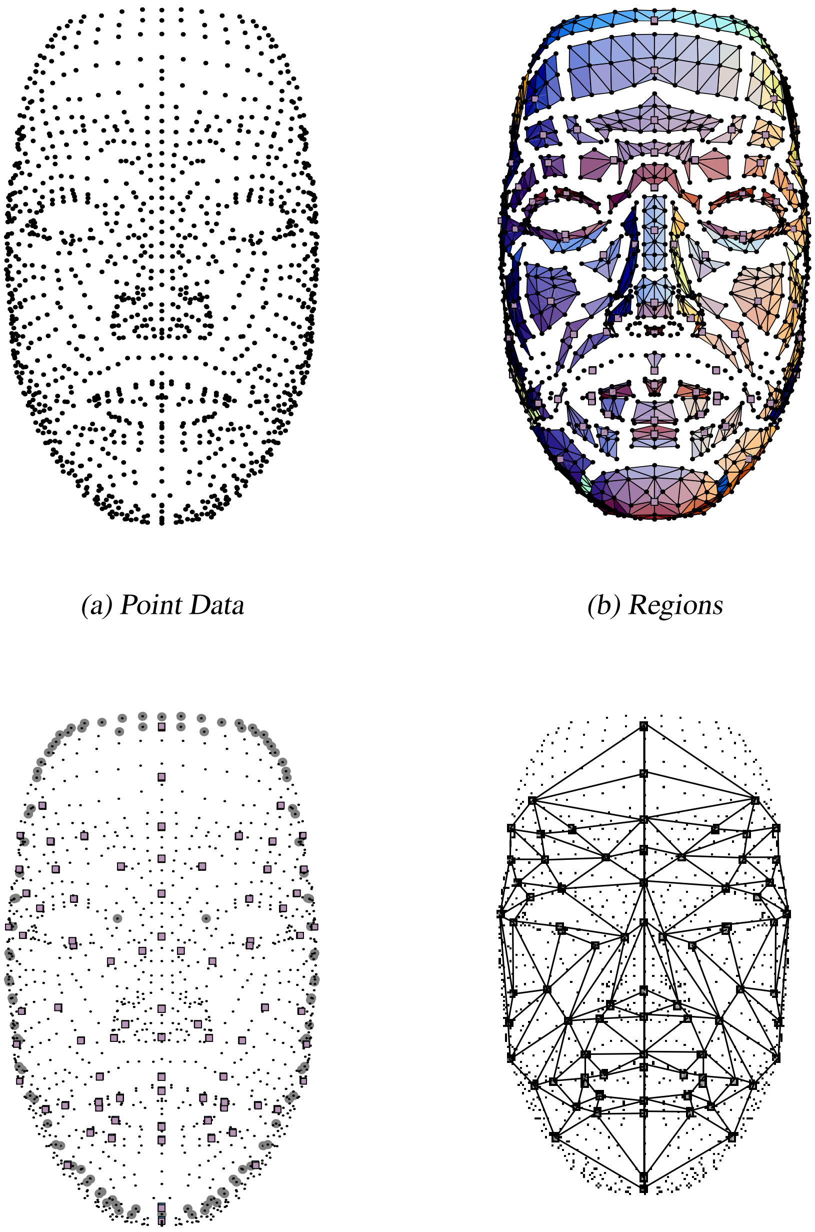


Figure 4-6: *Geometric Model of a Face (Polygons/Vertices).*

nodes and 2321 elements. Of these 1223 nodes 134 are considered rigid nodes. This mesh is shown in Figure 4-8(a). Computing the finite element matrices for both the meshes is not only a time consuming problem but also a problem with dynamic memory allocation (these are huge matrices). For this reason all of this computation is done off-line and stored. For material properties of skin we have used the data used by Pieper [72] for his facial surgery simulator. The fine mesh gets its input from the higher frequency motion analysis. Additionally, the motion that has already been computed and applied to the coarse mesh is propagated down to this level.

Muscles (at present there are 36 muscles in our system) are attached to this mesh with insertion and origin points also extracted from Pieper. Pieper got this data by scanning in an anatomical atlas of a face and warping it to his data. We were able to accomplish a similar warping. The muscles in our model are shown in Figure 4-8(b).

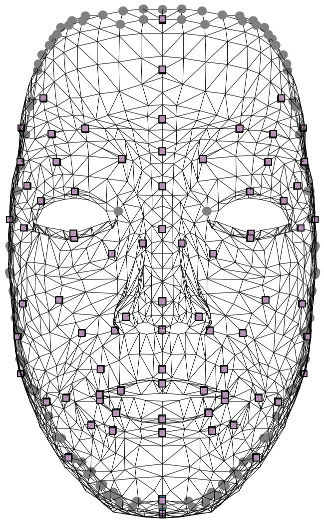
Figure 7-4 shows our face model in wireframe and with Gouraud shading. Also shown are the 80 regions of the face. In the next chapter we will bring together the concepts of motion estimation from Chapter 3 and the modeling concepts discussed in this chapter.



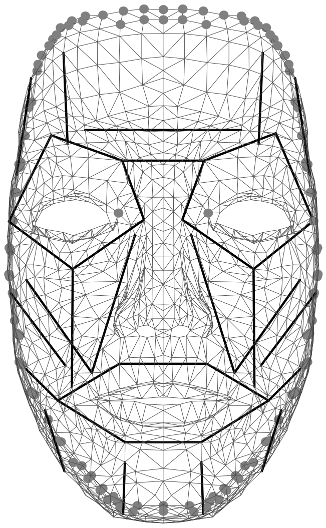
(a) Point Data *(b) Regions*

(c) Region Centers *(d) Coarse Mesh*

Figure 4-7: *(a) Shows the point data for a generic face in 3-D, (b) shows the different facial regions, (c) shows the centroids of the facial regions and the fixed nodes (shown as bigger grey dots) and (d) shows a coarse mesh using region centroids.*



(a) Fine Mesh



(b) Mesh + Muscles

Figure 4-8: *(a) Shows the fine mesh, and (b) show mesh with muscles.*

Chapter 5

Dynamics: Control and Estimation

Just as we learn to interpret certain types of changes as representing the motions of objects in the physical realm, we learn to classify other types of changes as signifying mental events; these are what we call "gestures" and "expressions." ... the destiny of each central "trajectory-type" agent is to learn to recognize, not a particular person but a particular type of gesture or expression.

Marvin Minsky, *The Society of Mind* (pp 312-313) [60]

5.1 Introduction

Muscles are the real effectors of facial motion, and all facial expressions are formed by actuation of dynamic muscles spread over the whole face. This suggests the need for modeling the face and the facial patterns dynamically. Modeling of the dynamic state evolution is important since visual motion sensing measures the change in facial appearance resulting from expression changes over time. Additionally, as the state measurements and the states evolve over time, they are constrained by the time-varying muscle actuations that are the cause of changes in the facial pattern. For this reason we model our facial analysis system as a dynamic state-space system with state variables that describe the muscle actuations and facial patterns in space and time. In this chapter we present the underlying concepts that combine the visual motion estimation and analysis methods discussed in Chapter 3 with the synthesis and modeling aspects presented in Chapter 4. We will first describe how we initialize the system and then discuss how the data from images is applied to a 3-D model of a face. Then we will introduce our *control theoretic* framework for detailed extraction of facial motion by control and estimation of motion with communication and feedback between our image measurements and the 3-D face model.

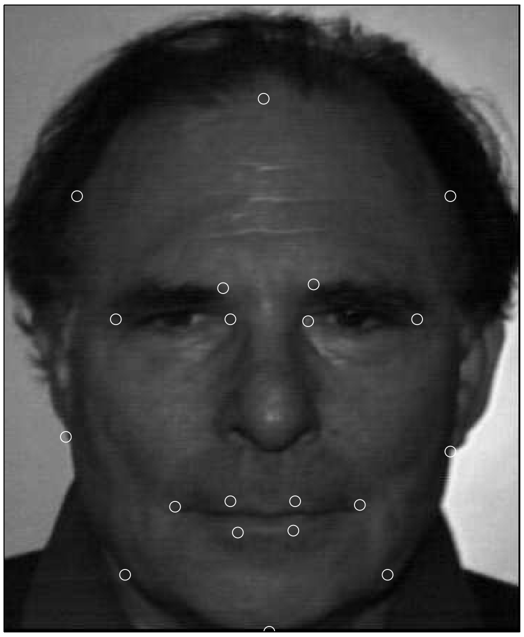
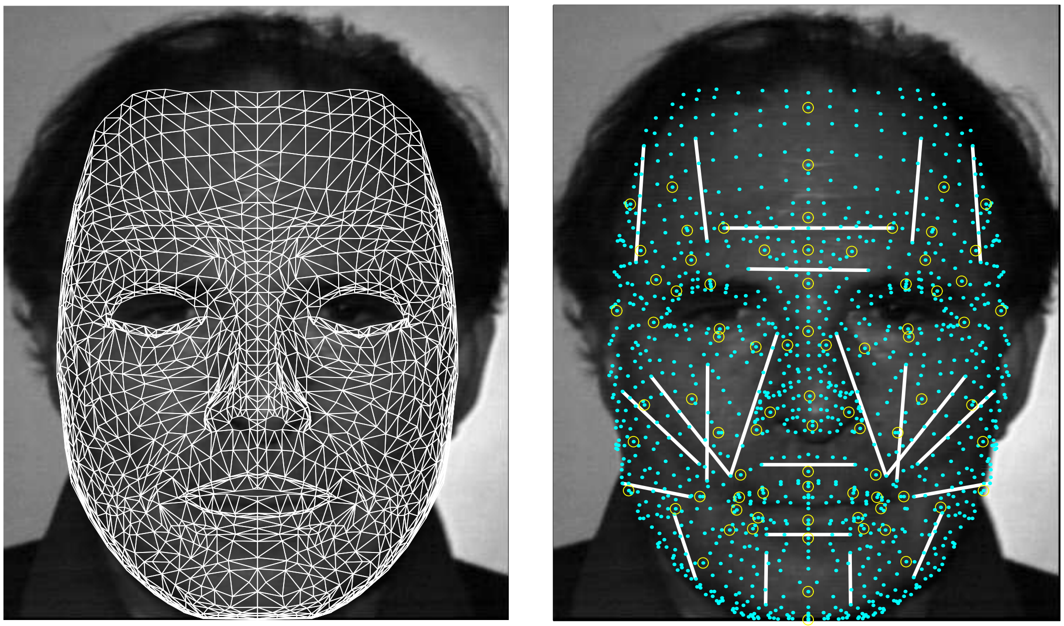


Figure 5-1: *Face showing landmarks used in initial placements and warping of facial mesh.*

5.2 Initialization of a facial mesh on an image

Extracting data from a vision system and mapping it onto a polygonal mesh requires good initial estimates of structure and location of the polygonal mesh. Therefore, initialization of a face mesh onto an image of a face is an important issue. A number of researchers are working on locating faces in images and extracting orientations [87, 68, 100, 15]. It is apparent that the feature-based deformable template work by Yuille *et al.* [100] or the pattern recognition aspects of Turk and Pentland *et al.* [87, 68] can be used to achieve a semi-automatic placement of the deformable facial mesh on a face in an image.

In our system we initialize the facial mesh on the frontal face image by hand. Using an interactive computer interface, a deformable mesh is placed on the image, then scaled, positioned and rotated to fit the general shape of the face. Then nonrigid warping of the mesh is done to conform the mesh to some well defined features on the face, *e.g.,* eyes, nose and mouth. The landmarks near the eyes, nose and mouth used for interactive placement of the mesh are shown in Figure 5-1. Some of these landmarks, with the points on the outer edge of the mesh, are then used as fixed nodes to distinguish between rigid and nonrigid motion as discussed in the previous chapter. Since these features are well-defined and can be easily extracted (unlike cheeks and eyeballs) it is feasible that a template matching or a feature extraction algorithm can be applied to locate the general location of the mesh. After the initial placement, the system can accommodate the deformations and the translations on the basis of its observer mechanics, thus determining the global and local motion in an image. Figure 5-2 shows the mesh, with all the muscles, accurately placed over a face.



(a) Mesh (b) Muscles

Figure 5-2: *(a) Face image with a mesh placed accurately over it and (b) Face image with muscles (white lines), region centers (circles) and nodes (dots).*

This mesh after initial placement can also be used as a mask to "mask" out all motion, except the motion related to the face (Figure 5-5).

5.3 Images to Facial Mesh

2-D to 3-D mapping

All the computations of facial motion are performed on 2-D images, however we are specifically interested in representating and analyzing facial motions in 3-D. Therefore, extracting two-dimensional data from images and then mapping it onto three-dimensional geometric models is an important part of our analysis framework. We accomplish this by using a simple projection mapping from two-dimensional images to three-dimensional geometric models. We can afford to apply this simplified method as all the facial images we are interested in processing are frontal views and we have a very good representation of a generic face model. Now we present the formulations of our method for mapping data from images to 3-D models.

In our formulations of the motion field we defined the velocity computed from vision sensing as $\mathbf{v}_i$. Now using the a mapping function, $\mathcal{M}$, we would like to compute velocities for the vertices of the geometric

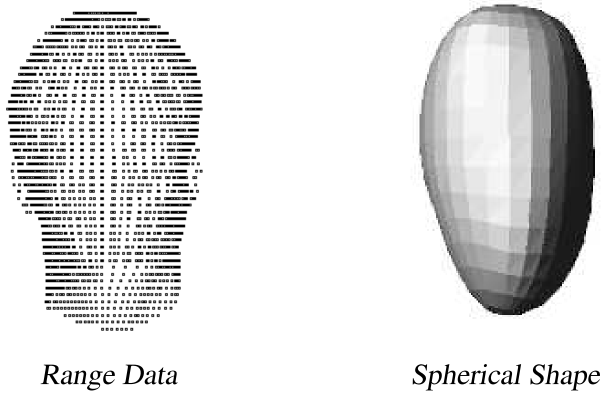


Figure 5-3: *Range data of a head and a 3-D spherical shape (Extracted from [32, 69]).*

model $\mathbf{v}_g$:

$$\mathbf{v}_g(x, y, z) = \boldsymbol{\mathcal{M}}(x, y, z)\mathbf{v}_i(x, y). \tag{5.1}$$

Using the physically-based modeling techniques of Section 4.2 and the relevant geometric and muscle models we can extract from $\mathbf{v}_g$ the forces that caused the observed motion.

Since we are mapping global information from an image to a geometric model, we have to concern ourselves with translations (vector $\boldsymbol{\mathcal{T}}$), and rotations (matrix $\boldsymbol{\mathcal{R}}$). Because our geometric model is deformable, we also need to account for its nonrigid behavior. The polynomial interpolation function $\mathbf{H}$ and the strain-displacement function $\boldsymbol{\mathcal{B}}$, from the finite element method (Equation (4.5)) provide information about the deformable behavior of the model [1].

Face shapes may be accurately parameterized using spherical coordinates (see Figure 5-3) and we use this constraint to extract motion in the third axis. For this we define a function that does a spherical mapping $\boldsymbol{\mathcal{S}}(u, v)$, where u and v are the spherical coordinates. The spherical function is computed by first selecting a standard three-dimensional geometric model of a face, with a spherical parameterization, and then using this canonical 3-D model to wrap the image onto the shape. Figure 5-4 illustrates how the intensity values for a face and a geometric face model may be projected onto the (u, v) coordinates. Using this approach, we determine the image-to-model equation to be:

$$\mathbf{v}_g(x, y, z) \approx \mathbf{H}\boldsymbol{\mathcal{S}}\boldsymbol{\mathcal{R}}\left(\mathbf{v}_i(x, y) + \boldsymbol{\mathcal{T}}\right). \tag{5.2}$$

[1] For more details on this kind of mapping see [32, 80, 69]. We could improve this mapping using structure from motion (*e.g.,* [58, 4]), but that takes us beyond the scope of this work.

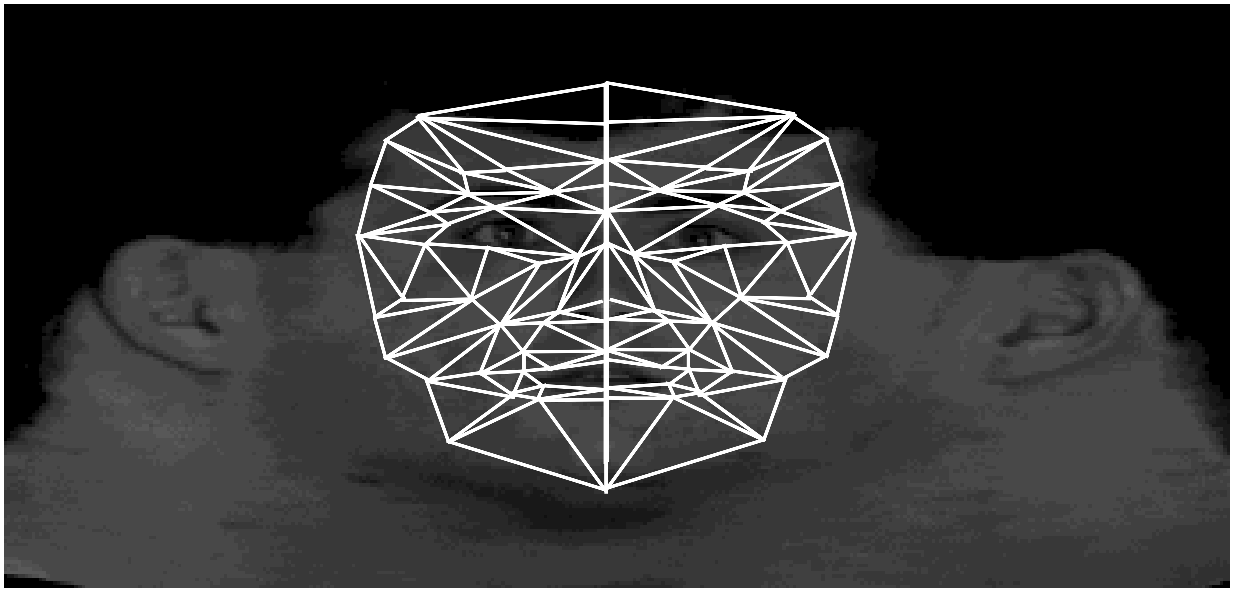


Figure 5-4: *Facial Mesh placed on intensity values of a face and projected on spherical coordinates* (u, v).

Mapping motion to a discretized mesh

Optical flow computation, done at full image resolution, provides us with velocity information at each and every pixel of the image. On the other hand our facial model is a 3-D mesh that has sampling points at discrete locations, namely at its vertices. We need to determine a method to apply the motions from an image to all the nodes and distribute the motion in a way that avoids chaotic behavior.

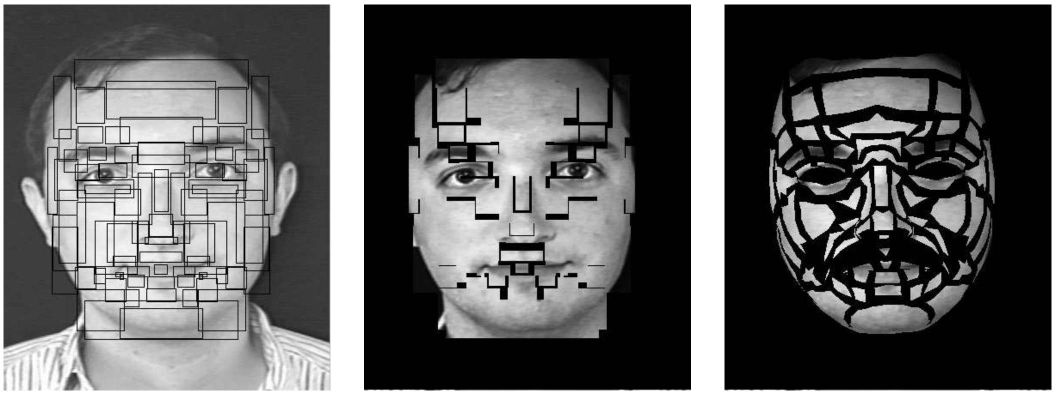
As it was described in the last chapter, we use a multi-grid mesh to model facial motion. We also have available to us a complete description of the facial structure including constraints on motion on the different parts of a face due to connectivity of the facial regions (*i.e.,* lower lip is below the upper lip *etc.*). To initialize the coarsest level of our multigrid representation, we use the bounding boxes that encompass these regions to describe windows of observation for these regions (see Figure 5-6 (a) and (b)) and then compute the flow under the area of the region (see Figure 5-6 (c)). The flow is then averaged across the region and applied to the centroid of the region. In this way all regions apply their motions separately and the energy surface defined by the finite element mesh deforms proportionately to the combined motion of all regions.

After propagating this motion across the coarse mesh, it is transformed to the fine mesh. The remaining higher-frequency motion is applied to the fine mesh using the distributed weighting functions $\mathcal{F}$, within each triangle of the fine mesh.

These weighting functions assign weights to the specific regions of the face based on the number of nodes of the facial geometry in that region. This weighting is similar to one used in computing concentrated loads on a mesh when loaded by a distributed load; here the same approach is used to compute motion at the

(a) Face Mask *(b) Masked Face Image*

Figure 5-5: *Face mask to compute motion.*



(a) Region Windows *(b) Region Window Images* *(c) Regions*

Figure 5-6: *Facial region to compute motion.*

vertices. This weighting function is also used to account for the confidence information as provided by the optical flow computation, described earlier. Naturally, it can also be used to calculate the forces and stress at the nodes of the FEM model, and the resulting stress distributions. The final mapping equation using this

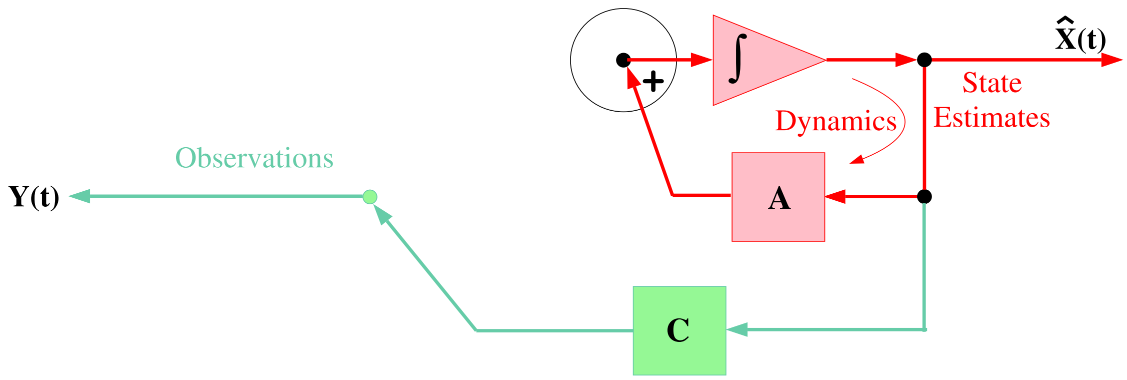


Figure 5-7: *Block diagram for dynamics.*

weighting function is:

$$\mathbf{v}_g(x, y, z) = \boldsymbol{\mathcal{F}}\mathbf{H}\boldsymbol{\mathcal{S}}\boldsymbol{\mathcal{R}}\left(\mathbf{v}_i(x, y) + \boldsymbol{\mathcal{T}}\right). \tag{5.3}$$

From now on, whenever, we talk about velocities, we will assume that the above mapping has already been applied, unless otherwise specified.

5.4 Estimation and Control

Despite the care taken in extracting motion from flow, the process of driving a physical system with the inputs from video is prone to errors (due to intensity singularities, lighting conditions, temporal aliasing, bad video quality, *etc.*) and can result in divergence or a chaotic physical response. This is why an estimation and control framework needs to be incorporated in order to obtain stable and well-proportioned results.

Figure 5-11 shows the estimation and control elements of our active facial expression modeling and analysis framework. The next few sections will explain the details of this framework and construct the Figure 5-11 by discussing each and every segment of the dynamic system separately.

In considering a dynamic system in a state-space notation, we will employ notation that is consistent with the notation in control theory literature. The notation used for motion estimation and Kalman filtering is of the standard form used by the vision and image-coding community.

Dynamic system with noise

In a dynamic system with state $\mathbf{X}$, control input vector $\mathbf{U}$, and measurement vector $\mathbf{Y}$, we may write the dynamic system in continuous time state-space form:

$$\dot{\mathbf{X}}(t) = \mathbf{AX}(t) + \mathbf{BU}(t) + \mathbf{Gn}_p(t), \tag{5.4}$$

This is known as a *dynamic state evolution equation* or a *process equation*. On the other hand, for observations and measurements we need a *measurement equation*, where the measurements $\mathbf{Y}$ are determined from states $\mathbf{X}$, and inputs $\mathbf{U}$:

$$\mathbf{Y}(t) = \mathbf{CX}(t) + \mathbf{DU}(t) + \mathbf{n}_m(t), \tag{5.5}$$

Equation (5.5) is called the *measurement model* corresponding to the *process model*. Throughout our formulation we assume that there is no relationship between control and observations (hence $\mathbf{D} = 0$). In both the above models we have incorporated a noise model to account for the process noise $\mathbf{n}_p$ and the measurement noise $\mathbf{n}_m$. The matrix $\mathbf{G}$ defines the coupling of the noise process $\mathbf{n}_p$ with the dynamic system. These noise models are aimed at characterizing systems that we are sure are prone to errors. We assume that $\mathbf{n}_p$ and $\mathbf{n}_m$ are *uncorrelated zero-mean white noise processes* with covariances $\boldsymbol{\Lambda}_p$ and $\boldsymbol{\Lambda}_m$, respectively. We also assume that the initial state is also uncorrelated with the measurement and process noises.

To assist in our discussion of the *control-theoretic* framework for analysis using separate models for observation (analysis) and control (synthesis), we will fragment our representation into separate domain of observation and control. We will conclude this chapter by combining them and discussing the complete framework.

First we will consider a simple observer dynamics system, in which change in state over time is observed and there is no input to the system ($\mathbf{U} = 0$). A simple graphical representation of such a system is shown in Figure 5-7. Most of the vision systems that are aimed at observing motion are described by this system. The systems for facial expression recognition presented by Mase [56] and Yacoob [99] also fall into this category of systems. These systems are simply observing the dynamic system (of motion from image sequences), and making rule-based decisions for recognition.

Prediction, Estimation, and Correction

The dynamic system described in the previous section does consider a noise model, but no attempt was made for error prediction and correction using this noise model. Motion estimation problems are highly prone to errors. Consequently, the technique of Kalman filtering has found wide application in motion estimation problems in the computer vision community [58]. We will also employ this technique for the prediction and error correction of motion with a dynamic model of a face.

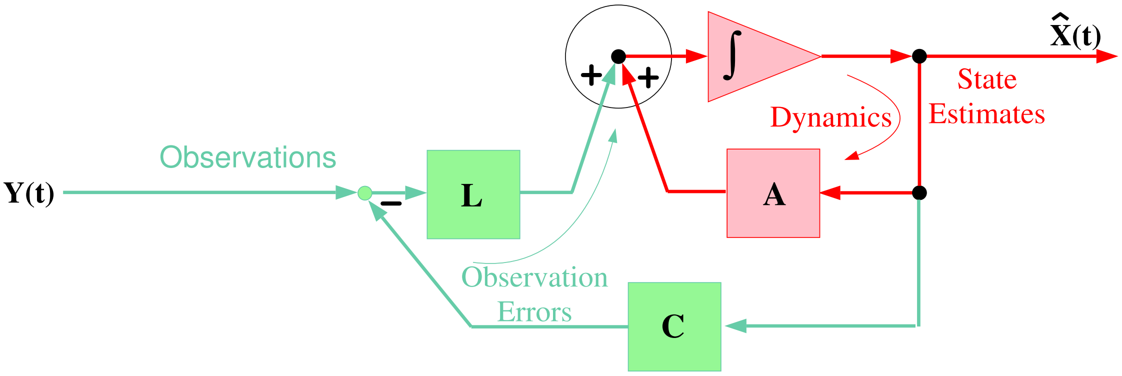


Figure 5-8: *Block diagram for dynamics with Kalman Filter for Error Prediction and Correction. Note the additional loop at the bottom.*

The continuous time Kalman filter (CTKF) allows us to estimate the uncorrupted state vector, and produces an optimal least-squares estimate under quite general conditions [12, 34]. The Kalman filter is particularly well-suited to this application because it is a recursive estimation technique, and so does not introduce any delays into the system (keeping the system active). Note that we have already used Kalman filters earlier, in our discussion of estimation of motion over scale and time (Chapter 3) but with a different dynamic model.

The CTKF for the above system is established by the following formulation:

$$\dot{\hat{\mathbf{X}}} = \mathbf{A}\hat{\mathbf{X}} + \mathbf{B}\mathbf{U} + \mathbf{L}\left(\mathbf{Y} - \mathbf{C}\hat{\mathbf{X}}\right), \tag{5.6}$$

where $\hat{\mathbf{X}}$ is the linear least squares estimate of $\mathbf{X}$ based on $Y(\tau)$ for $\tau < t$. Let Λ_e be the error covariance matrix for $\hat{\mathbf{X}}$ then:

$$\mathbf{L} = \Lambda_e \mathbf{C}^T {\Lambda_m}^{-1}, \tag{5.7}$$

is the Kalman gain matrix. The Kalman gain matrix $\mathbf{L}$ is obtained by solving the following Riccati equation to obtain the optimal error covariance matrix Λ_e:

$$\frac{d}{dt}\Lambda_e = \mathbf{A}\Lambda_e + \Lambda_e\mathbf{A}^T + \mathbf{G}\Lambda_p\mathbf{G}^T - \Lambda_e\mathbf{C}^T{\Lambda_m}^{-1}\mathbf{C}\Lambda_e. \tag{5.8}$$

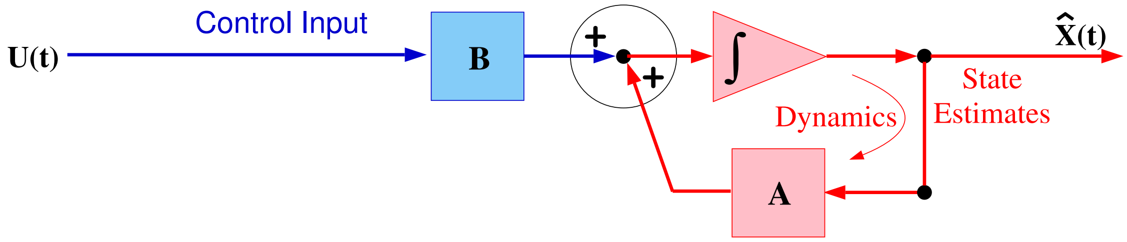


Figure 5-9: *Block diagram for dynamics with Control Input. This is a typical synthesis system in which regulated control is inputed to get desired motion.*

The Kalman filter, Equation (5.6), mimics the noise free dynamics and corrects its estimate with a term proportional to the difference $(\mathbf{Y} - \mathbf{C}\hat{\mathbf{X}})$, which is the innovations process. This correction in the estimate is between the observation and our best prediction based on previous data. Note that the above equation (Ricatti equation (Equation (5.8))) can be solved and the gain Equation (5.7) computed off-line as they do not depend on actual values of measurements. Figure 5-8 shows a dynamic system with a Kalman Filter. Comparison of Figures 5-7 and 5-8 graphically shows the addition of an estimation loop (the bottom loop) which is used to correct the dynamics based on the error predictions of the system.

As the visual sensing part of the process has already established a good probability distribution for the motion observations we can simply use Equation (3.12) and Equation (3.13), for each frame at time t, in our dynamic observations relationship in Equation (5.5). Hence using the mapping criteria as discussed earlier we obtain:

$$\begin{aligned}\boldsymbol{\Lambda}_m(t) &= \boldsymbol{\mathcal{M}}(x,y,z)\boldsymbol{\Lambda}_{\mathbf{v}}(t),\\ \mathbf{Y}(t) &= \boldsymbol{\mathcal{M}}(x,y,z)\hat{\mathbf{v}}_i(t)\end{aligned}$$

Control of Dynamic Motion

Now let us take a different view of a dynamic system and assume we have a dynamic system with a controlled input, $\mathbf{U}$ but no observations, ($\mathbf{Y} = 0$). This system, shown in Figure 5-9, represents a simple simulation and synthesis platform, in which an input is used to drive a system. This system is typical of facial animation systems that are dynamically controlled.

However, we are interested not only in simulation, but in observing a person and "mimicking" him/her. We are specifically interested in the *inverse dynamics* of the system; we want to extract the muscle actuations that caused the motion. This suggests the use of the *control feedback law* to obtain the muscle activations.

The control input vector $\mathbf{U}$ is provided by the control feedback law:

$$\mathbf{U} = -\mathcal{G}\mathbf{X}, \tag{5.9}$$

where $\mathcal{G}$ is the *control feedback gain matrix*. We assume that the instance of control under study falls into the category of an *optimal regulator*. This optimality criteria is needed to extract a unique set of control actuation to define the motion under analysis. An optimal state quadratic regulator has the following quadratic cost function [34, 46]:

$$\mathcal{J} = \int_{t_0}^{t_f} \frac{1}{2}\left[\mathbf{X}^T(t)\mathbf{Q}\mathbf{X}(t) + \mathbf{U}^T(t)\mathbf{R}\mathbf{U}(t)\right] dt \tag{5.10}$$

where $\mathcal{J}$ is the cost, $\mathbf{Q}$ is a real, symmetric, positive semi-definite *state weighting* matrix and $\mathbf{R}$ is a real, symmetric, positive definite *control weighting* matrix. This cost function enforces constraints of smooth state-evolution and minimizes energy. An optimal state regulator system can be controlled by a combined inverse dynamics and optimal regulator scheme.

The optimal control law $\mathbf{U}^*$ that minimizes the cost $\mathcal{J}$ is given by:

$$\mathbf{U}^* = -\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_c\mathbf{X}^* \tag{5.11}$$

where $\mathbf{X}^*$ is the optimal state trajectory and $\mathbf{P}_c$ is given by solving yet another *matrix Riccati equation* [34]:

$$-\dot{\mathbf{P}}_c = \mathbf{A}^T\mathbf{P}_c + \mathbf{P}_c\mathbf{A} + \mathbf{Q} - \mathbf{P}_c\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_c \tag{5.12}$$

$\mathbf{P}_c$ can be determined from numerical integration of Equation (5.12). In case of a steady-state system, $\dot{\mathbf{P}}_c = 0$, making the above equation quadratic in $\mathbf{P}_c$, making it easy to solve. Comparing with Equation (5.9) we obtain

$$\mathcal{G} = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_c$$

Figure 5-10 shows the dynamic system with feedback control input. Comparison between Figure 5-10

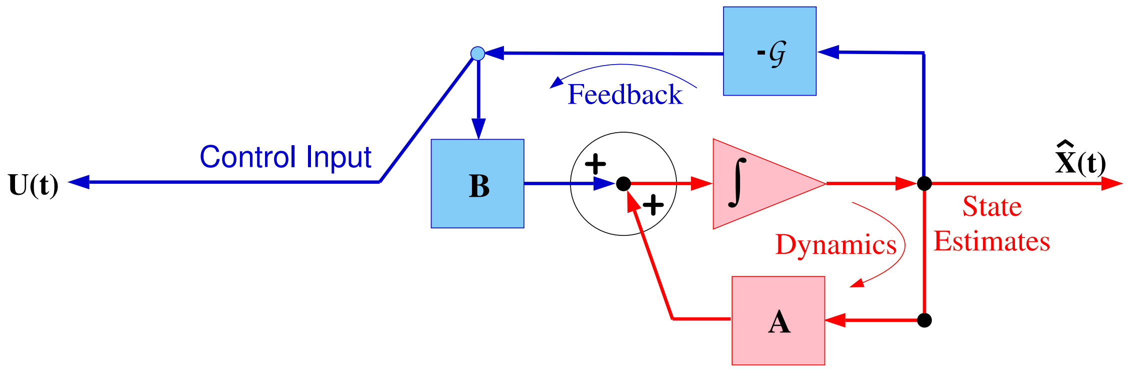


Figure 5-10: *Block diagram for dynamics with Control Input with an added Feedback.*

and Figure 5-9 shows an additional loop for feedback control in Figure 5-10.

Complete System

Combining the observation and estimation system of Figure 5-8 with the feedback control system of Figure 5-10, we get the complete system that we use for analysis of facial expressions. This system is shown in Figure 5-11. In our system the state variable, $\mathbf{X}$, is represents a vector of positions, velocities and accelerations of the facial mesh. Hence $\mathbf{A}$ is filled with equations of dynamic motion. Our observations, $\mathbf{Y}$, are velocities, and therefore, $\mathbf{C}$ is mainly filled with zeros and ones to extract velocities from the state vector. The control vector $\mathbf{U}$ is the muscle actuations, *i.e.,* the load vector $\mathbf{R}$ from Equation (4.1), and $\mathbf{B}$ transforms these forces using the stiffness, mass and damping matrices.

An important point to note about this system is that it runs in the opposite direction from traditional animation systems; this system is driven by observations from images; control measurements are then extracted and fed back to actively drive the system. We have found this system to be quite stable throughout our analysis and experimentation.

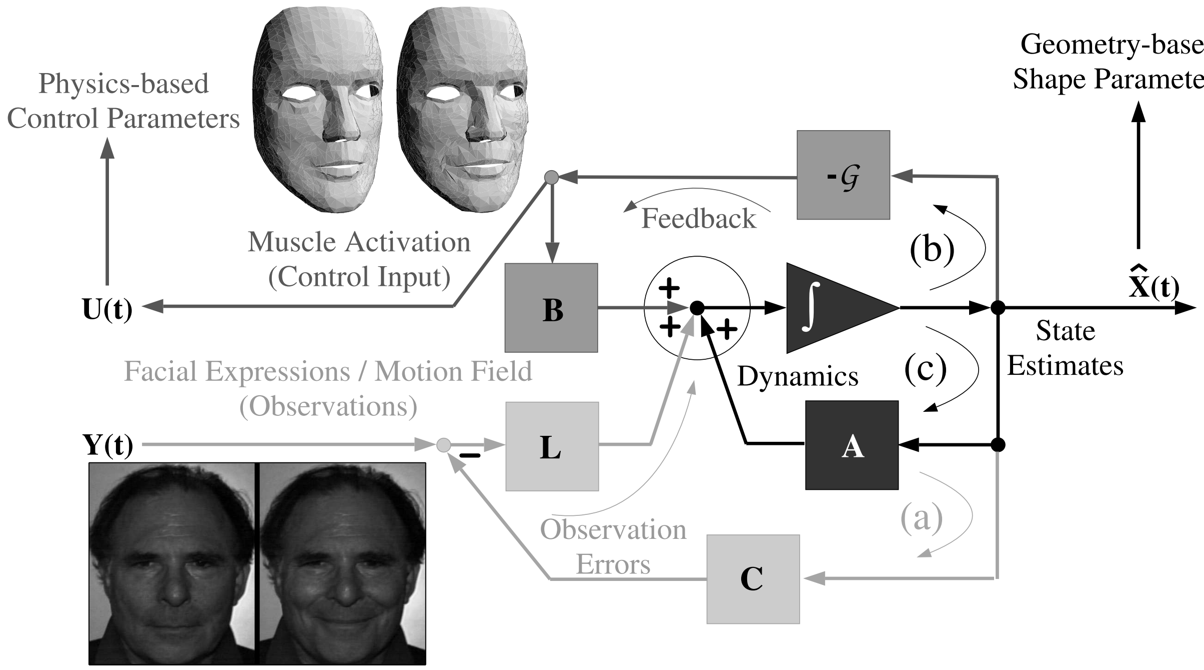


Figure 5-11: *Block diagram of the control-theoretic approach. Showing the estimation and correction loop (a), the dynamics loop (b), and the feedback loop (c).*

Chapter 6

Analysis, Identification and Synthesis

> *The first [zygomatic major] obeys the will but the second [orbicularis oculi] is only put in play by the sweet emotions of the soul; the ... fake joy, the deceitful laugh, cannot provoke the contradiction of this latter muscle.* (p 126) *[This muscle] ... does not obey the will; it is only brought into play by a true feeling.* (p 72)
>
> B. Duchenne, *The Mechanism of Human Facial Expressions*[24][1].

6.1 Analysis of Facial Motion

One of the main goals of this work is to devise methods that can be used to develop a more accurate model of facial action (FACS+). The current state-of-the-art for facial description (either FACS itself or the muscle-control versions of FACS) has the following limitations and weaknesses:

- FACS, a system designed to code facial movements is entirely based on studying static photographs of subjects after an expression has been made. No *quantitative* information is available to describe the facial motion leading to the expression. *Qualitative* information does exist [28, 7], however, such information is difficult to make a part of a representation and modeling paradigm.
- The action units are purely local spatial patterns. Real facial motion is almost never completely localized; Ekman himself has described some of these action units as an "unnatural" type of facial movement.
- There is no time component of the description, only a heuristic one. From EMG studies it is known that most facial actions occur in three distinct phases: *application*, *release* and *relaxation*. In contrast, current systems typically use simple linear ramps to approximate the actuation profile.

[1] extracted from [27]

- FACS does not have the ability to describe fine eye and lip motions. Perhaps most limiting is the inability to describe the coarticulation effects found commonly in speech.

Consequently, a major focus of experimental work in this thesis has been to characterize the functional form of the actuation profile, and to determine a newer (near-orthogonal) basis set of "action units" that better describes the spatial and temporal properties of real facial motion.

6.1.1 An Improved Representation for Facial Motion

This need for an improved representation resulted in a method that extracts control parameters for facial motion by quantitative observation of real people making expressions. The extracted parameters can control much more complex motions than is typical using the current day FACS and/or muscle models. Experiments (described in detail in Chapter 7) show that a smile, even though primarily due to actuation of muscles in the lower part of the face, is not complete without some facial deformation in the upper part of the face. This result corroborates statements of Darwin [22], Duchenne [24] and Ekman [27] who observed that the actuation of upper muscles is a significant part of a "real" smile.

These new control parameters can be described mathematically as follows: Consider a basis set $\boldsymbol{\Phi}_g$, which has n vectors, $\phi_o, \ldots, \phi_n$. Each of the basis vectors, ϕ_i is a deformation profile of a face for a specific action. In a static case, with reduced sampling of facial motion, these vectors would typically be FACS-like action units (*i.e.,* the CANDIDE model, which is a computer graphics model for implementing FACS motions [79, 28]).

The visual observation and estimation process extracts information about the time-evolving deformation profile of the face by extracting a new dynamic basis set $\boldsymbol{\Phi}_g$, using principal component analysis. This new basis set $\boldsymbol{\Phi}_g$, can be used as a "rotation" matrix to "rotate" polygon vertex displacements, $\mathbf{u}$ (see Chapter 4) to a new *generalized set of displacements* $\tilde{\mathbf{u}}$:

$$\mathbf{u} = \boldsymbol{\Phi}_g \tilde{\mathbf{u}}. \tag{6.1}$$

After the analysis of several people making expressions (see Chapter 7) we see that the resulting generalized displacements show distinct characteristics for different expressions. In the range of all expressions, the characteristics of each expression are easily identifiable. Similar results for lip reading and expression recognition using FACS were obtained by Mase and Pentland [56, 57], although within a static estimation framework. Deformations and actuations in both space and time for different expressions are quite distinct and will be discussed later. This basis set also functions as a set of constraints on the system and is used to determine the control input for different expressions.

Another important transformation is the transformation of the nodal forces $\mathbf{R} = \mathbf{H}\bar{\mathbf{R}}$ to a set of *generalized loads* $\tilde{\mathbf{R}}$ [2].This transformation requires another basis set $\boldsymbol{\Phi}_p$, with each of its vectors defining muscle actuations causing different facial expressions.

$$\mathbf{R} = \boldsymbol{\Phi}_p \tilde{\mathbf{R}}. \tag{6.2}$$

This *force-based* basis set is obtained by mapping nodal forces and the causal nodal deformations to the parametric representation of muscles actuations (rather than just geometric deformations as in the case of deformation basis $\boldsymbol{\Phi}_g$). Let $\mathbf{G}$ be this mapping function, using which we obtain:

$$\boldsymbol{\Phi}_p = \mathbf{G}\boldsymbol{\Phi}_g. \tag{6.3}$$

Application of principal component analysis on a system of equations defined by the finite element method requires *Eigenvalue Analysis*. Using concepts defined in [8, 32] for Modal Analysis on Equation (4.1) the generalized eigenvalue problem is set up as:

$$\mathbf{K}\boldsymbol{\Phi}_g = \mathbf{M}\boldsymbol{\Phi}_g \boldsymbol{\Omega}^2, \tag{6.4}$$

where $\boldsymbol{\Phi}_g$ has for its columns the eigenvectors ϕ_i, corresponding to the eigenvalues ω_i^2, which fill the diagonal eigenvalue matrix $\boldsymbol{\Omega}^2$. This results in (see [8, 32]):

$$\tilde{\mathbf{R}} = \boldsymbol{\Omega}^2 \tilde{\mathbf{u}}. \tag{6.5}$$

Using Equations (6.1), (6.2) and (6.5), with $\mathbf{R} = \mathbf{K}\mathbf{u}$ on Equation (6.3) we get:

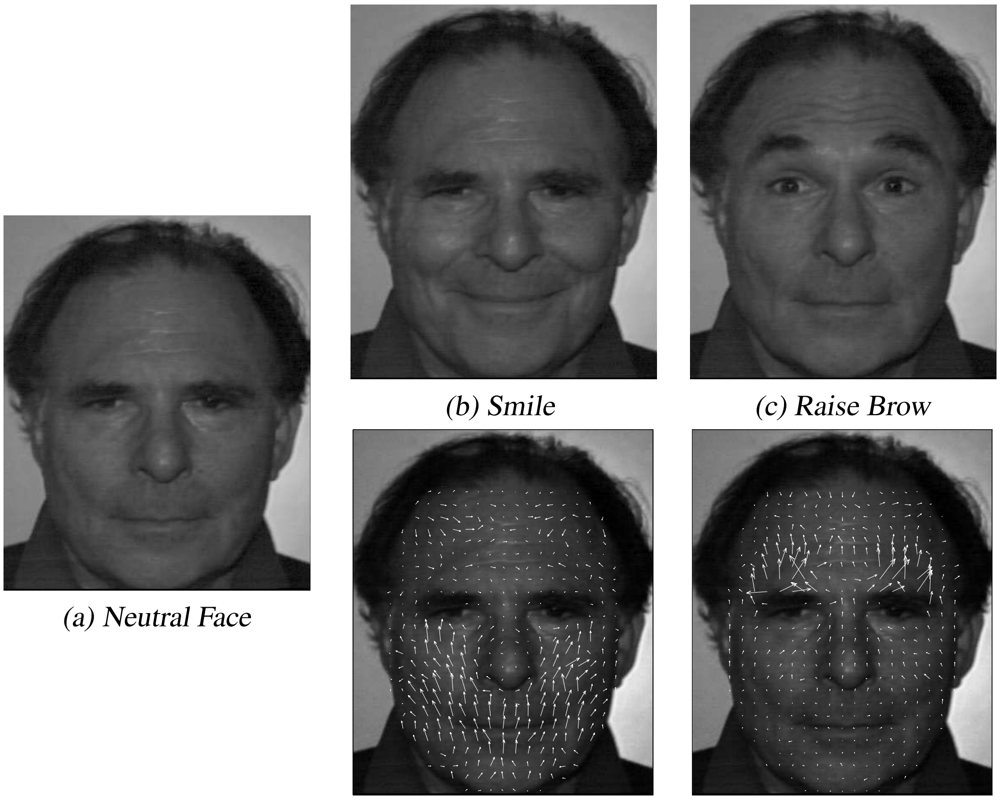
$$\mathbf{G} = \frac{\mathbf{R}}{\tilde{\mathbf{R}}}\frac{\tilde{\mathbf{u}}}{\mathbf{u}} = \frac{\mathbf{K}}{\boldsymbol{\Omega}^2} = \mathbf{K}\left(\boldsymbol{\Omega}^2\right)^{-1}. \tag{6.6}$$

Using these relationships the mapping function $\mathbf{G}$ is precomputed for the model and then the basis set, $\boldsymbol{\Phi}_p$ is computed actively during the control loop. The application of such a "near-orthogonal" basis set is less computationally expensive as there are fewer muscle descriptors than geometric descriptors and the muscle descriptors are independent of the topology of the face. The following sections present details on important features of this new representation.

6.1.2 Evaluation of the improved representation

For the purpose of discussion we have chosen two expressions for detailed analysis: *eyebrow raising* (or $AU2$ in FACS), is traditionally known to be an expression predominantly resulting in motion on the upper regions of the face and *smiling* (or $AU12$ in FACS), which is, in earlier models, limited to motion in the

[2] where $\mathbf{H}$ is the interpolation matrix and $\bar{\mathbf{R}}$ are loads at the nodes of the finite element mesh as described in Chapter 4



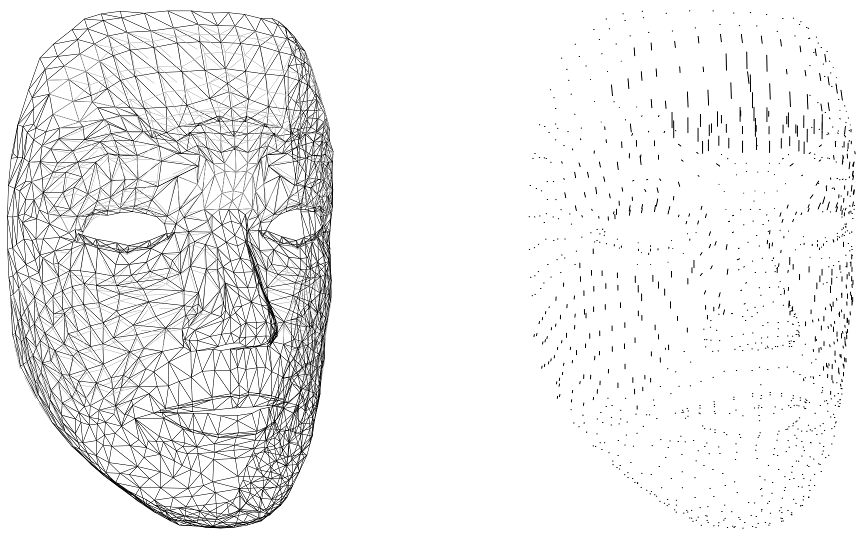
(a) Neutral Face

(b) Smile

(c) Raise Brow

(d) and (e) Motion Fields

Figure 6-1: *Expressions from video sequences for Paul Ekman. (a) neutral expression, (b) eighth frame from a smile sequence and (c) the eighth frame from a raising brow sequence, (d) and (e) motion fields for the smile and raising brow expressions*



(a) Overlayed (b) Difference

Figure 6-2: *(a) Vision-based FACS+ expression overlayed on top of a FACS expression of Raising Eyebrow, and (b) the differences between the two facial motions. It can be seen that modeling by visual measurement produces a more detailed pattern of motion.*

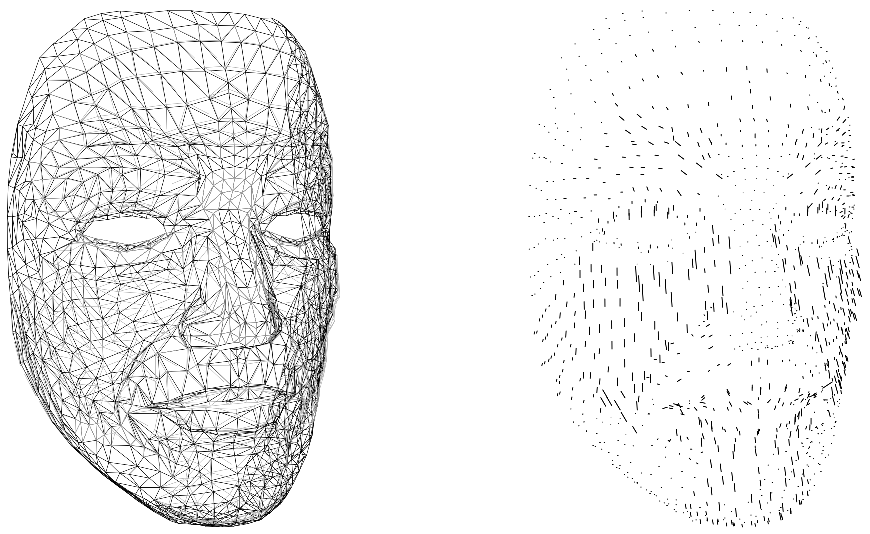
lower regions. Figure 6-1 shows examples of these expressions as performed by Paul Ekman, and recorded using the experimental set-up discussed in the next chapter. Results of the optical flow computation for these expressions are also shown.

Spatial Patterning

When fed into the control feedback loop shown in Figure 5-11, the dense motion measurements that are extracted from optical flow computations at every pixel, result in both geometric and physical (muscle-based) parameters. The physical parameters are the muscle activations that will produce the expression being analyzed.

To illustrate that the resulting parameters for facial expressions are more spatially detailed than FACS, comparisons of the expressions of *raising eyebrow* and *smile* produced by standard FACS-like muscle activations and our visually extracted muscle activations are shown in Figure 6-2 and Figure 6-3.

Figure 6-2 (a) compares the FACS motions with FACS+motions by overlaying the two for the raising eyebrow motion. Figure 6-2 (b) shows the difference in spatial motion. Similar comparison for the smile



(a) Overlayed (b) Difference

Figure 6-3: *(a) Vision-based FACS+ expression overlayed on top of a FACS expression of Smile, and (b) the differences between the two facial motions. It can be seen that modeling by visual measurement produces a more detailed pattern of motion.*

expression is shown in Figure 6-3. As expected, the two models are very similar in the primary FACS activation region. For the case of eyebrow raising, both models are similar in the area directly above the eyebrow. For the smiling example both models are similar in the area immediately adjacent to the mouth. However, as seen by these figures, there is a lot of additional motion across the face which the FACS model failed to account for. In both cases, the visual measurement model had significant additional deformations in distant areas of the face. In the case of eyebrow raising, the visual model has additional deformations high in the forehead, immediately above the eye, and in the lower cheek. In the case of smiling, there are additional deformations beneath and between the eyes, on the far cheek to either side of the mouth, and on the temples.

So far only qualitative descriptions have been described. To quantitatively evaluate these descriptions and their differences, we compare them to a FACS-based CANDIDE model. Figures 6-4 and 6-5 shows the difference in control parameters for both the expressions.

The top row of Figure 6-4 shows $AU2$ ("Raising Eyebrow") from the FACS model and the linear actuation profile of the corresponding geometric control points. This is the type of spatial-temporal patterning

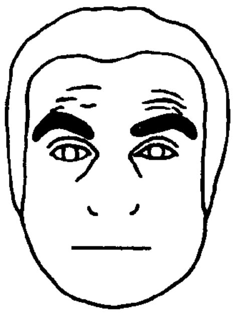
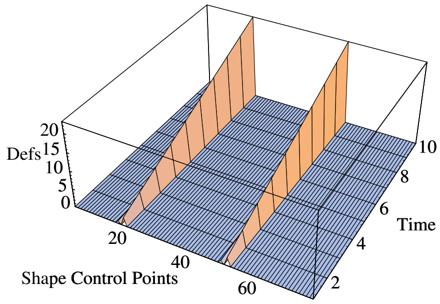
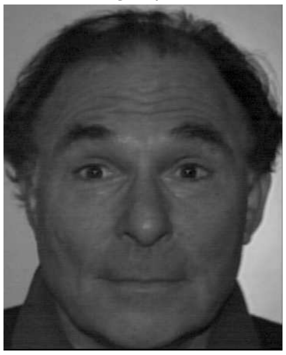
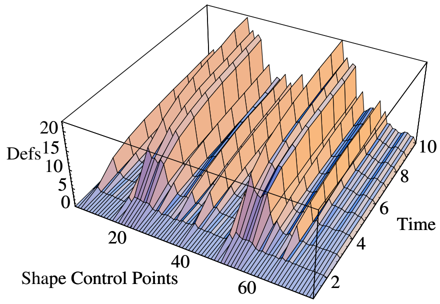
<table>
<tr><th>Expression</th><th>Magnitude of Control Point Deformation</th></tr>
<tr><td>$AU2$
</td><td>
</td></tr>
<tr><td>Raising Eyebrow.
</td><td>
</td></tr>
</table>

Figure 6-4: *FACS/CANDIDE deformation vs. Observed deformation for the Raising Eyebrow expression. Surface plots (top) show deformation over time for FACS actions $AU2$, and (bottom) for an actual video sequence of raising eyebrows.*

commonly used in computer graphics animation. The bottom row of Figure 6-4 shows the observed motion of these control points for the expression of *raising eyebrow* by Paul Ekman. This plot was achieved by mapping the motion onto the FACS model and the actuations of the control points measured. As can be seen, the observed pattern of deformation is very different than that assumed in the standard computer graphics implementation of FACS. There is a wide distribution of motion through all the control points, and the temporal patterning of the deformation is far from linear. It appears, very roughly, to have a quick linear rise, then a slower linear rise and then a constant level (*i.e.,* may be approximated as piece-wise linear).

Similar plots for smile expression are shown in Figure 6-5. By using these observed distributed patterns of motion more realistic image synthesis and computer animations can be produced.

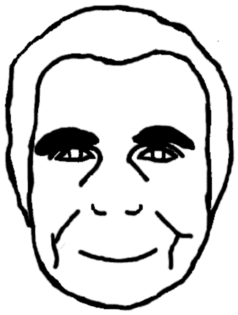
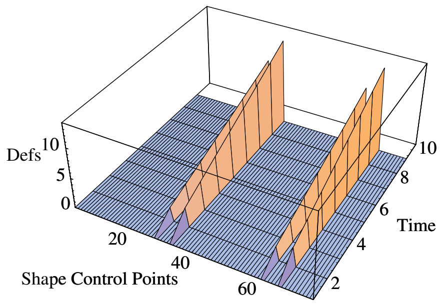
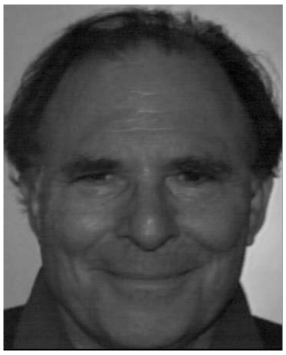
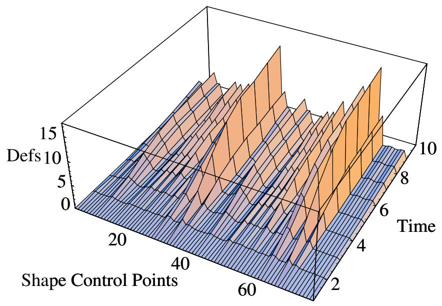
Expression	Magnitude of Control Point Deformation
$AU12$ 	
Smile. 	

Figure 6-5: *FACS/*C*ANDIDE deformation vs. Observed deformation for the Smile expression. Surface plots (top) show deformation over time for FACS action* $AU12$*, and (bottom) for an actual video sequence of smile.*

Temporal Patterning

Another important observation about facial motion that is apparent in Figures 6-4 and 6-5 is that the facial motion is far from linear in time. This observation becomes much more important when facial motion is studied with reference to muscles. Figure 6-6 and Figure 6-7 shows plots of facial muscle actuations for the same smile and eyebrow raising expressions. For the purpose of illustration, in this figure the 36 face muscles were combined into seven local groups on the basis of their proximity to each other and to the regions they effected. As can be seen, even the simplest expressions require multiple muscle actuations.

Of particular interest is the temporal patterning of the muscle actuations. We have fit exponential curves to the activation and release portions of the muscle actuation profile to suggest the type of rise and decay seen in EMG studies of muscles. From this data we suggest that the relaxation phase of muscle actuation is mostly due to passive stretching of the muscles by residual stress in the skin.

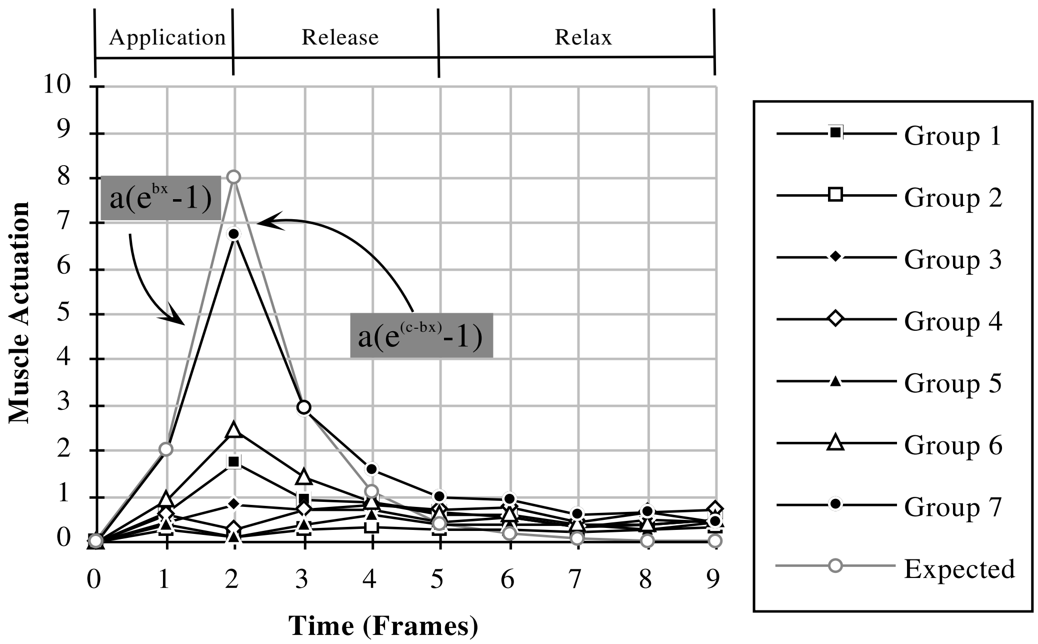


Figure 6-6: *Actuations over time of the seven main muscle groups for the expressions of raising brow. The plots shows actuations over time for the seven muscle groups and the expected profile of application, release and relax phases of muscle activation.*

Note that Figure 6-7 for the smile expression also shows a second, delayed actuation of muscle group 7, about 3 frames after the peak of muscle group 1. Muscle group 7 includes all the muscles around the eyes and as can be seen in Figure 6-6 is the primary muscle group for the raising eye brow expression. This example illustrates that coarticulation effects can be observed by our system, and that they occur even in quite simple expressions. By using these observed temporal patterns of muscle activation, rather than simple linear ramps, more realistic computer animations and synthetic images can be generated.

6.2 Identification and Recognition of Expressions

So far the emphasis has been on quantitative estimation of facial motion. This analysis produced a representation in which each expression is unique, suggesting that this same representation can be used for recognition and identification of facial expressions.

Recognition requires a unique "feature vector" to define each expression and a *similarity metric* to measure the differences between expressions. Since both temporal and spatial characteristics are extremely important we require a feature vector that can account for both of these characteristics. We must, however, normalize for the speed at which the expressions are performed. Since facial expressions occur in three distinct phases: *application*, *release* and *relaxation*, by dividing the data into these phases and by warping

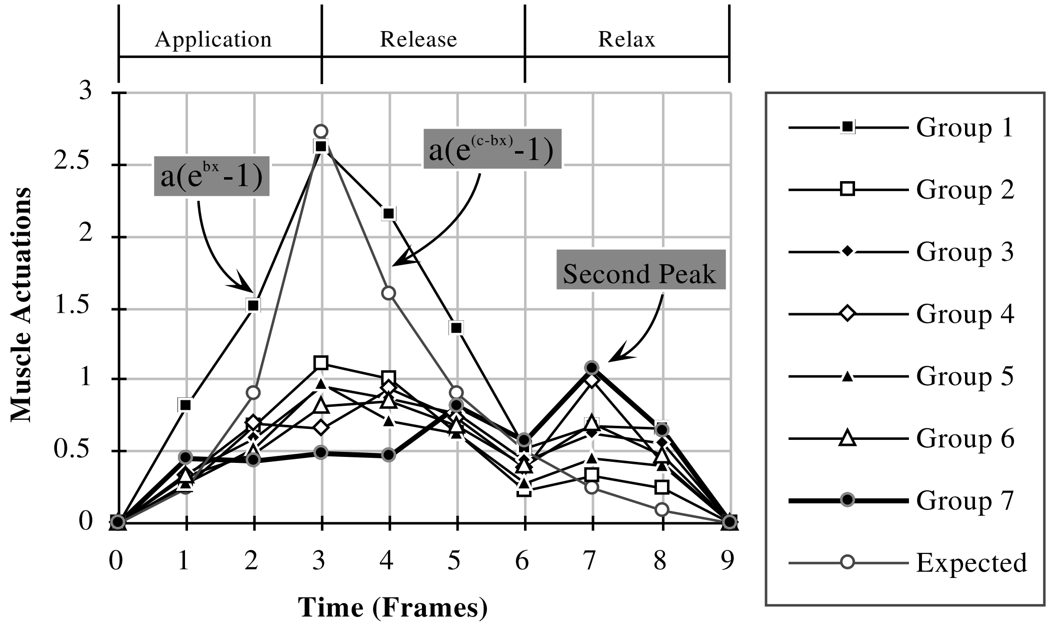


Figure 6-7: *Actuations over time of the seven main muscle groups for the expressions of smiling – lip motion. The plots shows actuations over time for the seven muscle groups and the expected profile of application, release and relax phases of muscle activation.*

it for all expressions into a fixed time period of ten discrete samples, we can take the temporal aspect out of the analysis. This normalization allows us to use the muscle actuation profiles to define a feature vector for a unique facial motion.

We define the peak actuation of each muscle between the application and release phases as the feature vector for each expression. These feature vectors are then used for recognition of facial expression using dot products, as presented in Chapter 7.

6.3 Synthesis and Animations of Expressions

One of the most important applications of the new and detailed facial motion representation is realistic synthesis and simulations of facial expressions. Once the control parameters are obtained (muscle-based or deformation based), we can impart these as control input to a generic polygonal face model to achieve synthesis of facial expressions. The resulting patterning of muscle activations and/or facial expressions can be more realistic and can provide significantly better simulations for facial animations.

One important aspect of the vision-based deformation and force extraction is that these control parameters can be *actively* imparted onto a facial model. This means that by using vision as well as dynamic estimation and control to develop the force or deformation basis, we can impart these "expression changes" as they are

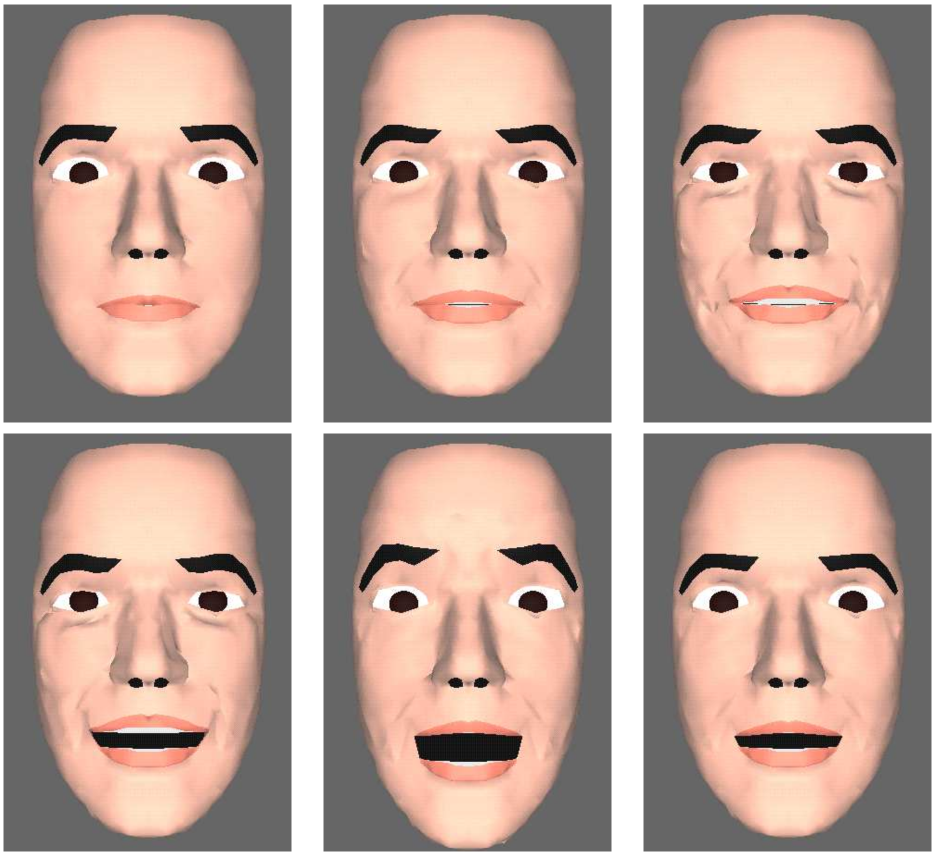


Figure 6-8: *Several Examples of Synthesis of Facial Expressions.*

being computed. This is of special interest, as it theoretically allows zero-lag animation, which is important in applications such as Teleconferencing. However, since the computations are time consuming and data transmissions bandlimited, it is doubtful that such a system is possible without new specialised hardware.

A simpler real-time system, discussed in Chapter 7, uses this representation for mimicking or tracking of facial expressions.

Chapter 7

Experiments and Results

The principle of science, the definition, almost, is the following: The test of all knowledge is experiment. Experiment is the sole judge of scientific "truth."

Richard Feynman, *Lectures on Physics* (p 1-1) [33]

This chapter describes experiments conducted using the method developed in the previous chapters. We present the data collection method and the system used to conduct these experiments and discuss the results in the areas of analysis, identification and synthesis. At the end of this chapter a real-time system (developed in collaboration with Trevor Darrell) is presented as an example of how the methods developed here can be used for interactive tracking and synthesis of facial expressions.

Data Acquisition

The first step in conducting these experiments was to acquire image sequences of subjects making expressions. For this purpose we set up an experimental rig with two calibrated and synchronized cameras to acquire front and side views of a subject. In the experiments described here we use only the frontal view, using the side view only to validate our algorithms during development. The data acquisition was done at several different times, and with several different purposes. All the images acquired and digitized were at video resolution (*i.e.,* 640 by 480). They were later cropped to a resolution of 380 by 450.

Initially subjects with experience with FACS actions were recorded. The primary volunteer in this experiment was Paul Ekman, the author of the FACS model and the leading expert on FACS actions. Another local expert was also asked to be a subject, however, only a limited amount of data from that person was used. A rig similar to the one shown in Figure 7-1 was employed for this data acquisition.

The second set of data was acquired using the set-up shown in Figure 7-1. Sixteen volunteers made expressions of *smile*, *surprise*, *fear*, *disgust*, *anger*, *raise eye brow*, and *sad*. Some of the subjects were also recorded making sounds to observe lip motions with speech. It was apparent during this phase of data acquisition that it is difficult for the subjects to make the expressions on demand, especially the expressions that are driven by emotion [1]. In order to conduct a detailed analysis of facial motion, we used this data to define average feature vectors for the expressions of *smile*, *surprise*, *disgust*, *anger* and *raise eye brow*. A majority of our subjects were unable to make realistic expressions of *sad* and *fear*, so we decided to drop these expressions from our study.

Figure 7-1: *Experimental set-up used to record subjects making expressions.*

In a third round of data acquisition, only one camera was used. The goal of this effort was to get data for comparison with the data acquired earlier, thus permitting us to test the recognition and identification capabilities of our method. This time subjects were asked to make expressions of *smile*, *surprise*, *disgust*, *anger* and *raise eyebrows*. Recordings were also made for eye blinks and winks, lip motions for vowels, and other sounds. All recognition results are based on the data from this data set. Our subjects were not able to make the all of the above expressions, in some cases they would smile (or laugh), when asked to be angry or to show surprise. The subjects would also comment that they were having a tough time making expression of disgust/contempt. This resulted in some "holes" in our database of expressions for

[1] We also invited a person with theatrical training to record expressions, so that we could get better representation of expression linked to emotions.

identification/recognition tasks.

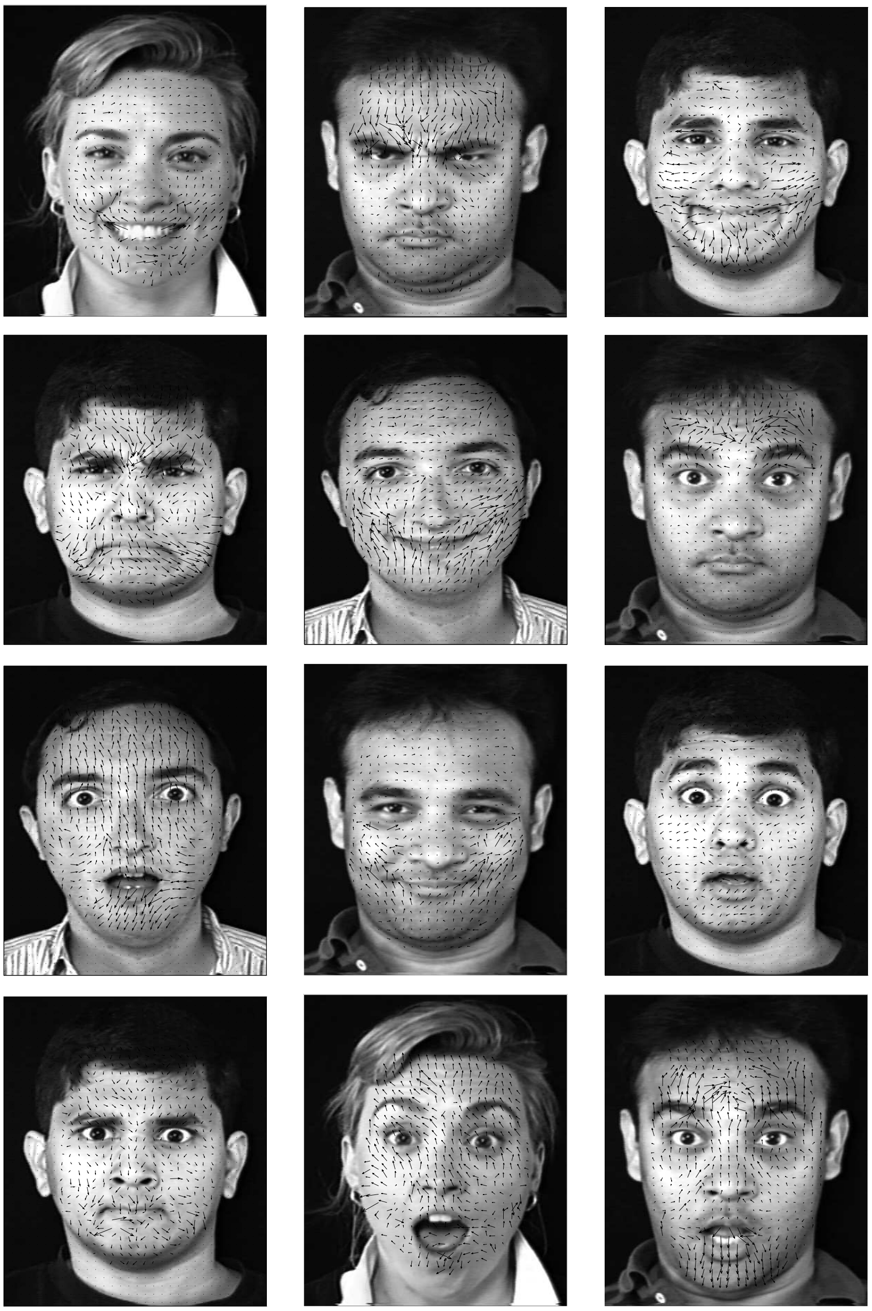


Figure 7-2: *Expressions from video sequences for various people making expression with overlaid motion fields after optical flow computation.*

7.1 The System

Flow Computations

After digitizing the acquired video sequences, optical flow was computed. Flow was computed at the full image resolution to ensure that small motions could be detected and accounted for. The coarse-to-fine algorithm used for this optical flow computation employs a *bicubic interpolation function* for unwarping the motion before extracting the motion at another level, and uses *Laplacian functions* for extracting different scales (to form a Laplacian Pyramid) of the image [82]. Our implementation of the coarse-to-fine algorithm is based on the work of Wang [91] and Simoncelli [82]. Most of our flow computation was performed by decomposing the image into 5 levels, with flow computation at each level. For some of our recognition experiments, we also analyzed the same data with 3 or 4 levels to compare the results. The covariance and confidence measures were combined from all the levels for estimation and correction of observations over time. Flow computation is the most compute-intensive part of our method and takes about an average of 60 seconds per frame for 380 by 450 images on an HP735 workstation or a DEC Alpha Workstation. Some of the results of optical flow computation are shown in Figure 7-2.

Modeling and Display

In these experiments the geometric model consisted of a 1226 node, 2512 polygon geometric model with 80 facial regions (based on [75]). This polygonal model is then used to generate a multi-layer finite element mesh as discussed in Chapter 4.

The stiffness, mass, damping and basis matrices are computed off-line and stored. The computation of the 36 muscle actuations takes about 20 seconds per frame on an SGI workstation after the optical flow has been computed. The effects of the 36 muscles on each and every vertex of the facial mesh are stored (*i.e.,* the motion on these points due to the muscle actuation) with the result that the simulation of an expression from the computed muscle actuations is in real-time. This part of the system, which runs on the SGI Iris Workstations using GL with a Motif interface, is shown in Figure 7-3. The face model, as is displayed by our system, is shown in Figure 7-4. The eyes and teeth were added separately and are not a part of the finite element mesh.

7.2 Identification and Recognition

Using the feature vectors for each expression as described earlier in Section 6.2, recognition tests were performed on a set of six subjects making about five expressions each. The plots show that the peak muscle actuations for each expression are unique, This supports the view that these are good features for recognition. For the expressions of smile, surprise, raise eyebrow, anger and disgust, the feature vectors (*i.e.,* peak muscle activations) are shown in Figure 7-5.

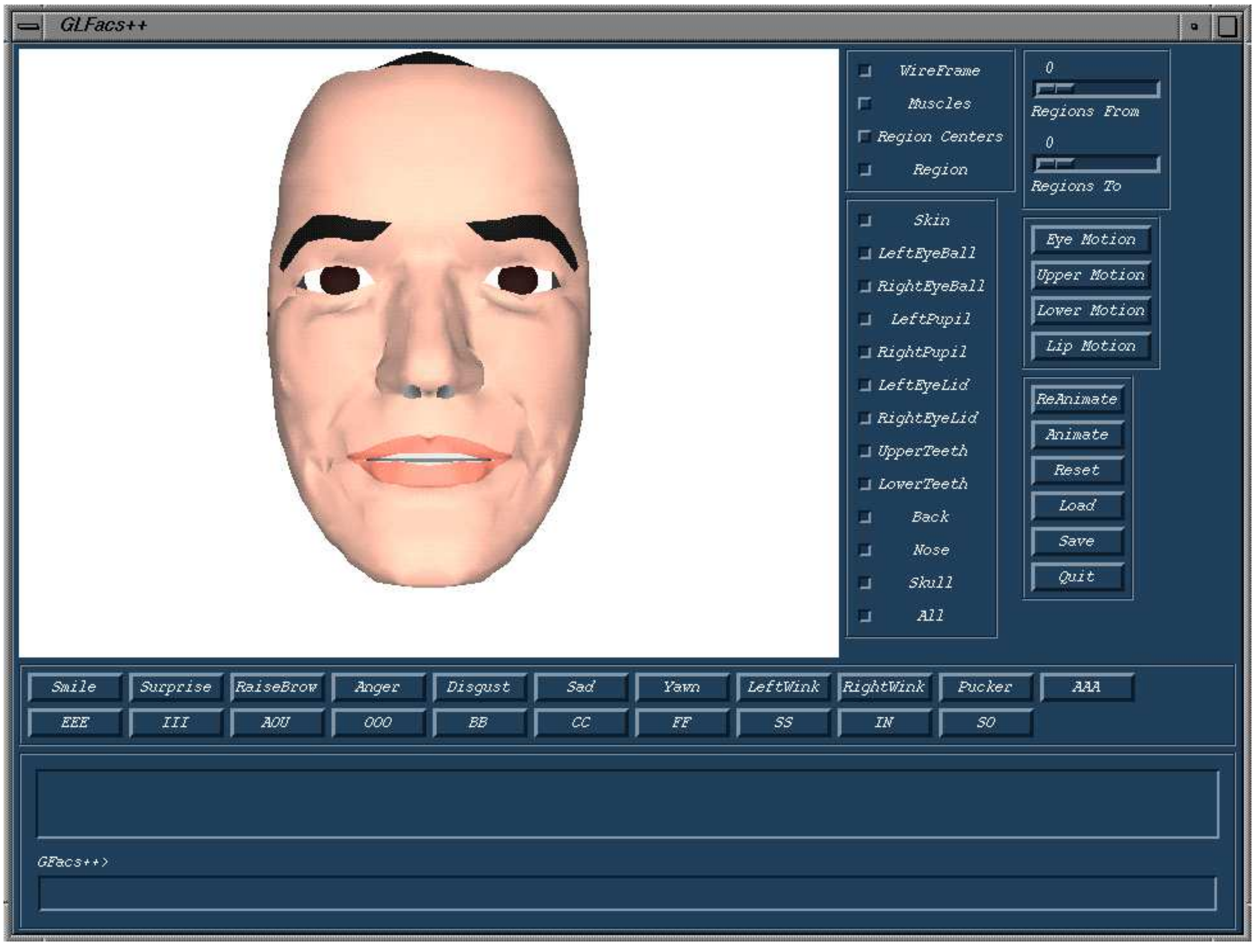
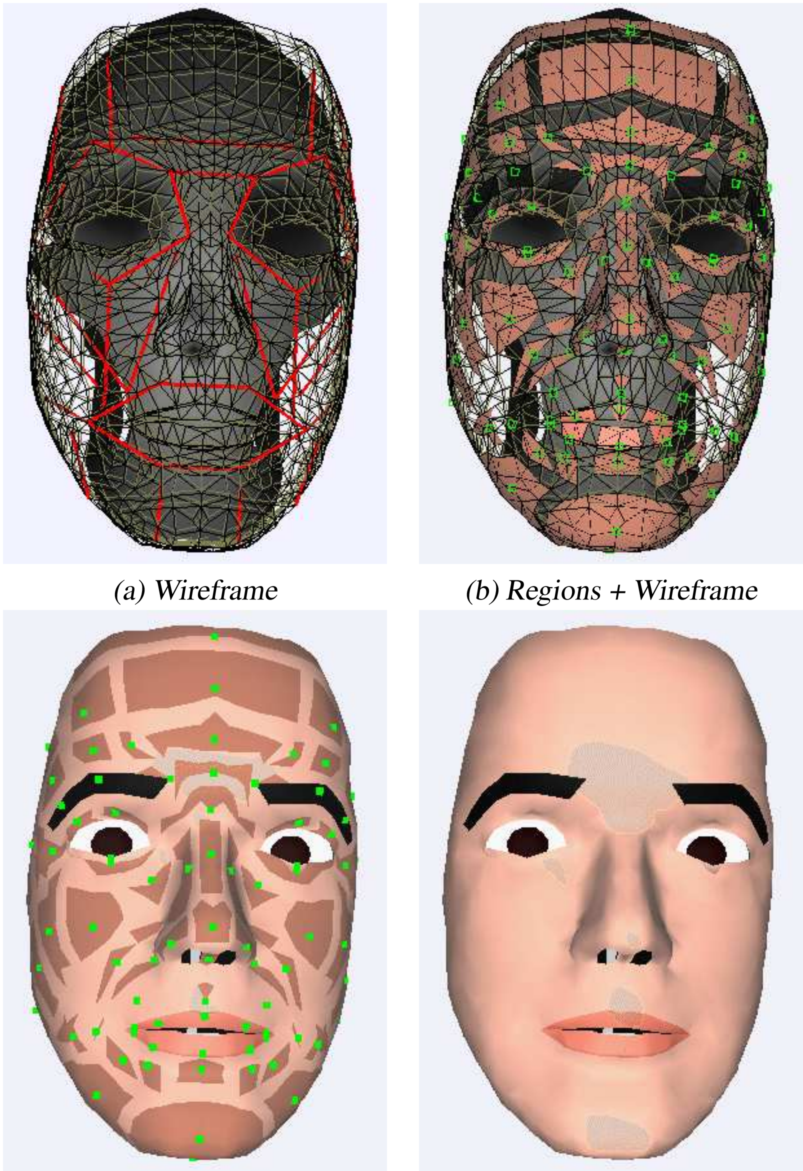


Figure 7-3: *The system running on an SGI Reality Engine II, using GL and Motif.*

Figure 7-6 shows the muscle activations for the smile expression for the six subjects. The dotted line shows the activation for the average smile expression. It can be seen that all six plots are quite similar. Similarly, Figures 7-7, 7-8, and 7-9 show the expressions of surprise, anger and disgust by different subjects. In each case it can be seen that plots are similar, demonstrating the reliability and repeatability of our system.

The major differences are due to facial asymmetry and average intensity of the actuation. The intensity difference is especially apparent in the case of the surprise expression where some people open their mouth less then others, as shown by the actuations of the first five muscles in Figure 7-7. Our analysis does not enforce any symmetry constraints and none of our data, including the averages in Figure 7-5, portray symmetric expressions.

The similarity in expression feature vectors across different individuals prompted us to attempt recognition of expressions as discussed in Section 6.2. Table 7.1 shows the results of dot products between some randomly chosen expression feature vectors with the average expressions in the database (our average



(a) Wireframe (b) Regions + Wireframe

(c) Regions + Shaded (d) Shaded Model

Figure 7-4: *Face Model used (a) shows the wireframe model with muscles, (b) show the different region and the centroid of these regions which form the coarse-grid finite element mesh, (c) and (d) show the Gouraud shaded model.*

feature vectors). It can be seen that for the five instances shown in this table, each expression is correctly identified.

Generally, the difference between the correct expression and the next most similar expression is quite large. However, in the case of anger the score for smile is also quite high. This is interesting, especially if you note the plots of smile and anger in Figure 7-5, as these plots do show muscles actuations that are similar. These are plotted together in Figure 7-10. This could mean that even though smile and anger are completely different emotions and expressions, they have similar peak muscle actuations.

Minsky [60] has suggested that the trajectory of motion is extremely important in distinguishing between anger and happiness. It is possible that our assumption of taking time and trajectory out of the picture for recognition, makes smile and anger appear similar. We will discuss this topic further when we discuss the limitation of our work and make suggestions for future work in Chapter 8.

Using this recognition method with all of our new data consisting of 6 different people making the expressions; *smile*, *surprise*, *raise eye brow*, *anger*, and *disgust*, we attempted a classification of the whole database. Since some of people did not make all expressions, we had four samples each for anger, disgust and raise eye brow and five for surprise.

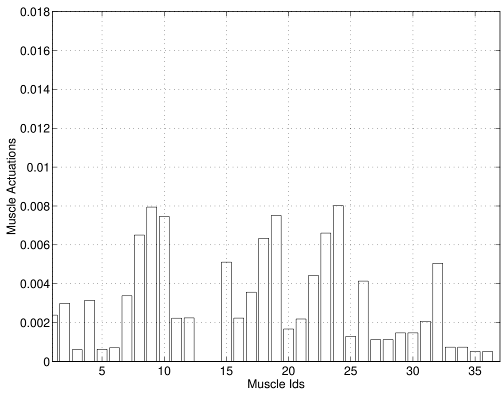
Applying our default analysis method of extracting motion using a 5 level coarse-to-fine algorithm with two-layer mesh we got a perfect recognition rate (100%) for all five expressions.

However, being aware of the limitations imposed by such a small database of expressions, we again analyzed all the expressions applying only a 3 layer coarse-to-fine algorithm. This method was unable to capture as much detail as before. Table 7.2 shows the classification results using this reduced resolution set of test cases. This table shows that by reducing the detail in the expression of anger we only got one error in recognition of the anger expression. The overall classification rate for our experiment was 97.8%, which is an excellent result, although it must be kept in mind that these results are based on an extremely small number of expressions. The similarity scores between smile and anger expressions were quite high throughout the experiment, however, as it is shown, there was only one false identification.

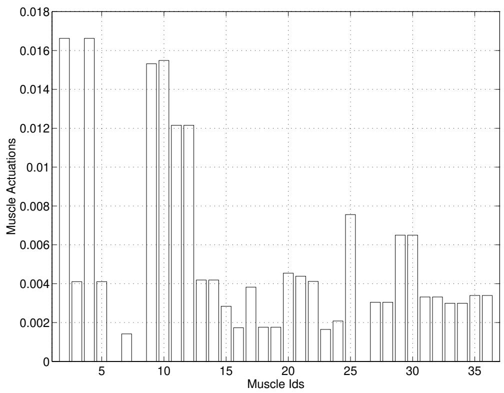
7.3 Simulations and Synthesis

Synthesis of facial expressions is a natural application of the methods developed so far. Since we utilize a facial model to extract facial motion, we have a complete representation of how to synthesize the same expression. Because an expression is reduced to a small set of muscle actuations and an actuation profile in time, the representation is extremely compact.

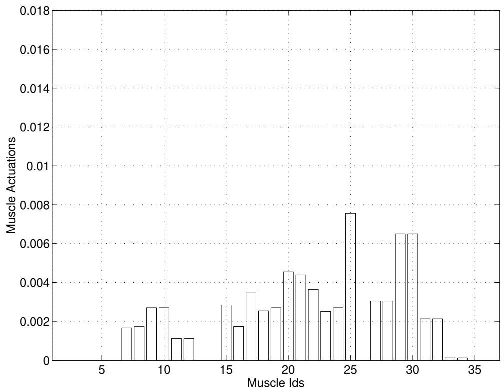
The FACS+ system being developed in this thesis would be a useful model of facial motion for the computer animation community, as it provides a parametric representation of overall facial motion that is based on motion extracted from real people, and furnishes data of how the motion evolves over time. Also, each expression is reduced to a representation that is non-specific to model geometry and the range



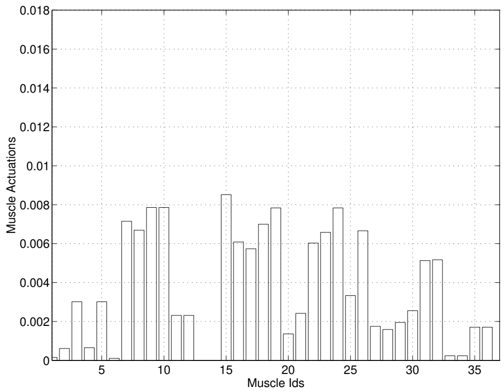
(a) Smile



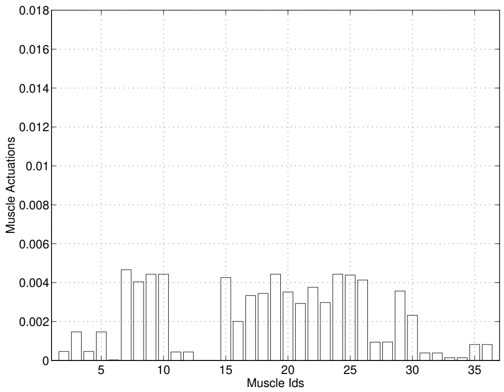
(b) Surprise



(c) Raise Eyebrows

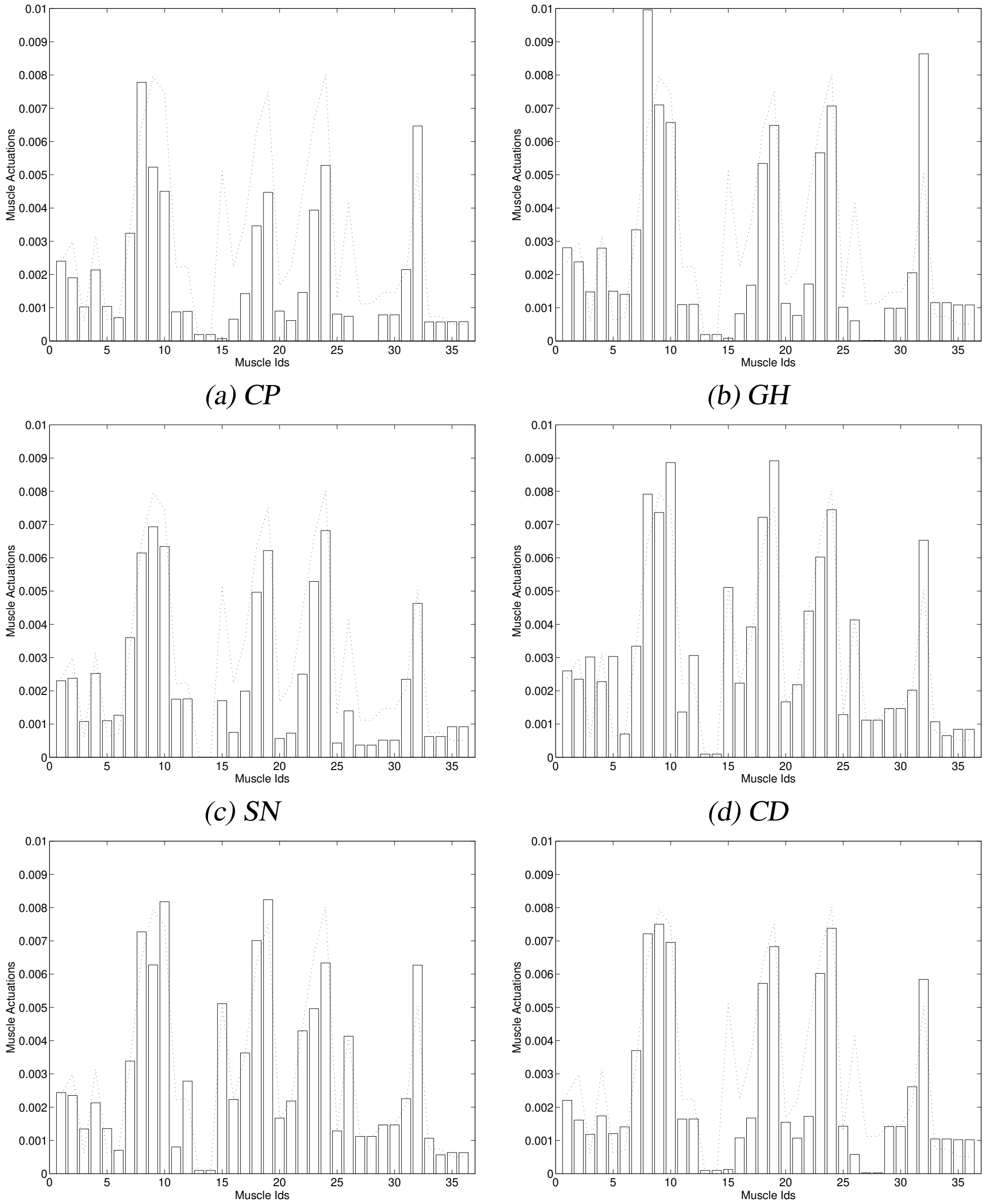


(d) Anger



(c) Disgust

Figure 7-5: *Feature vectors for different expressions*

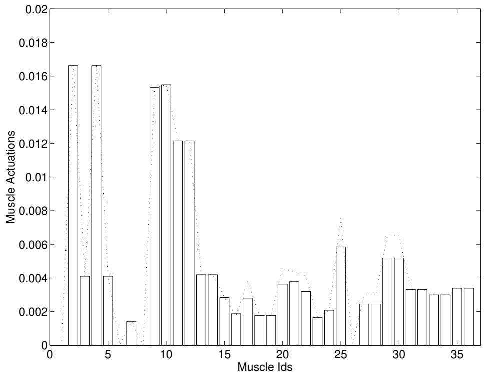


(a) CP (b) GH

(c) SN (d) CD

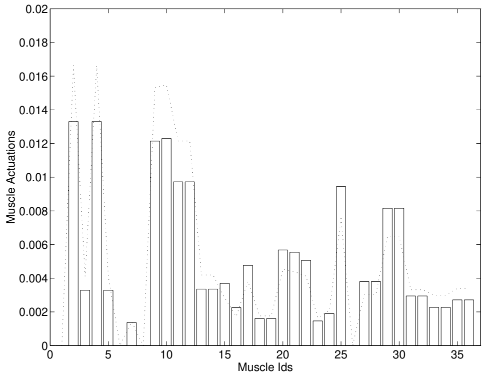
(c) IE (d) KR

Figure 7-6: *Feature vectors for smile expression for different people. The dotted line shows the average across all subjects in our experiments*

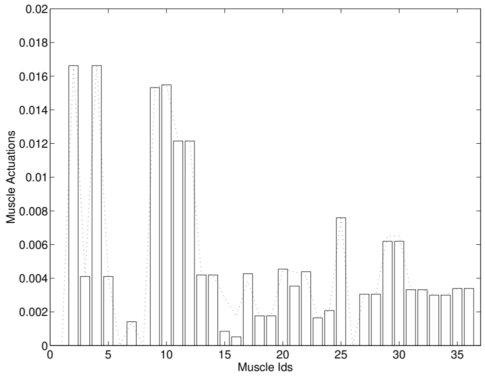


(a) CP

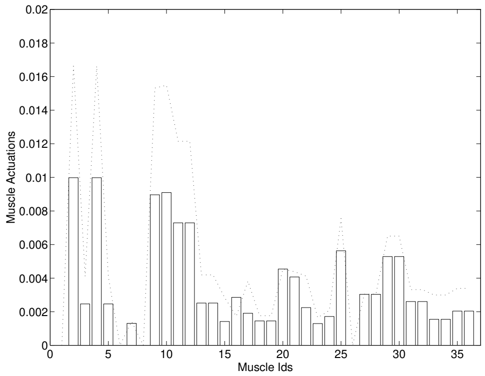
(b) GH



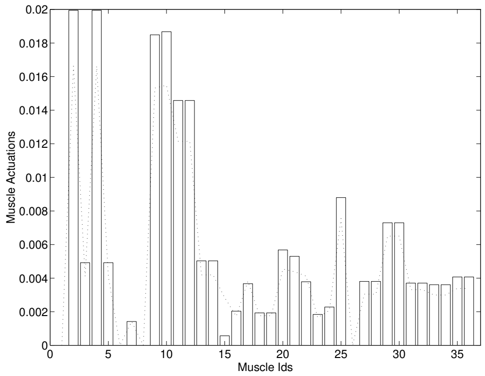
(c) SN



(d) CD

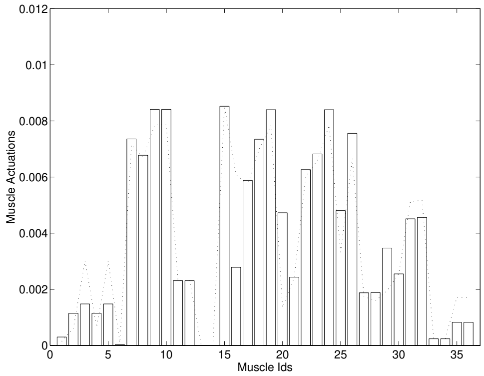


(c) IE



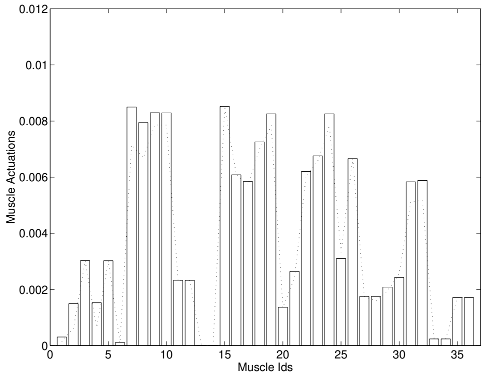
(d) KR

Figure 7-7: *Feature vectors for surprise expression for different people. The dotted line shows the average across all subjects in our experiments*

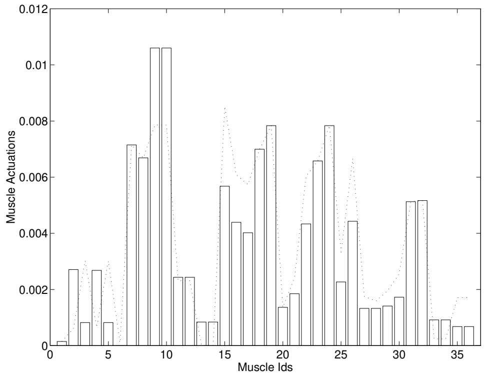


(a) CP

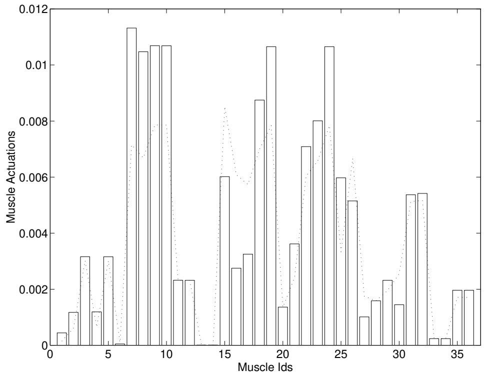
(b) GH



(c) SN



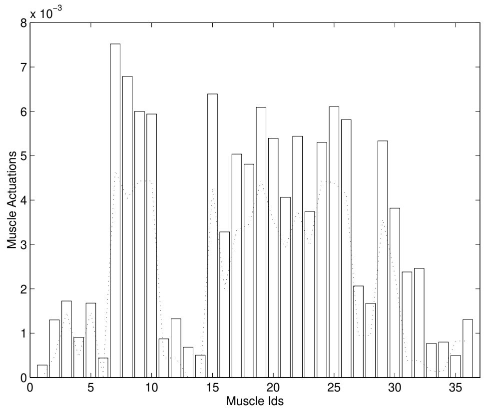
(d) CD



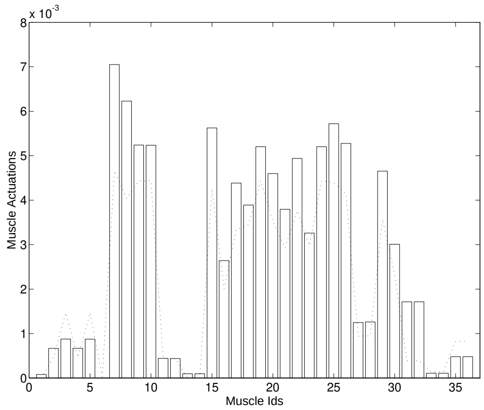
(c) IE

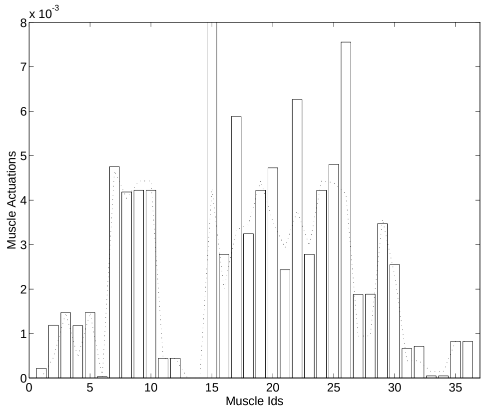
(d) KR

Figure 7-8: *Feature vectors for anger expression for different people. The dotted line shows the average across all subjects in our experiments*

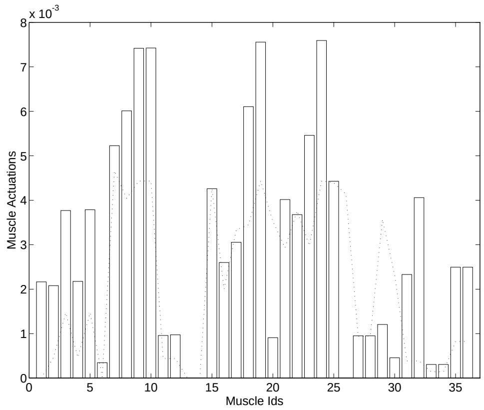


(a) CP *(b) GH*





(c) SN *(d) CD*



(c) IE *(d) KR*

Figure 7-9: *Feature vectors for disgust expression for different people. The dotted line shows the average across all subjects in our experiments*

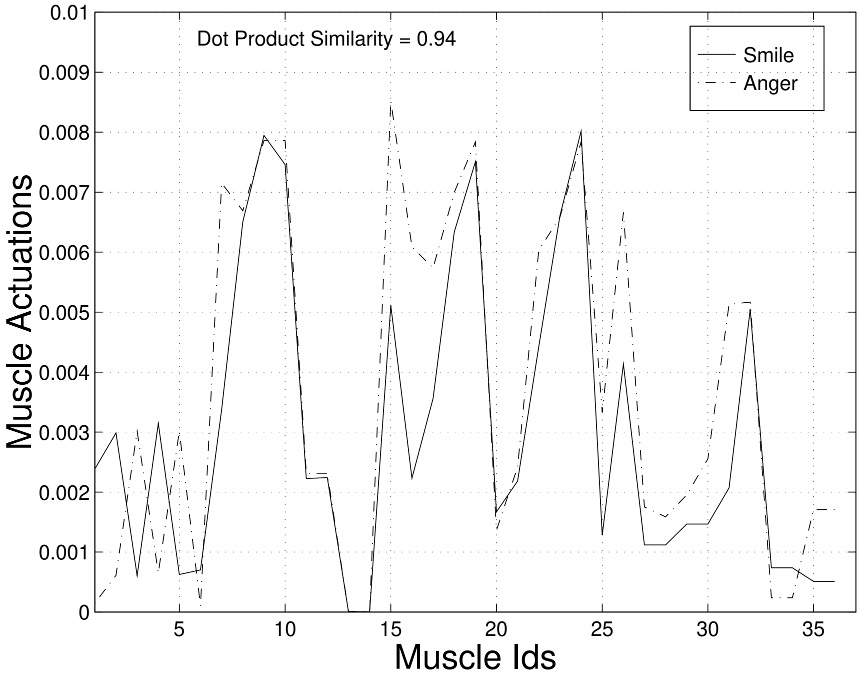


Figure 7-10: *Comparison between peak muscle actuations for Smile and Anger Expressions. The dot product of these two vectors is 0.94.*

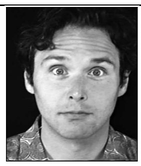
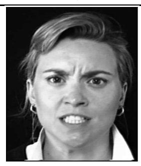
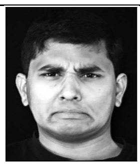
	Smile	Surprise	Raise Brow	Anger	Disgust
					
Smile	**0.91**	0.36	0.17	0.91	0.75
Surprise	0.32	**0.99**	0.20	0.34	0.28
Raise Brow	0.32	0.22	**0.88**	0.43	0.66
Anger	0.75	0.27	0.24	**0.98**	0.84
Disgust	0.62	0.28	0.47	0.81	**0.99**

Table 7.1: *Recognition of Facial Expressions*

of expressions that can be produced is constrained only by expressions that can be visually analyzed by this system.

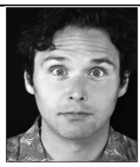
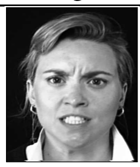
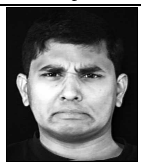
	Smile	Surprise	Raise Brow	Anger	Disgust
					
Smile	**12**	0	0	1	0
Surprise	0	**10**	0	0	0
Raise Brow	0	0	**10**	0	0
Anger	0	0	0	**7**	0
Disgust	0	0	0	0	**8**
Success	100%	100%	100%	88.8%	100%

Table 7.2: *Results of Facial Expression Recognition. This result is on based on 6 image sequences of smile, 5 image sequences of surprise and 4 each for anger disgust and raise eyebrow. The feature vector for each expression was computed using a reduced level of accuracy. Success rate for each expression is shown in the bottom row. The overall recognition rate is 97.8%.*

7.4 Real Time Tracking

Because face models have a large number of degrees of freedom, facial modeling requires dense, detailed geometric measurements in both space and time. Currently such dense measurement is both computationally expensive and noisy; consequently it is more suitable to undertake off-line analysis of discrete facial movements than to conduct real-time analysis of extended facial action. Tracking of facial expressions, in contrast, typically involves temporally sequencing between a fixed set of predefined facial actions. For instance, an extended sequence of facial expressions might consist of the lip movements associated with speech plus a few eye motions plus eyeblinks and eyebrow raises.

The number of degrees of freedom required for tracking facial articulations is limited, especially as most of the facial expressions are linear combinations of simpler motions. One can think of tracking being limited to a fixed, relatively small set of "control knobs," one for each type of motion, and then tracking the change in facial expression by moving these control knobs appropriately. The *muscle* parameters associated with these control knobs are determined by off-line modeling of each individual type of facial action as described in the previous section.

The major question, of course, is when and how much to move each control knob (*face muscle*). In this system the actuation of each muscle control parameter is determined using sparse, real-time geometric measurements from video sequences.

One way to obtain these measurements would be to locate landmarks on the face and then adjust the

Figure 7-11: *2-D Full-Face templates of neutral, smile and surprise expressions used for tracking facial expressions. See Figure 7-14 and Figure 7-15(a).*

control parameters appropriately. The difficulty with this approach is first that landmarks are difficult to locate reliably and precisely, and second that there are no good landmarks on the cheek, forehead, or eyeball.

Image Measurement

An alternative method is to *teach* the system how the person's face looks for a variety of control parameter settings, and then measure how similar the person's current appearance is to each of these known settings. From these similarity measurements we can then interpolate the correct control parameter settings. Darrell and Pentland have successfully used this general approach to describe and recognize hand gestures [21], and in our experience this method of determining descriptive parameters is much more robust and efficient than measuring landmark positions.

By constraining the space of expressions to be recognized we can match and recognize predefined expressions rather than having to derive new force controls for each new frame of video input. This can dramatically improve the speed of the system.

This thesis has so far described a method to acquire detailed muscle actuation and timing information for a set of expressions, using the optical flow method. After acquiring detailed motor controls for facial modeling, we then acquire training images of each expression for which we have obtained detailed force and timing information. This training process allows us to establish the correspondence between motor controls and image appearance.

Given a new image, we compute the peak normalized correlation score between *each* of the training views and the new data, thus producing $\mathbf{V}(t)$, vector-valued similarity measurements at each instant. Note

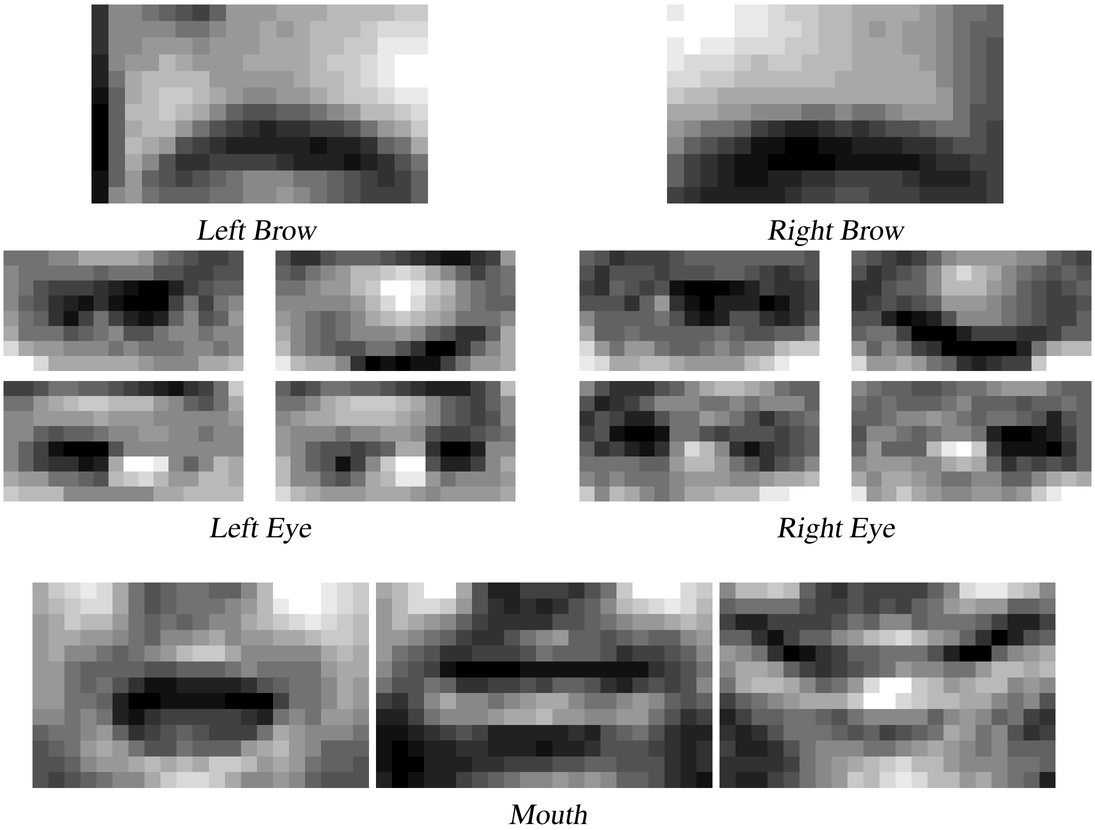


Figure 7-12: *2-D Eye-brows [Raised], Left and Right Eyes [Open, Closed, Looking Left, and Right], and Mouth templates [Open, Closed and Smiling] used for tracking facial expressions. See Figure 7-15(b).*

that the matching process can be made more efficient by limiting the search area to the neighborhood of where we last saw the eye, mouth, *etc*. Normally there is no exact match between the image and the existing expressions, so an interpolated motor observation $\mathbf{Y}(t)$ must be generated based on a weighted combination of expressions (our training examples).

In our system, we interpolate from vision scores to motor observations, using the Radial Basis Function (RBF) method [76] with linear basis functions. We define the observed motor state $\mathbf{Y}$ for a set of correlation scores $\mathbf{V}$ to be the weighted sum of the distance of the scores to a set of exemplars.

$$\mathbf{Y} = \sum_{i=1}^{n} c_i \mathcal{G}(\mathbf{V} - \mathbf{V_i}), \tag{7.1}$$

where $\mathbf{V_i}$ are the example correlation scores, $\mathcal{G}$ is an RBF (and in our case was simply a linear ramp $\mathcal{G}(\S) = ||\S||$), and the weights c_i are computed using the pseudo-inverse method given in [76]. The details of using this interpolation method for real-time expression analysis and synthesis appear in [20].

The RBF training process associates the set of view scores with the facial state, *e.g.,* the motor control parameters for the corresponding expression. If we train views using the entire face as a template, the appearance of the entire face helps determine the facial state. This provides for increased accuracy, but the generated control parameters are restricted to lie in the convex hull of the examples. View templates that correspond to parts of the face are often more robust and accurate than full-face templates, especially when several expressions are trained. This allows local changes in the face, if any, to have local effect in the interpolation.

Figure 7-12 shows the eye, brow, and mouth templates used in some of our tracking experiments, while Figure 7-11 shows full-face templates of neutral, smile and surprise expressions. (The normalized correlation calculation is carried out in real-time using commercial image processing hardware from Cognex, Inc.) The normalized correlation matching process allows the user to move freely side-to-side and up-and-down, and minimizes the effects of illumination changes. The matching is also insensitive to small changes in viewing distance ($\pm 15\%$) and small head rotations ($\pm 15°$).

Dynamic Estimation

Estimating motor controls and then driving a physical system with the inputs from a noisy video source is prone to errors, and can result in divergence or in a chaotic physical response. This is why an estimation framework needs to be incorporated to obtain stable and well-proportioned results. Similar considerations motivated the framework used in Chapter 5. Figure 7-13 shows the whole framework of estimation and control of our facial expression tracking system.

This framework uses a continuous time Kalman filter (CTKF) that allows us to estimate the uncorrupted state vector, and produces an *optimal least-squares estimate* under quite general conditions [12]. The CTKF for the above system is established by the following formulation:

$$\dot{\hat{\mathbf{X}}} = \mathbf{A}\hat{\mathbf{X}} + \mathbf{L}\left(\mathbf{Y} - \mathbf{C}\hat{\mathbf{X}}\right), \tag{7.2}$$

where $\hat{\mathbf{X}}$ is the linear least squares estimate of the state $\mathbf{X}$, which are the motor controls of facial motion. $\mathbf{A}$ is a state evolution matrix and contains elements of $\mathbf{K}$, $\mathbf{M}$ and $\mathbf{D}$ from Equation (4.1) to relate the changes in facial mesh with muscle actuation. $\mathbf{Y}$ is the observed motor state ($= \mathbf{X}$ here) for a set of correlation scores $\mathbf{V}$. Using the Riccati equation [12] to obtain the optimal error covariance matrix $\boldsymbol{\Lambda}_e$ with $\boldsymbol{\Lambda}_e$ as the error covariance matrix for $\hat{\mathbf{X}}$ and $\boldsymbol{\Lambda}_m$ the error covariance matrix for measurement $\mathbf{Y}$, the Kalman Gain matrix $\mathbf{L}$ is simply: $\mathbf{L} = \boldsymbol{\Lambda}_e \mathbf{C}^T {\boldsymbol{\Lambda}_m}^{-1}$.

The Kalman filter, Equation (7.2), mimics the noise free dynamics and corrects its estimate with a term proportional to the difference $(\mathbf{Y} - \mathbf{C}\hat{\mathbf{X}})$. This correction is between the observation and our best prediction based on previous data. Figure 7-13 shows the estimation loop (the bottom loop (a)) which is used to correct the dynamics based on the error predictions.

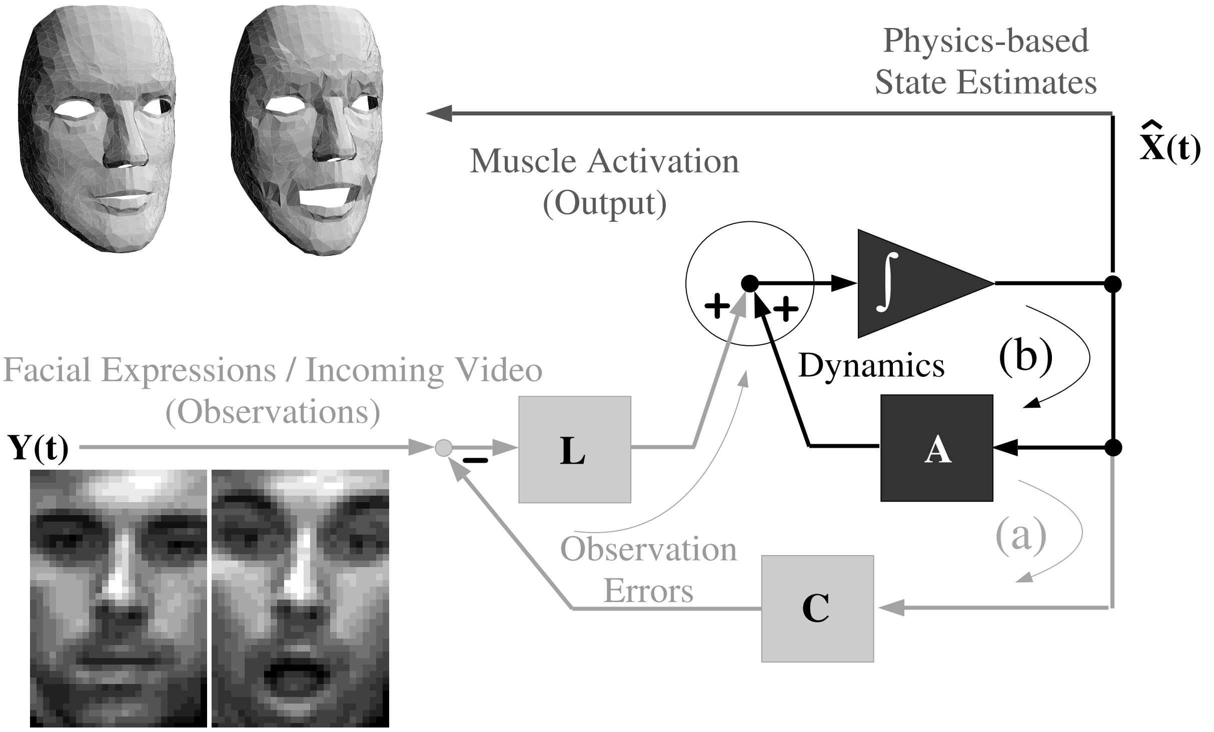


Figure 7-13: *Block diagram of the estimation and correction loop loop (a) for real-time tracking.*

Experiments

Figure 7-14 illustrates an example of real-time facial expression tracking using this system. Across the top, labeled (a), are five video images of a user making an expression. Each frame of video is then matched against all of the templates shown in Figure 7-11, and peak normalized correlation scores ($\mathbf{V}(t)$) are measured. These scores are then converted to motor observations ($\mathbf{Y}(t)$) and fed into the muscle control loop, to produce the muscle control parameters (state estimates; $\hat{\mathbf{X}}(t)$). Five images from the resulting sequence of mimicking facial expressions in 3-D are shown in (d). Figure 7-14 (b) and (c) show the correlation scores ($\mathbf{V}(t)$) and the converted motor observations ($\mathbf{Y}(t)$) respectively. This example ran in real time, with 5 frames processed per second, which is sufficient to capture general detail of expression onset and change from one expression to the other.

Figure 7-15 (c), (d) and (e) show some of the live shots of the system in use. Figure 7-15 (a) and (b) show the video feed with the regions of interest on the face for both full-face and local region templates. We have tested this system for video sequences of up to several minutes without noticeable failure in tracking. We have also tested the system successfully for tracking lip motions for speech. The major difficulty encountered is that increasing the number of templates slows down the processing and creates a lag of about half a second to a second, which is unacceptable for some applications. We are working on reducing the lag

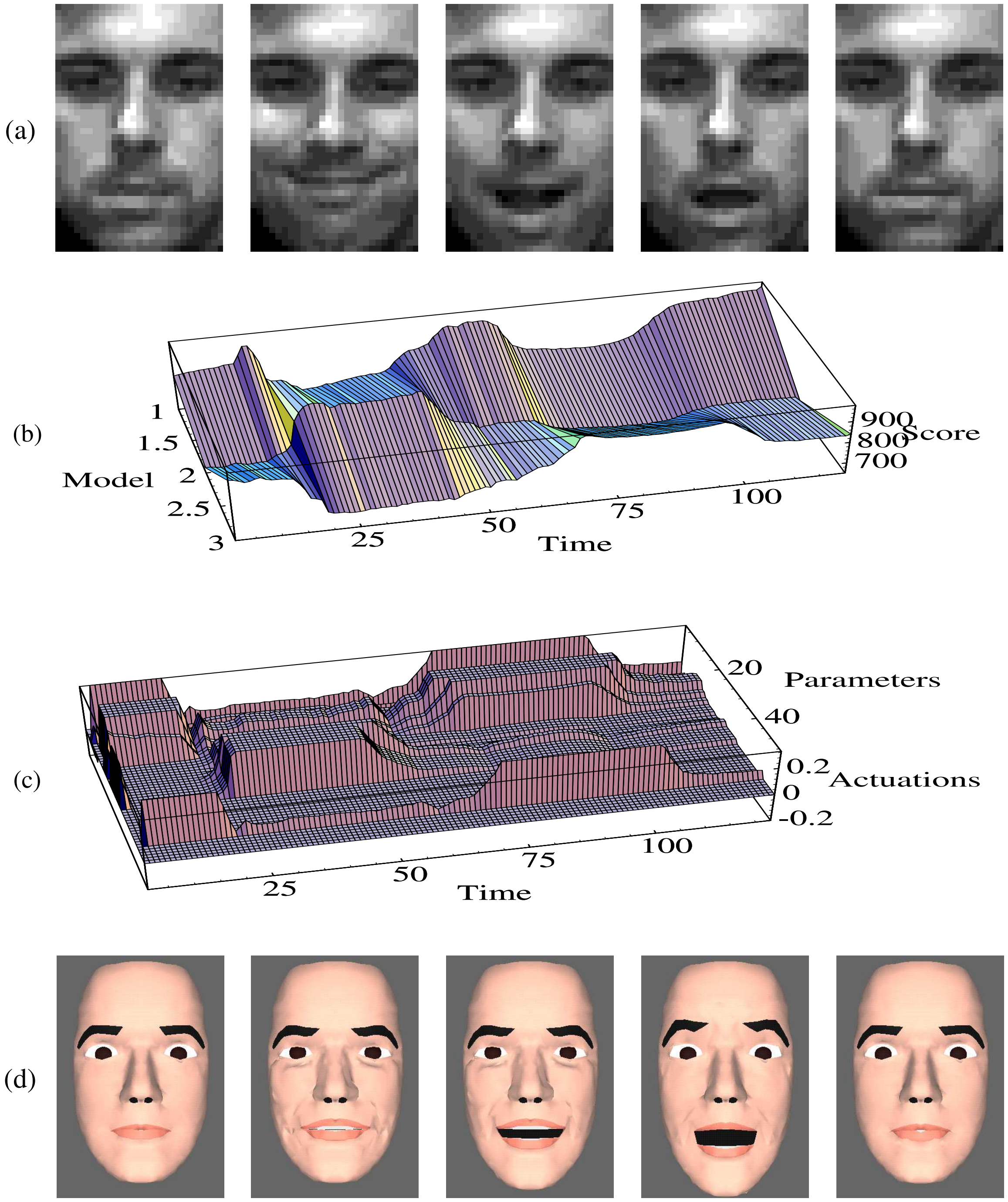
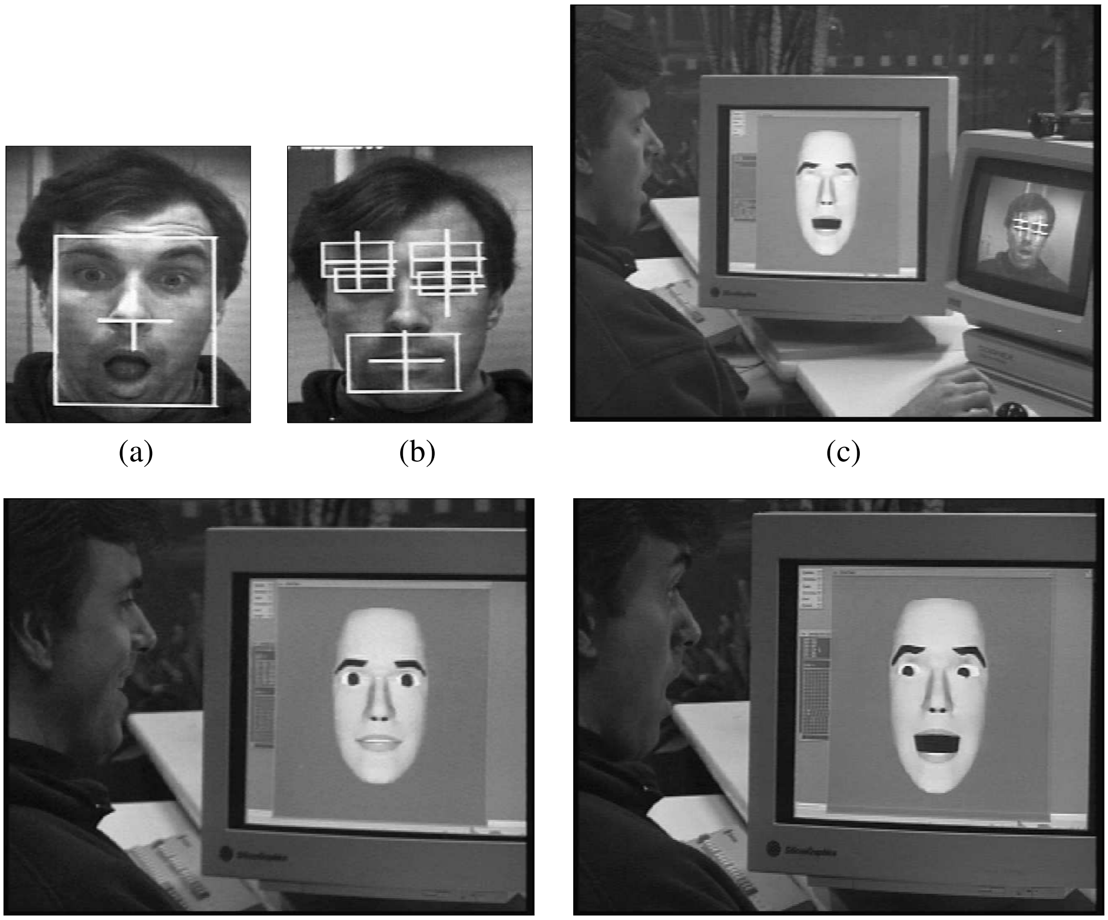


Figure 7-14: *(a) Face images used as input, (b) normalized correlation scores* $\mathbf{V}(t)$ *for each 2-D template, (c) resulting muscle control parameters* $\mathbf{X}(t)$*, (d) images from the resulting tracking of facial expressions.*



(a) (b) (c)

(d) (e)

Figure 7-15: *(a) Face with single template, (b) Face with multiple templates. (c) Complete system tracking eyes, mouth, eyebrows., (d) tracking a smile and (e) a surprise expression.*

time by incorporating a more sophisticated prediction algorithm.

In the next two chapters we will discuss the limitations of methods presented here. We also present some suggestions for future work. Following that is a brief summary of contributions of this work and the conclusions.

Chapter 8

Limitations and Future Work

> *Results! Why man, I have gotten a lot of results, I know several thousand things that won't work.*
>
> Thomas A. Edison

The experiments of the previous chapter suggest some limitations of our method. Here we discuss these limitations and evaluate their effect on results. We also make suggestions for future work.

8.1 Data Acquisition

Our experimental results rely on acquiring data of people making facial expressions on demand. We found it extremely hard to get our subjects to make expressions, even the so-called seven "universal" expressions. Expressions of happiness and surprise were easy to elicit, however, expressions of anger, disgust and fear were difficult for most subjects. A true expression for sadness was acted out by only 2 out of our 20 subjects ("trueness" of an expression was evaluated on the basis of expected facial motion as suggested by Bassili [7] and Ekman [28]). We also tried using a subject with expertise in theater to act out the various expressions. This subject did a good job of generating the correct expressions, but had a difficult time getting the timing of the expression to be close to normal (*i.e.,* spontaneous). This problem with data acquisition seriously limited our ability to characterize human emotional expressions. However, our aim is to analyze facial motion, and then check the validity of our analysis by evaluating the similarity or reproducibility of different people generating the same expressions. As our results suggest, we have done quite well at this task.

An interesting direction for future research would be to use the data acquisition method in a controlled environment as a part of a psychology experiment. Visual stimuli (*e.g.,* Paul Ekman's emotion eliciting videotapes) can be presented to the subjects and their expressions recorded and analyzed. Similar experiments could also be conducted on speech movements.

8.2 Flow Computation

Optical flow computation inherently suffers from problems like intensity singularities, brightness changes, temporal aliasing and multiple motions. Our use of a coarse-to-fine algorithm with a probabilistic framework is aimed at minimizing such problems. However, we must still expect that such problems will affect our analyses. Faces are non-Lambertian surfaces; we have observed specular reflections causing large brightness changes near the cheeks and on the forehead region during facial motion. Additionally, it is difficult for optical flow to deal with wrinkles and furrows on a face model, as optical flow usually follows the valley formed by the wrinkle/furrow.

In our analysis, we employ a facial model to measure facial motion, and we use error covariance information from the flow computation to assign a confidence level to different regions of the face model. Our multi-grid approach and an energy-based FEM mesh enforce continuity and smoothness of motion when applied to the facial model. We have also defined regions where wrinkles and furrows form, hence predicting where the flow computation can fail. However, it is still possible that motion information is incorrect because of the above-mentioned problems with image sequences. We have so far avoided these problems by controlling the lighting conditions during our data acquisition.

8.3 Initialization

As discussed in Chapter 5, we manually initialize our model onto the face image. This hand initialization precludes the possibility of automatic facial analysis. Adding to our system a method for locating faces in images and extracting their pose (orientation) can allow automatic initialization. There are several methods that are aimed at exactly this task. We believe that the feature-based deformable template work by Yuille *et al.* [100] or the pattern recognition aspects of Turk and Pentland *et al.* [87, 68] would be ideal extensions for our system.

8.4 Rigid motion and orientation

We have not concerned ourselves with rigid motion (translation and orientation) of the head during expression generation. We have constrained our study to have only frontal views of faces, and asked each of our subjects to avoid moving his/her head during an expression. We were also helped by the fact that most of the expressions take less then a second to generate, and the rigid motion in that period was small enough to estimate with our optical flow computation.

In addition to translation and rotation of the head, scaling (*i.e.,* moving towards or away from the camera) is also a major concern. Consider expressions of surprise and fear; both of them are normally accompanied by a backward motion of the head. Because we used an orthographic model of projection, we cannot deal with this type of a motion.

If facial analysis is desired in spontaneous and dynamic settings, which in some ways will alleviate our data acquisition problems, then we need to address this issue of extracting global motions of the head. There is much work aimed at extraction and analysis of body motion from video sequences [1, 53]. Thus an interesting future extension would be the addition of either a feature-based or a template-based method to extract such global motions.

Koch [47] presents a method to extract parametric description of dynamic objects and time-varying scene parameters in image sequences, especially image sequence involving heads in the teleconferencing domain. His method, which employs analysis-by-synthesis approach, on similar lines to our approach would form a very interesting precursor to our method.

The methods suggested as future extensions for initialization could be used for the purpose of extracting rigid motions too, however they would do this in a static framework; one image at a time.

8.5 Mapping from 2-D to 3-D

Another major assumption in our method was the simplified mapping of motion from 2-D images to a 3-D model. This assumption results in an unrealistic analysis of expressions with a significant "into the image" motion. Lip pucker motions suffer the most due to this limitation.

One way of resolving this problem would be to use two cameras so that depth information is available. We have used two cameras providing orthographic views to analyze some of our images. Using two cameras for a stereo setup has also been pursued extensively by vision researchers to extract 3-D information. Another method would be to use structure from motion techniques to extract the 3-D structure of the face using only one camera [58, 4].

Some of the recent work by Terzopoulos [85] addresses the issue of registration of a 3-D scanned data of a face to a model of a face automatically.

8.6 Modeling Limitations

One of the contributions of this thesis is that a face model is used to analyze facial motion within a analysis-synthesis framework. To our knowledge this is the most detailed face model to date. However, this model does not even begin to capture the complex anatomical details of the face which includes assembly of bones, cartilage, muscles, nerves, blood vessels, glands, fatty tissue, connective tissue, skin and hair. The need to model such exact detail is not apparent.

Despite its high level of detail, our model is limited to a generic facial structure with generic material properties and generic muscles and attachments. This limits the validity of our anatomically-based model. By demonstrating that we can extract similar muscle actuations for different people and reproduce similar expressions, we have shown the reasonability of our representation as a modeling tool. However, further studies are needed to validate the use of generic muscle and skin descriptions.

A possible future research direction is to acquire more anatomically accurate facial models from biomedical data. A better facial model can be developed by using MRI [1] scans of faces and using that information to model the shape of the face and location of muscles. Similarly, biomechanical studies of the material properties of the different parts of a face could lend more credibility to our FEM model of skin. EMGs could be used to validate the muscle descriptions that are extracted in our analysis,

Another important limitation of our method is the lack of "real" oris muscles. As can be seen in Figure 5-2, our face model uses simple muscles that directly connect the insertion and origin points. However, the muscles around the eyes and the lips are oris muscles, which are round and act towards and away from their center, to cause opening and closing motions. Because it is difficult to model an oris muscle, we have attempted to model these oris muscles using a series of simple muscles connected to a FEM mesh. This limitation in modeling has resulted in a somewhat unrealistic motion around the lips.

This problem can be addressed in two ways; first, by increasing the number of finite elements (*i.e.,* sampling polygons) around the lips, and second to model the oris muscle directly rather than by using discrete muscles. This is an important extension, without which it will be difficult to analyze speech motions.

Additional modeling for the tongue, teeth, eyelid, eyeball, and hair is also needed to complete the visual perception of a synthetic face.

8.7 Real-time Tracking

The real-time tracking system described in Chapter 7 also has some limitations worth mentioning. At present the system is user-specific and can handle rotations of only $\pm 15°$ and changes in viewing distance of ($\pm 15\%$).

This system can be more robust by using generic templates and not user-specific templates. Additionally, the system could also be used to recognize pose (orientation) of the head and to account for that in computing the similarity metric. The Eigenvalue approach of Pentland *et al.* [68] seems to be a good approach for solving these problems.

8.8 Emotion versus Motion

Throughout this work, we have attempted to maintain a distinction between emotion and facial motion. While it is certain that facial motion is related to emotion, the exact "rules" of this relationship are still a

[1] Magnetic Resonance Imaging

matter of study. Consequently, even though we can extract facial motion, we are unable to make serious claims about emotion recognition. Our system does, however, appear to be a good platform to study facial motion and its relationship to emotion.

This is perhaps the most interesting future research direction for the methods presented here. It is hoped that psychology researchers will use these tools in automating their analysis of facial motion, incorporating the spatial and temporal information that it provides.

8.9 Recognition of facial expressions

We have presented preliminary results that have shown the potential usefulness of our model for recognition of facial expressions. However our results are based on a very small set of experiments. Further experiments are required to confirm the high level of recognition accuracy reported here.

8.10 Modeling and control paradigm

A major contribution of our work is the method by which motor control parameters of motion are extracted from image sequences. Although we have dealt only with facial motion, the method is in no way restricted to faces. The robustness of the framework indicates that it is a promising method to apply to other problems. An interesting application would be in analysis of articulated shapes where rigid and nonrigid motions are simultaneously important and where occlusion poses a significant problem for visual observations. However, it is important to note that this method is far from real-time and computationally very expensive.

8.11 Possible Applications

In addition to the above suggestions for research, there are interesting possible applications of our work. For instance:

- We have already discussed the applicability of our system for coding images with facial expression. An interesting application would be to use this system for a teleconferencing system.
- There are also possible applications for our facial model with our extracted control parameters for expression generation, caricature and cartoon animations. The facial structure can also be used for morphing.
- Cassell *et al.* [17] has presented a system for rule-based animation of conversation between multiple agents. We can imagine combining the above two applications with Cassell's method to address the problems of conversations between a human and an animated agent.

Chapter 9

Conclusions

> .... *people will converse with a face on the computer screen. Both the computer and the person will be able to read each other's facial expressions, gleaning the understanding that can be communicated through a smile or a scowl, a nod of the head, an arched eyebrow or a piercing gaze. .... computers, now known for their cold logic will use their faces to convey emotions*
>
> A. Pollack, *Japanese Put a Human Face on Computers*, New York Times, June 28 1994, (p. C1)

Faces are much more than just keys to individual identity. Facial gestures serve both affective and communicative functions during interaction. They provide signals for the underlying emotional states and help to disambiguate speech. The automatic analysis and synthesis of facial expressions is becoming especially important in light of the increasing human-machine interaction.

Towards this end we have developed a mathematical formulation and implemented a computer vision system capable of detailed analysis of facial expressions within an active and dynamic framework. The purpose of this system to to analyze real facial motion in order to derive an improved model (FACS+) of the spatial and temporal patterns exhibited by the human face.

This system analyzes facial expressions by observing expressive articulations of a subject's face in video sequences. The visual observation (sensing) is achieved by using an *optimal optical flow* method. This motion is then coupled to a physical model describing the skin and muscle structure, and the muscle control variables estimated.

By observing the control parameters over a wide range of facial motions, we can then extract a minimal parametric representation of facial control. We can also extract a minimal parametric representation of facial patterning, a representation useful for static analysis of facial expressions.

Our experiments to date have demonstrated that we can indeed extract FACS-like models that are more accurate than existing models. We have used this representation for recognition of facial expressions over diffrent individuals and with 97.8% accuracy. We are now processing data from a wider range of facial expression in order to develop a model that is adequate for all facial expressions.

We have also developed a mathematical formulation and implemented a computer system capable of real-time tracking of facial expressions through extended video sequences.

This system analyzes facial expressions by observing expressive articulations of a subject's face in video sequences. The primary visual measurements are a set of peak normalized correlation scores using a set of previously-trained 2-D templates. These measurements are then coupled to a physical model describing the skin and muscle structure, and the muscle control variables estimated.

Our experiments to date have demonstrated that we can reliably track facial expressions, including independent tracking of eye and eyebrow movement, and the mouth movements involved in speech. We are currently extending our system so that it can handle large head rotations, and are working to remove lags in estimation/generation by use of sophisticated prediction methods.

Bibliography

[1] K. Akita. Analysis of body motion image sequences. In *Proceedings of the 6th International Conference on Pattern Recognition*, pages 320–327, October 1982.

[2] Amir Wadi Al-Khafaji and John R. Tooley. *Numerical Methods in Engineering Practice*. Holt, Rinehart and Winston Inc., 1986.

[3] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2:283–310, 1989.

[4] A. Azarbayejani, B. Horowitz, and A. Pentland. Recursive estimation of structure and motion using the relative orientation constraint. In *Computer Vision and Pattern Recognition*, 1993.

[5] M. Basseville, A. Benveniste, K. C. Chou, S. A. Golden, R Nikoukhah, and A S Willsky. Modeling and estimation of multiresolution stochastic processes. Technical Report CICS-P-283, MIT Center for Intelligent Control Systems, February 1991.

[6] J. N. Bassili. Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psyschology*, 4:373–379, 1978.

[7] J. N. Bassili. Emotion recognition: The role of facial motion and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–2059, 1979.

[8] Klaus-Jürgen Bathe. *Finite Element Procedures in Engineering Analysis*. Prentice-Hall, 1982.

[9] H. L. Bennett. F. A. C. E.: a sensitive and specific monitor for adequacy of anesthesia. In P. Sebel, G. Winograd, and B. Bonke, editors, *Memory and Awareness in Anesthesia*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[10] P. J. Benson, D. I. Perrett, and D. N. Davis. Towards a quantitative understanding of facial caricatures. In Vicki Bruce and Mike Burton, editors, *Processing Images of Faces*. Ablex Publishing Corporation, 1992.

[11] P. Bergeron and P. Lachapelle. Techniques for animating characters. In *Advanced Computer Graphics and Animation*, number 2 in SIGGRAPH 85 Tutorial Notes, pages 61–79. ACM, 1985.

[12] R. G. Brown. *Introduction to Random Signal Analysis and Kalman Filtering*. John Wiley & Sons Inc., 1983.

[13] V. Bruce. *Recognising Faces*. Lawrence Erlbaum Associates, 1988.

[14] V. Bruce. *Face Recognition*. A Special Issue of the European Journal of Cognitive Psychology. Lawrence Erlbaum Associates, 1991.

[15] R. Brunelli and T. Poggio. Face Recognition: Features versus Templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, October 1993.

[16] J. S. Bruner and R. Taguiri. The perception of people. In *Handbook of Social Psyschology*. Addison-Wesley, 1954.

[17] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, C. Seah, and M. Stone. Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple conversational agents. *ACM SIGGRAPH Conference Proceedings*, pages 413–420, 1994. Annual Conference Series.

[18] K. C. Chou, A. S. Willsky, A. Benveniste, and M. Basseville. Recursive and iterative estimation algorithms for multi-resolution stochastic processes. Technical Report LIDS-P-1857, MIT Laboratory for Information and Decision Sciences, March 1989.

[19] M. M. Cohen and D. W. Massaro. Modeling and coarticulation in synthetic visual speech. In N. Magnenat-Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*. Springer-Verlag, 1993.

[20] T. Darrell, I. A. Essa, and A. Pentland. Correlation and interpolation networks for real-time expression analysis/synthesis. In *Neural Information Processing Systems Conference*. NIPS, 1994.

[21] T. Darrell and A. Pentland. Space-time gestures. In *Computer Vision and Pattern Recognition*, 1993.

[22] C. Darwin. *The expression of the emotions in man and animals*. University of Chicago Press, 1965. (Original work published in 1872).

[23] B. deGraf. Notes on facial animation. In *SIGGRAPH 89 Tutorial Notes: State of the Art in Facial Animation*, pages 10–11. ACM, 1989.

[24] Guillaume-Benjamin Duchenne. *The Mechanism of Human Facial Expression*. Studies in Emotion and Social Interaction. Cambridge University Press ; Editions de la Maison des Sciences de l'Homme, 1990. Translation of: Mecanisme de la Physionomie Humaine.

[25] P. Ekman. Facial signs: Facts, fantasies and possibilities. In T. Sebeok, editor, *Sight, Sound and Sense*. Indiana University Press, 1978.

[26] P. Ekman. *Emotions in the Human Faces*. Studies in Emotion and Social Interaction. Cambridge University Press, second edition edition, 1982.

[27] P. Ekman. Facial expression of emotion: An old controversy and new findings. *Philosophical Transactions: Biological Sciences (Series B)*, 335(1273):63–69, 1992.

[28] P. Ekman and W. V. Friesen. *Facial Action Coding System*. Consulting Psychologists Press Inc., 577 College Avenue, Palo Alto, California 94306, 1978.

[29] P. Ekman, T. Huang, T. Sejnowski, and J. Hager (Editors). Final Report to NSF of the Planning Workshop on Facial Expression Understanding. Technical report, National Science Foundation, Human Interaction Lab., UCSF, CA 94143, 1993.

[30] A. Emmett. Digital portfolio: Tony de peltrie. *Computer Graphics World*, 8(10):72–77, October 1985.

[31] I. A. Essa, S. Sclaroff, and A. Pentland. A unified approach for physical and geometric modeling for graphics and animation. *Computer Graphics Forum, The International Journal of the Eurographics Association*, 2(3), 1992.

[32] I. A. Essa, S. Sclaroff, and A. Pentland. Physically-based modeling for graphics and vision. In Ralph Martin, editor, *Directions in Geometric Computing*. Information Geometers, U.K., 1993.

[33] Richard P. Feynman, Robert B. Leighton, and Matthew Sands. *Lectures on Physics*, volume 1. Addison-Wesley, 1977.

[34] B. Friedland. *Control System Design: An Introduction to State-Space Methods*. McGraw-Hill, 1986.

[35] N. H. Frijda. Facial expression processing. In H. D. Ellis, M. A. Jeeves, F. Newcombe, and A. Young, editors, *Aspects of Face Processing*, chapter 9, pages 319–325. Martinus Nijhoff Publishers, 1986.

[36] A. Gelb, editor. *Applied Optimal Estimation*. MIT Press, 1989.

[37] C. Ghez. Muscles: Effectors of the motor systems. In E. R. Kandel, J. H. Schwartz, and T. M. Jessell, editors, *Principles of Neural Science*, chapter 36, pages 548–563. Elsevier Science Publishing Co. Inc., third edition, 1991.

[38] S. Glenn. VActor animation system. In *ACM SIGGRAPH Visual Proceedings*, page 223, SimGraphics Engineering Corporation, 1993.

[39] B. Le Goff, T. Guiard-Marigny, M. Cohen, and C. Benoit. Real-time analysis-synthesis and intelligibilty of talking faces. In *Proceedings of the 2nd International Conference on Speech Synthesis*, 1994.

[40] A. J. Goldschen. *Continuous Automatic Speech Recognition by Lipreading*. PhD thesis, George Washington University, Washington, D.C., 1993.

[41] D. R. Hill, A. Pearce, and B. Wyill. Animating speech: An automated approach using speech synthesis by rules. *Visual Computer*, 3(5):277–289, 1988.

[42] C. E. Izard. The Maximally Discriminative Facial Movement Coding System (MAX). Technical report, Instructional Resource Center, University of Delaware, 1979.

[43] C. E. Izard. *The Psychology of Emotions*. Plenum Press, 1991.

[44] T. Kanade. *Computer recognition of human faces*. Birkhauser Verlag, 1977.

[45] E. R. Kandel, J. H. Schwartz, and T. M. Jessell (Editors). *Principles of Neural Science*. Elsevier Science Publishing Co. Inc., third edition, 1991.

[46] D. E. Kirk. *Optimal Control Theory*. Prentice-Hall, 1970.

[47] R. Koch. Dynamic 3-d scene analysis through synthesis feedback control. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):556–568, June 1993.

[48] T. Kurihara and K. Arai. A transformation method for modeling and animation of the human face from photographs. In Nadia Magnenat Thalmann and Daniel Thalmann, editors, *Computer Animation '91*, pages 45–58. Springer-Verlag, 1991.

[49] Doris Lessing. *The Four Gated City*. MacGibbon and Kee, 1969.

[50] J. P. Lewis and F. I. Parke. Automated lipsynch and speech synthesis for character animation. In *Proceedings of CHI+CG '87*, pages 143–147, 1987.

[51] H. Li, P. Roivainen, and R. Forchheimer. 3-d motion estimation in model-based facial image coding. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):545–555, June 1993.

[52] P. Litwinowitcz and L. Williams. Animating images with drawings. *ACM SIGGRAPH Conference Proceedings*, pages 409–412, 1994. Annual Conference Series.

[53] Pattie Maes. ALIVE: an artificial life interactive video environment. In *ACM SIGGRAPH Visual Proceedings*, page 189, MIT Media Laboratory, 1993.

[54] N. Magnenat-Thalmann, E. Primeau, and D. Thalmann. Abstract muscle action procedures for face animation. *The Visual Computer*, 3:290–297, 1988.

[55] D. Marr. *Vision*. W. H. Freeman and Company, 1984.

[56] K. Mase. Recognition of facial expressions for optical flow. *IEICE Transactions, Special Issue on Computer Vision and its Applications*, E 74(10), 1991.

[57] K. Mase and A. Pentland. Lipreading by optical flow. *Systems and Computers*, 22(6):67–76, 1991.

[58] L. Matthies, T. Kanade, and R. Szelsiki. Kalman filter-based algorithms for estimating depth from image sequence. *International Journal of Computer Vision*, 3(3):209–236, 1989.

[59] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):581–591, 1993.

[60] Marvin Minsky. *The Society of Mind*. A Touchstone Book, Simon and Schuster Inc., 1985.

[61] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1965.

[62] F. Parke. Parameterized modeling for facial animation. *IEEE Computer Graphics and Applications*, 2(9):61–68, 1982.

[63] F. I. Parke. Control parameterization for facial animation. In Nadia Magnenat Thalmann and Daniel Thalmann, editors, *Computer Animation '91*, pages 3–13. Springer-Verlag, 1991.

[64] F. I. Parke. Techniques of facial animation. In Nadia Magnenat Thalmann and Daniel Thalmann, editors, *New Trends in Animation and Visualization*, chapter 16, pages 229–241. John Wiley and Sons, 1991.

[65] F. I. Parke and K. Waters. *Computer Facial Animation*. AK Peters, 1994.

[66] E. C. Patterson, P. C. Litwinowitcz, and N. Greene. Facial animation by spatial mapping. In Nadia Magnenat Thalmann and Daniel Thalmann, editors, *Computer Animation '91*, pages 31–44. Springer-Verlag, 1991.

[67] C. Pelachaud, N. I. Badler, and M. Steedman. Linguistic issues in facial animation. In Nadia Magnenat Thalmann and Daniel Thalmann, editors, *Computer Animation '91*, pages 15–30. Springer-Verlag, 1991.

[68] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Computer Vision and Pattern Recognition Conference*, pages 84–91. IEEE Computer Society, 1994.

[69] A. Pentland and S. Sclaroff. Closed form solutions for physically based shape modeling and recovery. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(7):715–729, July 1991.

[70] E. Petajan. Automatic lipreading to enhance speech recognition. In *Computer Vision and Pattern Recognition Conference*. IEEE Computer Society, 1985.

[71] S. Pieper. *CAPS: Computer-Aided Plastic Surgery*. PhD thesis, The Media Laboratory, Massachusetts Institute of Technology, 1991.

[72] S. Pieper, J. Rosen, and D. Zeltzer. Interactive graphics for plastic surgery: A task level analysis and implementation. *Computer Graphics, Special Issue: ACM Siggraph, 1992 Symposium on Interactive 3D Graphics*, pages 127–134, 1992.

[73] I. Pilowsky, M. Thornton, and B. B. Stokes. Towards the quantification of facial expression with the use of a mathematical model of the face. In H. D. Ellis, M. A. Jeeves, F. Newcombe, and A. Young, editors, *Aspects of Face Processing*, chapter 9, pages 340–349. Martinus Nijhoff Publishers, 1986.

[74] S. Platt. *A Structural Model of the Human Face*. PhD thesis, University of Pennsylvania, Department of Computer and Information Science, Philadelphia, PA 19104, 1986. TR #: MS-CIS-86-11.

[75] S. M. Platt and N. I. Badler. Animating facial expression. *ACM SIGGRAPH Conference Proceedings*, 15(3):245–252, 1981.

[76] T. Poggio and F. Girosi. A theory of networks for approximation and learning. Technical Report A.I. Memo No. 1140, Artificial Intelligence Lab, MIT, Cambridge, MA, July 1989.

[77] Lyn Quam. Hierarchical warp stereo. In *Proceedings of the DARPA Image Understanding Workshop*, September 1984.

[78] M. J. T. Reinders, F. A. Odijk, J. C. A. van der Lubbe, and J. J. Gerbrands. Tracking of global motion and facial expressions of a human face in image sequences. In *Visual Communication and Image Processing*, pages 1516–1527. SPIE, 1993.

[79] M. Rydfalk. *CANDIDE: A Parameterized Face*. PhD thesis, Linköping University, Department of Electrical Engineering, Oct 1987.

[80] S. Sclaroff and A. Pentland. Generalized implicit functions for computer graphics. *ACM SIGGRAPH Conference Proceedings*, 25(4):247–250, 1991.

[81] Larry J. Segerlind. *Applied Finite Element Analysis*. John Wiley and Sons, 1984.

[82] E. P. Simoncelli. *Distributed Representation and Analysis of Visual Motion*. PhD thesis, Massachusetts Institute of Technology, 1993.

[83] Richard Szeliski. Bayesian modeling of uncertainty in low-level vision. *International Journal of Computer Vision*, 5(3):271–301, December 1990.

[84] A. Takeuchi and K. Nagao. Speech dialogue with facial displays. In *ACM CHI '94 Conference Proceedings*, pages 449–450, 1994.

[85] D. Terzopoulus and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):569–579, June 1993.

[86] M. Turk. *Interactive-Time Vision: Face Recognition as a Visual Behavior*. PhD thesis, M.I.T. Media Laboratory, 1991. Also available as MIT Perceptual Computing Group TR No. 183.

[87] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[88] M. Viaud and H. Yahia. Facial animation with muscle and wrinkle simulation. In *IMAGECON 1993: Second International Conference on Image Communication*, pages 117–121, 1993.

[89] C. T. Waite. The facial action control editor for computer generated animation. Master's thesis, M. I. T. Media Laboratory, 1989.

[90] E. F. Walther. *Lipreading*. Nelson-Hall Inc., 1982.

[91] J. Y. A. Wang and E. Adelson. Layered representation for motion analysis. In *Computer Vision and Pattern Recognition*, 1993.

[92] K. Waters. A muscle model for animating three-dimensional facial expression. *ACM SIGGRAPH Conference Proceedings*, 21(4):17–23, 1987.

[93] K. Waters. A physical model of facial tissue and muscle articulation derived from computer tomography data. *SPIE, Visualization in Biomedical Computing*, 1808:574–583, 1992.

[94] K. Waters and D. Terzopoulos. Modeling and animating faces using scanned data. *The Journal of Visualization and Computer Animation*, 2:123–128, 1991.

[95] K. Waters and D. Terzopoulos. The computer synthesis of expressive faces. *Phil. Transacations, Royal Society, London*, 1992.

[96] K. Wayers and T. M. Levergood. DECface: An automatic lip-synchronization algorithm for synthetic faces. Technical Report CRL 93/4, DEC Cambridge Research Lab, 1993.

[97] B. Welsh. *Model-based Coding of Images*. PhD thesis, British Telecom Research Labs, 1991.

[98] L. Williams. Performance-driven facial animation. *ACM SIGGRAPH Conference Proceedings*, 24(4):235–242, 1990.

[99] Y. Yacoob and L. Davis. Computing spatio-temporal representations of human faces. In *Computer Vision and Pattern Recognition Conference*, pages 70–75. IEEE Computer Society, 1994.

[100] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111, 1992.

Appendix A

Facial Structure

The model used to describe the facial model was provided to us by the computer graphics group at University of Pennsylvania. After some of out modifications, this model now has 1223 points, of which 136 are fixed to distinguish between rigid and non-rigid motion. There are 2321 polygons (triangles), each of the is described as a isoparametric finite element. The face is divided into 80 regions that are described below:

Region Number	Region Name
0	ABOVE-FOREHEAD-CENTRAL
1	ABOVE-FOREHEAD-LATERAL-LEFT
2	ABOVE-FOREHEAD-LATERAL-RIGHT
3	ABOVE-UPPER-LIP-CENTRAL
4	ABOVE-UPPER-LIP-LATERAL-LEFT
5	ABOVE-UPPER-LIP-LATERAL-RIGHT
6	BETWEEN-ABOVE-BROW
7	BETWEEN-BROW
8	CHIN-BOSS
9	FOREHEAD-CENTRAL
10	FOREHEAD-MEDIAL-LEFT
11	FOREHEAD-MEDIAL-RIGHT
12	LEFT-ABOVE-BROW-CENTRAL
13	LEFT-ABOVE-BROW-LATERAL
14	LEFT-ABOVE-BROW-MEDIAL
15	LEFT-BELOW-LOWER-LID
16	LEFT-BEYOND-EYE-CORNER
17	LEFT-BEYOND-LIP-CORNER
18	LEFT-BROW-CENTRAL
19	LEFT-BROW-LATERAL
20	LEFT-BROW-MEDIAL
21	LEFT-EYE-COVER-FOLD-CENTRAL
22	LEFT-EYE-COVER-FOLD-LATERAL
23	LEFT-INFRAORBITAL-TRIANGLE-ABOVE-NASOLABIAL-FURROW
24	LEFT-INFRAORBITAL-TRIANGLE-CENTRAL
25	LEFT-INFRAORBITAL-TRIANGLE-LATERAL

26	LEFT-INFRAORBITAL-TRIANGLE-MEDIAL
27	LEFT-INSIDE-NASOLABIAL-FURROW
28	LEFT-LIP-CORNER
29	LEFT-LOWER-LID
30	LEFT-NOSTRIL-WING
31	LEFT-OUTER-CHEEK-LOWER
32	LEFT-TEMPLE
33	LEFT-OUTER-CHEEK-UPPER
34	LEFT-SIDE-OF-CHIN
35	LEFT-SIDE-OF-NOSE
36	LOWER-LIP-CENTRAL
37	LOWER-LIP-LATERAL-LEFT
38	LOWER-LIP-LATERAL-RIGHT
39	LOWER-LIP-MEDIAL-LEFT
40	LOWER-LIP-MEDIAL-RIGHT
41	NOSE-RIDGE
42	NOSE-STRUT
43	NOSE-TIP
44	RIGHT-ABOVE-BROW-CENTRAL
45	RIGHT-ABOVE-BROW-LATERAL
46	RIGHT-ABOVE-BROW-MEDIAL
47	RIGHT-BELOW-LOWER-LID
48	RIGHT-BEYOND-EYE-CORNER
49	RIGHT-BEYOND-LIP-CORNER
50	RIGHT-BROW-CENTRAL
51	RIGHT-BROW-LATERAL
52	RIGHT-BROW-MEDIAL
53	RIGHT-EYE-COVER-FOLD-CENTRAL
54	RIGHT-EYE-COVER-FOLD-LATERAL
55	RIGHT-INFRAORBITAL-TRIANGLE-ABOVE-NASOLABIAL-FURROW
56	RIGHT-INFRAORBITAL-TRIANGLE-CENTRAL
57	RIGHT-INFRAORBITAL-TRIANGLE-LATERAL
58	RIGHT-INFRAORBITAL-TRIANGLE-MEDIAL
59	RIGHT-INSIDE-NASOLABIAL-FURROW
60	RIGHT-LIP-CORNER
61	RIGHT-LOWER-LID
62	RIGHT-NOSTRIL-WING
63	RIGHT-TEMPLE
64	RIGHT-OUTER-CHEEK-LOWER
65	RIGHT-OUTER-CHEEK-UPPER
66	RIGHT-SIDE-OF-CHIN
67	RIGHT-SIDE-OF-NOSE
68	ROOT-OF-NOSE
69	UNDER-CHIN
70	UNDER-LOWER-LIP-CENTRAL
71	UNDER-LOWER-LIP-LATERAL-LEFT
72	UNDER-LOWER-LIP-LATERAL-RIGHT
73	UNDER-LOWER-LIP-MEDIAL-LEFT
74	UNDER-LOWER-LIP-MEDIAL-RIGHT
75	UPPER-LIP-CENTRAL
76	UPPER-LIP-LATERAL-LEFT

77	UPPER-LIP-LATERAL-RIGHT
78	UPPER-LIP-MEDIAL-LEFT
79	UPPER-LIP-MEDIAL-RIGHT