Automatic Real-Time Facial Expression Recognition for Signed Language Translation

Jacob Richard Whitehill

A thesis submitted in partial fulfillment of the requirements for the degree of Magister Scientiae in the Department of Computer Science, University of the Western Cape.

May 2006

Keywords

Machine learning

Facial expression recognition

Sign language

Facial action units

Segmentation

Support vector machines

Boosting

Adaboost

Haar

Gabor

Abstract

Automatic Real-Time Facial Expression Recognition for Signed Language Translation

Jacob Richard Whitehill
M.Sc. thesis, Department of Computer Science, University of the Western Cape

We investigated two computer vision techniques designed to increase both the recognition accuracy and computational efficiency of automatic facial expression recognition. In particular, we compared a local segmentation of the face around the mouth, eyes, and brows to a global segmentation of the whole face. Our results indicated that, surprisingly, classifying features from the whole face yields greater accuracy despite the additional noise that the global data may contain. We attribute this in part to correlation effects within the Cohn-Kanade database. We also developed a system for detecting FACS action units based on Haar features and the Adaboost boosting algorithm. This method achieves equally high recognition accuracy for certain AUs but operates two orders of magnitude more quickly than the Gabor+SVM approach. Finally, we developed a software prototype of a real-time, automatic signed language recognition system using FACS as an intermediary framework.

22 May 2006

Declaration

I declare that *Automatic Real-time Facial Expression Recognition for Signed Language Translation* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Jacob Whitehill 22 May 2006

Signed:

Foreword and Acknowledgment

Conducting this research at the University of the Western Cape (UWC) was a challenging and demanding experience, especially because of the limited material resources that UWC possesses and the small research staff that it hosts. It was exactly through overcoming these challenges, however, that I matured as an aspiring scientist while writing my MSc thesis. As my adviser so often reminds his students, this is *my* thesis, and any problems that arose during its completion were mine alone to solve. Learning to convert my moments of confusion into well-posed questions, and learning where to begin searching for answers to these questions, are lessons even more valuable than the considerable knowledge of automatic facial expression recognition I have amassed.

During this learning process I was aided by several people whom I would like to thank. First, Mr. David Petro of the Bastion Center for the Deaf in Cape Town generously volunteered his time and native knowledge of South African Sign Language. Without his help, the pilot study on SASL recognition in this thesis would not have been possible. The three examiners of this thesis provided useful feedback on improving the thesis presentation as well as several useful references on support vector machines (SVMs). Mr. Steve Kroon from the University of Stellenbosch kindly answered numerous questions on SVMs and statistics. Professor Marian Stewart Bartlett of the Machine Perception Laboratory (MPLab) at the University of California at San Diego gave me detailed and insightful feedback on my analysis of local versus global face analysis. To Dr. Gwen Littlewort, also of the MPLab, I express my particular gratitude for her generous, patient, encouraging, and helpful responses to my many email queries about Gabor filters, Adaboost, and FACS AU recognition. Finally, I thank my research adviser, Professor Christian W. Omlin, now at the University of the South Pacific in Fiji, for his faith in me as a researcher, his encouragement at times of frustration, his enthusiasm, and his high-level wisdom on this challenging research project.

This research was partially funded by the Telkom/Cisco Centre for Excellence for IP and Internet Computing at the University of the Western Cape.

Contents

1 Introduction 2

1.1 Thesis Objectives . . . 3

1.2 Outline . . . 3

2 Facial Action Coding System 5

2.1 Purpose of FACS . . . 5

2.2 The Design of FACS . . . 6

2.2.1 AU Combinations . . . 6

2.2.2 AU Intensity . . . 7

2.3 Suitability of FACS for Sign Language Recognition . . . 7

2.4 Alternative Systems for Facial Expression Description . . . 7

2.5 Why Use FACS for SASL? . . . 8

2.6 Summary . . . 8

3 Literature Review 9

3.1 Comparing the Accuracy of FER Systems . . . 9

3.2 Local versus Global Segmentation . . . 10

3.3 Feature Extraction for FER: The Two Approaches . . . 11

3.4 Geometry-based Features . . . 11

3.4.1 Locations and Relative Distances . . . 12

3.4.2 Parameter Estimation . . . 13

3.4.3 Models of Face Musculature . . . 14

3.4.4 Dimensionality Reduction . . . 14

3.5 Appearance-based Features . . . 15

3.5.1 Optical Flow . . . 15

3.5.2 Pixel Intensity Values . . . 16

3.5.3 Dimensionality Reduction in Appearance-Based Systems . . . 17

3.5.4 Gabor Filters . . . 18
3.5.5 Haar Wavelets . . . 21
3.6 Comparing the Two Approaches . . . 24
3.7 Combining Geometric and Appearance-based Features . . . 25
3.8 Conclusions . . . 26
3.9 Summary . . . 26

4 Support Vector Machines **27**
4.1 Premise . . . 27
4.2 Training Phase . . . 28
4.2.1 The Lagrangian Method and the Wolfe Dual Form . . . 29
4.2.2 Determining b . . . 31
4.3 Test Phase . . . 31
4.4 Linear Inseparability . . . 32
4.5 Non-linear Decision Surfaces . . . 33
4.5.1 Kernel Functions and Mercer's Condition . . . 35
4.6 Polychotomous Classification . . . 35
4.7 Summary . . . 36

5 Experimental Results **37**
5.1 Preliminary Parameters and Techniques . . . 37
5.1.1 Facial Expression Database . . . 37
5.1.2 Image Normalization . . . 38
5.1.3 AU Classification . . . 38
5.1.4 Metric of Accuracy . . . 38
5.1.5 Cross Validation . . . 39
5.2 Local versus Global Face Segmentation . . . 39
5.2.1 Feature Extraction . . . 39
5.2.2 Segmentations . . . 39
5.2.3 Results . . . 40
5.2.4 Discussion . . . 40
5.3 Haar Features and Adaboost for AU Recognition . . . 43
5.3.1 Feature Selection . . . 43
5.3.2 Face Region Segmentation . . . 44
5.3.3 Feature Extraction . . . 44

5.3.4 Classification . . . 44
5.3.5 Results . . . 45
5.3.6 Theoretical Performance Analysis . . . 45
5.3.7 Empirical Performance Analysis . . . 47
5.4 Summary . . . 48

6 Real-Time SASL Video Analysis **49**
6.1 Uses of Facial Expressions in Signed Languages . . . 49
6.1.1 Lexical Functionality . . . 50
6.1.2 Adverbial Functionality . . . 50
6.1.3 Syntactic Functionality . . . 50
6.2 Expression Intensity . . . 51
6.3 Implications for Automatic Translation . . . 52
6.4 Recognizing Facial Expressions of SASL . . . 52
6.4.1 Test Case: A Simple Story . . . 54
6.5 Approach . . . 55
6.5.1 Method 1: Exact Matching . . . 57
6.5.2 Method 2: Cosine Similarity . . . 57
6.6 System Design . . . 57
6.7 Experiment . . . 58
6.8 Results . . . 59
6.9 Discussion . . . 59
6.10 Summary and Conclusions . . . 62

7 Conclusions and Directions for Further Research **64**
7.0.1 Facial Expression Recognition . . . 65
7.0.2 Automatic Signed Language Recognition . . . 65

A Mathematical Fundamentals and Computer Vision Algorithms **66**
A.1 Distance between a hyperplane H and the origin . . . 66
A.2 Time Complexity of 2-D FFT . . . 66
A.3 Principle Component Analysis . . . 67
A.4 Optic Flow Analysis . . . 68
A.5 Haar Wavelets . . . 69
A.5.1 One-dimensional Haar Wavelet Decomposition . . . 69
A.5.2 Two-dimensional Haar Wavelet Decomposition . . . 70

B Representative ROC Curves 71

B.1 Local Gabor+SVM 71

B.2 Global Gabor+SVM 73

B.3 Local Haar+Adaboost 75

Chapter 1

Introduction

In human-to-human dialogue, the articulation and perception of facial expressions form a communication channel that is supplementary to voice and that carries crucial information about the mental, emotional, and even physical states of the conversation partners. In their simplest form, facial expressions can indicate whether a person is happy or angry. More subtly, expressions can provide either conscious or subconscious feedback from listener to speaker to indicate understanding of, empathy for, or even skepticism toward what the speaker is saying. Recent research has shown that certain facial expressions may also reveal whether an interrogated subject is attempting to deceive her interviewer [Ekm01].

One of the lesser known uses of facial expression in human interaction is signed communication, i.e., "sign language." In signed languages, facial expressions are used to denote the basic emotions such as "happy" and "sad". Even more importantly, however, they also provide lexical, adverbial, and syntactic information. In some instances, a signer may use a facial expression to strengthen or emphasize an adverb which is also gestured through the hands. In others, the facial expression may serve to differentiate two nouns from each other. Any computer system designed to recognize a signed language must thus be able to recognize the facial expressions both accurately and efficiently.

Throughout the world, but especially in developing countries such as South Africa, deaf people face severely limited educational and occupational opportunities relative to a hearing person. The existence of a computer system that could automatically translate from a signed language to a spoken language and vice-versa would be of great benefit to the deaf community and could help to alleviate this inequality. In the South African Sign Language Project at the University at the Western Cape, of which this research is a part, we envision the development of a small, unobtrusive, hand-held computing device that will facilitate the translation between signed and spoken languages. This computer system will need to recognize both hand gestures and facial expressions simultaneously; it must then analyze these two channels linguistically to determine the intended meaning; and it will need to output the same content in the target language.

All three stages must operate in real-time. In this thesis are interested in the facial expression recognition aspects of this translation device. We believe that the Facial Action Coding System (FACS, by Ekman and Friesen[EF78]), a well-known framework which objectively describes human facial expressions in terms of facial "action units", will serve as a useful intermediary representation for SASL expression recognition. In the section below, we describe our particular thesis goals.

1.1 Thesis Objectives

The goals of this thesis are two-fold:

- First, we wish to construct an automatic FACS action unit recognition system that supports the automated recognition and translation of South African Sign Language (SASL). Automatic FACS action unit recognition is useful in its own right and has numerous applications in psychological research and human-computer interaction.
- Second, using the action unit recognition system that we build, we will construct a software prototype for the recognition of facial expressions that occur frequently in SASL and evaluate this prototype on real SASL video.

Automatic facial expression recognition (FER) takes place during three phases: (1) image preprocessing, face localization and segmentation; (2) feature extraction; and (3) expression classification. This thesis investigates techniques across all three stages with the goal of increasing both accuracy and speed. In our first main experiment, we investigate the effect of local segmentation around facial features (e.g., mouth, eyes, and brows) on recognition accuracy. In our second experiment, we assess the suitability of using Haar features combined with the Adaboost boosting algorithm for FACS action unit recognition. We conduct both experiments using the Cohn-Kanade database [KCIT00] as our dataset, and using the area under the Receiver Operator Characteristics (ROC) curve, also know as the A' statistic, as the metric of accuracy. For statistical significance, we use matched-pairs, two-tailed t-tests across ten cross-validation folds.

1.2 Outline

The rest of this thesis is constructed as follows: in Chapter 2 we describe the Facial Action Coding System and motivate our decision to use this framework. In Chapter 3 we conduct a wide-ranging survey of historical and contemporary FER systems in order to discover which techniques and algorithms already exist. We place particular emphasis on the feature types that each surveyed FER system uses. Chapter 4 provides a derivation of the support vector machine (SVM) due to its importance in the FER literature. In Chapter 5 we assess whether local analysis of the face around particular features such as the mouth and

eyes can improve recognition accuracy as well as increase run-time performance. We use support vector machines and Gabor features for this study. The results of this experiment underline the importance of establishing a large, publicly available facial expression database in which individual facial actions occur independently of others. Later in Chapter 5 we depart from the Gabor+SVM approach in order to test a new method of detecting FACS AUs: Haar wavelet-like features classified by an Adaboost strong classifier. Our results show that this new technique achieves the same recognition accuracy for certain AUs but operates two orders of magnitude more quickly than the Gabor+SVM method.

In Chapter 6 we use FACS as an intermediary expression coding framework and apply the FER system developed in Chapter 5 to our target application domain of SASL recognition. While the actual recognition results of this pilot study are unsatisfactory, we believe that the system architecture as well as the particular problems we encountered will be useful when designing future such systems. Finally, Chapter 7 suggests directions for future research.

With regards to the pilot project on signed language recognition we make one disclaimer: This thesis does *not* constitute *linguistic* research on South African Sign Language or signed communication in general. The purpose of this pilot application is to assess whether a simple object recognition architecture can support viable automatic signed language recognition, and to discover the most pressing problems that need to be solved in support of this goal. By implementing a software prototype of a SASL expression recognizer, we also provide future researchers of the South African Sign Language Project a firm starting point from which to conduct further research.

Chapter 2

Facial Action Coding System

In this thesis we use the Facial Action Coding System (FACS) [EF78] as an intermediary framework for recognizing the facial expressions of South African Sign Language (SASL). Two other research groups also use a FACS-based approach for their signed language recognition systems: the group of Professors Ronnie Wilbur and Aleix Martinez at Purdue University [Wil], and Ulrich Canzler [Can02] at the RWTH-Aachen. In order to motivate our own decision to use FACS, we must first describe the purpose and design of FACS and compare it to other representations that describe human facial expression. Later in this chapter we discuss the advantages and disadvantages of using FACS for our end-goal of automated SASL recognition.

2.1 Purpose of FACS

The primary goal of FACS was "to develop a comprehensive system which could distinguish all possible visually distinguishable facial movements" ([EFH02], p. 2). In contrast to other systems for facial expression coding, the development of FACS was governed by "the need to separate inference from description." In other words, the investigation of which emotion caused a particular facial expression should be determined independently from the description of the facial expression itself.

FACS is based on an eight-year, highly-detailed anatomical study of the muscles which control the face. It was designed to measure every *visible* movement of the face due to the contraction of facial muscles. In contrast to certain intrusive methods such as electromyography, in which wires must be connected to subjects' faces, FACS was designed for use on humans who are perhaps unaware of the fact they are being studied; coding of facial expression is therefore performed using only visual measurements. For this reason, FACS is not intended to measure muscle movements which result in no appearance change or whose effect on the face is too subtle for reliable human perception. FACS also does not register changes in facial appearance due to factors unrelated to muscles, e.g., blushing or sweating [EFH02].

2.2 The Design of FACS

FACS' approach is to specify the minimal units of facial behavior. These units are known as *action units* (AUs). Some AUs have a one-to-one correspondence with a particular facial muscle. AU 13, for example, corresponds solely to the *caninus* muscle. Other AUs may be generated by any one of a set of face muscles whose effects on the face are indistinguishable from each other. In yet other cases, multiple AUs may be linked to the same muscle if different parts of that muscle can be activated independently. Both AUs 7 and 8, for example, pertain to *orbicularis oris* [EFH02].

Each AU is assigned a number to facilitate coding of faces. In the original FACS definition in 1978 [EF78], there were 44 AUs whose numbers ranged from 1 through 46 (numbers 3 and 40 are not used). The updated 2002 edition [EFH02], which incorporated movements of the eyeball and head, contains an additional 12 AUs numbered 51 and higher. In both editions, AUs 1 through 7 pertain to the upper-face actions whereas AUs numbered 8 through 46 relate to the lower face.

For each AU in FACS, the *FACS Manual* [EFH02] provides the following information:

- The muscular basis for the AU, both in words and in illustrations.
- A detailed description of facial appearance changes supplemented by photographs and film examples.
- Instructions on how to perform the AU on one's own face.
- Criteria to assess the intensity of the AU.

2.2.1 AU Combinations

As AUs represent the "atoms" of facial expressions, multiple AUs often occur simultaneously. Over 7000 such combinations have been observed [Ekm82]. Most such combinations are *additive*, meaning that the appearance of each AU in the combination is identical to its appearance when it occurs alone. Some combinations, however, are *distinctive* (sometimes also called *non-additive*) - in such cases, some evidence of each AU is present, but new appearance changes due to the joint presence of the AUs arise as well. In the *FACS Manual*, the distinctive AUs are described in the same detail as the individual AUs.

Further relationships among multiple AUs exist as well. For instance, in certain AU combinations, the *dominant* AU may completely mask the presence of another, *subordinate* action unit. For certain such combinations, special rules have been added to FACS so that the subordinate AU is not scored at all.[1] Another relationship among AUs is that of *substitutive* combinations. In these cases, one particular AU

[1] Most such rules were removed in 1992 after it had been determined that they they were mostly confusing.

combination cannot be distinguished from another, and it is up to the FACS coder to decide which is more appropriate.

2.2.2 AU Intensity

In addition to determining which AUs are contained within the face, the intensity of each AU present must also be ascertained. Intensity is rated on a scale from A (least intense) through E (most intense). Criteria for each intensity level are given in the *FACS Manual* for each AU.

2.3 Suitability of FACS for Sign Language Recognition

In this project we chose FACS as our intermediary framework for facial expression recognition because of the level of detail it provides in describing expressions; because of its ability to code expression intensity; and because FACS is a standard in the psychology community. As we will describe in Chapter 6, we conducted a preliminary FACS analysis of 22 facial expressions that occur within SASL and determined that no pair of facial expressions contained exactly the same set of AUs. Although this study will have to be extended over more subjects and more expressions, it does support our belief that FACS is sufficiently detailed to enable sign language recognition.

2.4 Alternative Systems for Facial Expression Description

We are aware of only a few other systems designed to describe facial expressions in detail. One such system is the *Maximally Discriminative Facial Movement Coding System* (MAX), which was developed by C.E. Izard in 1979 [Iza79] and later updated in 1995. MAX was developed for psychological research on infants and small children, though with modification it can also be applied to persons of other age groups. Face analysis under MAX is performed using slow-motion video and proceeds in two stages. In the first stage, the face is divided into three regions: (1) the brows, forehead, and nasal root; (2) the eyes, nose, and cheeks; and (3) the lips and mouth. Each region is then analyzed independently for the occurrence of facial movements known as *appearance changes* (ACs). In the second stage, the ACs in each face region are classified either as one of eight distinct emotional states (interest, joy, surprise, sadness, anger, disgust, contempt, and fear), or as a complex expression comprising multiple simultaneous affects [Iza79]. Like FACS AUs, the MAX ACs are rooted anatomically in the muscles of the face. Unlike AUs, however, the set of ACs is not comprehensive of the full range of visually distinct human facial movement, nor does it distinguish among certain anatomically distinct movements (e.g., inner- and outer-brow movement) [OHN92]. MAX is therefore less appealing for signed language translation than FACS.

Another approach is the Moving Pictures Expert Group Synthetic/Natural Hybrid Coding (MPEG-4 SNHC) [Mov] standard. MPEG-4 SNHC uses 68 *facial animation parameters* (FAPs) to describe movements of the face. The purpose of MPEG-4 SNHC, however, is to animate computer-generated graphics, not to recognize the expression on an actual human's face. Correspondingly, the set of FAPs is not comprehensive of all visible human face movement, nor do the individual FAPs correspond to the actual muscle groups of the human face. As with MAX, it is unlikely to be of use in sign language recognition.

2.5 Why Use FACS for SASL?

In this thesis we endeavor to build an automated system for the recognition of SASL facial expressions by first determining the set of AUs present in a particular face image, and then mapping these AUs to a particular SASL expression. While we have already explained the advantages of FACS over other expression recognition frameworks, we have not yet motivated why we need an intermediary framework at all.

Using an intermediary expression description framework does add an additional layer of complexity to a translation system that recognizes SASL expressions directly from the input images. However, the advantage of using a framework for expression description such as FACS is that linguistic research on SASL and machine learning research on expression recognition can be de-coupled. For example, if a new expression is discovered in SASL, it can be accommodated simply by adding an additional AU-to-expression mapping to the translation system. The AU recognition code, on the other hand, remains completely unchanged. In systems that are trained on individual SASL expression directly, on the other hand, a whole new set of training examples containing this newly-found expression must be collected, and a new classifier must be trained - this requires significant time and effort. We thus believe that the use of an intermediary framework, especially FACS, is a worthwhile component of our system design.

2.6 Summary

We have described the purpose and basic architecture of FACS, including its set of action units and intensity ratings. We have explained some of the advantages of FACS over other expression coding systems for the task of signed language translation. Finally, we justified our use of an intermediary framework such as FACS in our SASL expression recognition system.

Chapter 3

Literature Review

Automatic facial expression recognition (FER) is a sub-area of face analysis research that is based heavily on methods of computer vision, machine learning, and image processing. Many efforts either to create novel or to improve existing FER systems are thus inspired by advances in these related fields.

Before describing our own contributions to the field of automatic FER, we first review the existing literature on this subject. This survey includes the major algorithms that have significantly impacted the development of FER systems. We also describe more obscure algorithms of FER both for the sake of comprehensiveness, and to highlight the subtle benefits achieved by these techniques that may not be offered by more mainstream methods. In accordance with the experiments we perform in Chapter 5, we place particular emphasis in our survey on the role of feature type, and on the effect of local versus global face segmentation on classification performance.

3.1 Comparing the Accuracy of FER Systems

Objectively comparing the recognition accuracy of one FER system to another is problematic. Some systems recognize prototypical expressions, whereas others output sets of FACS AUs. The databases on which FER systems are tested vary widely in number of images; image quality and resolution; lighting conditions; and in ethnicity, age, and gender of subjects. Most databases include subjects directly facing the camera under artificial laboratory conditions; a few (e.g., [KQP03]) represent more natural data sets in which head posture can vary freely. Given such vastly different test datasets used in the literature, only very crude comparisons in accuracy between different FER systems are possible. However, for the sake of completeness, we do quote the reported accuracy of the systems we reviewed.

The most common metric of recognition accuracy used in the literature is the percentage of images classified correctly. An accuracy of 85% would thus mean that, in 85 out of 100 images (on average), the

expression was predicted correctly, and in 15 images it was not. This metric is natural for characterizing a face as belonging to one of a fixed set of k emotions. For FACS AU recognition, however, this metric can be highly misleading: some expressions occur so rarely in certain datasets that a classifier could trivially always output 0 ("absent") for the expression and still score high accuracy. In such a system, even though the hit rate (% of positively labelled images classified correctly) would be low (0%), the percentage of images correctly classifier would still be high. A more sophisticated measure of recognition accuracy is the area under the ROC curve, also called the A' statistic, which takes into account both the true positive and false positive rates of a classifier. We use the A' metric in our own experimental work in Chapter 5. Most previous literature on FER presents results only as percent-correct, however, and in this literature review we are thus constrained to do the same.

3.2 Local versus Global Segmentation

The first issue we investigate, both in this survey and in Chapter 5, is whether analyzing a local subregion of the face around particular facial muscles can yield a higher recognition accuracy of certain FACS AUs than analyzing the face as a whole. Little research has been conducted on this issue for prototypical expressions, and no study, to our knowledge, has assessed the comparative performance for FACS AUs. Results for prototypical expressions are mixed:

Lisetti and Rumelhart developed neural networks to classify faces as either smiling or neutral [LR98]. They compared two networks: one which was trained and tested on the whole face, and one which was applied only to the lower half of the face (containing the mouth). For their application, local analysis of the lower face-half outperformed the global, whole-face analysis.

Padgett and Cottrell compared global to local face analysis for the recognition of six prototypical emotions. In particular, they compared principle component analysis (PCA) on the whole face (*eigenfaces*) to PCA on localized windows around the eyes and mouth (*eigenfeatures*). The projections onto the eigenvectors from each analysis were submitted to neural networks for expression classification. As in Lisetti and Rumelhart's study, the localized recognition clearly outperformed global recognition. Padgett and Cottrell attribute these results both to an increased signal-to-noise ratio and to quicker network generalization due to fewer input parameters [PC97].

However, Littlewort, et al [LFBM02] compared whole-face, upper-half, and lower-half face segmentations for the recognition of prototypical facial expressions. They classified Gabor responses (described later in this chapter) using support vector machines. In contrast to the other literature on this subject, their whole-face segmentation clearly outperformed the other two segmentation strategies by several percentage points [LFBM02].

From the literature, there seems to be no definite answer as to which segmentation - local or global - yields higher accuracy. As we shall show in Chapter 5, the issue depends on the particular facial expression database on which the system is tested. It may also depend on the particular *feature type* that is used. In the rest of this chapter, we describe the many kinds of features that have been deployed for FER as well as the systems that deploy them.

3.3 Feature Extraction for FER: The Two Approaches

Research on automatic FER can largely be divided into two categories: *appearance-based* and *geometry-based* methods. The former uses color information about the image pixels of the face to infer the facial expression, whereas the latter analyzes the geometric relationship between certain key points (*fiducial points*) on the face when making its decision. We describe geometry-based methods in Section 3.4 and appearance-based methods in Section 3.5.

3.4 Geometry-based Features

Many modern FER systems use the geometric positions of certain key facial points as well as these points' relative positions to each other as the input feature vector. We refer to such FER systems as *geometry-based* systems. The key facial points whose positions are localized are known as *fiducial points* of the face. Typically, these face locations are located along the eyes, eyebrows, and mouth; however, some FER systems use dozens of fiducial points distributed over the entire face.

The motivation for employing a geometry-based method is that facial expressions affect the relative position and size of various facial features, and that, by measuring the movement of certain facial points, the underlying facial expression can be determined. In order for geometric methods to be effective, the locations of these fiducial points must be determined precisely; in real-time systems, they must also be found quickly. Various methods exist which can locate the face and its parts, including optic flow, elastic graph matching, and Active Appearance Models ([CET98]). Some FER systems (e.g., [TKC01]) require manual localization of the facial features for the first frame in a video sequence; thereafter, these points can be tracked automatically. Other approaches to fiducial point location do not actually *track* the points at all, but instead re-locate them in each frame of the video sequence.

The exact type of feature vector that is extracted in a geometry-based FER systems depends on: (1) which points on the face are tracked; (2) whether 2-D or 3-D locations are used; and (3) the method of converting a set of feature positions into the final feature vector. The first question (1) has no definitive best answer, but it is influenced by several factors, including (a) how precisely each chosen fiducial point can be tracked; and (b) how sensitive is the position of a particular fiducial point to the activation of the