ACTIVE APPEARANCE MODELS FOR AFFECT RECOGNITION USING FACIAL EXPRESSIONS

Matthew Stephen Ratliff

A Thesis Submitted to the
University of North Carolina Wilmington in Partial Fulfillment
of the Requirements for the Degree of
Master of Science

Department of Computer Science
Department of Information Systems and Operations Management

University of North Carolina Wilmington

2010

Approved by

Advisory Committee

Curry Guinn

Thomas Janicki

Eric Patterson
Chair

Accepted by

DN: cn=Robert D. Roer, o=UNCW, ou=Dean of the Graduate School & Research, email=roer@uncw.edu, c=US
Date: 2010.10.27 11:30:14 -04'00'

Dean, Graduate School

ABSTRACT

This paper will explore the effectiveness of active appearance models (AAM) at extracting emotional information as well as usefulness of AAMs in the role of emotion classification. This paper examines what has been accomplished recently in this field by reviewing the various types of classification schemes and databases used. Three types of classification techniques are presented: Euclidean Distance Measure, Gaussian Mixture Models, and the Support Vector Machine. Each will be gauged by its effectiveness in the classification of emotional expressions.

The technique presented involves the creation of an Active Appearance Models (AAM) trained on face images taken from a publicly available database. This model is a representation of the shape and texture variation of the image, which is key to expression recognition. In each experiment model parameters from the AAM are used as input into a classification scheme, which is used for expression identification. The results of this study will demonstrate the effectiveness of AAMs in capturing the important facial structure for expression identification and also help suggest a framework for future development.

TABLE OF CONTENTS

ABSTRACT ii
TABLE OF CONTENTS iii
LIST OF TABLES v
LIST OF FIGURES vi
1 INTRODUCTION 1
1.1 Problem Statement 3
1.1.1 Hypothesis I 4
1.1.2 Hypothesis II 4
2 BACKGROUND MATERIAL 5
2.1 Eigenfaces 6
2.2 Facial Action Coding System 7
2.3 Active Appearance Models 10
3 PREVIOUS APPROACHES 14
3.1 Expression Databases 16
3.1.1 JAFFE Database 17
3.1.2 Cohn-Kanade Database 18
3.1.3 FG-NET Database 21
3.1.4 Kinetic Facial Expressions 22
3.2 Data Collection Methods 25
3.3 Feature Extraction Methods 29
3.3.1 Template-based Methods 29
3.3.2 Feature-based Methods 31
3.4 Classification Techniques 34
4 PROPOSED SYSTEM 41
4.1 Architecture 43

4.1.1 Data Collection 43
4.1.2 Data Analysis and Segmentation 44
4.1.3 Model Creation 47
4.2 Classification 49
4.2.1 Euclidean Distance 49
4.2.2 Gaussian Mixture Model 53
4.2.3 Support Vector Machine 58
5 DISCUSSION 64
6 CONCLUSIONS 72

LIST OF TABLES

4.1 Images were evaluated for their use in the study. 46
4.2 Total number of images collected for the study. 46
4.3 Euclidean classification results by subject. 51
4.4 Confusion matrix for Euclidean classification. 52
4.5 GMM classification results by subject. 57
4.6 Confusion matrix for GMM MLE classification. 58
4.7 Confusion matrix for SVM classification. 62
4.8 SVM Classification results by subject. 63

5.1 Comparison of results from previous studies. 65
5.2 Comparison of results from previous studies. 66
5.3 Comparison for number chosen samples 67
5.4 Recognition results reported by previous studies for Ekman's six emotions. 68
5.5 Confidence Interval for correct classification of "sad". 69
5.6 Confidence Interval for correct classification of "joy". 69
5.7 Confidence Interval for correct classification of "fear". 69
5.8 Confidence Intervals across all emotions for the GMM classification scheme. 70

LIST OF FIGURES

2.1 Eigenfaces provided by ATT Laboratories Cambridge 7
2.2 Conceptual image representing the principal components of a distribution. 11

3.1 Some of the expressions available in the JAFFE database. 18
3.2 Some of the expressions available in the Cohn-Kanade database. 19
3.3 Some of the expressions available in the FG-Net database. 21
3.4 Some of the expressions available in the DaFEx database. 24

4.1 Landmarked location points used in training the AAM. 43
4.2 Sample of faces in the FG-Net database. 45
4.3 Face being labeled during model creation. 47
4.4 Model created from training data. 48
4.5 Conceptual diagram demonstrating euclidean distance measure. 50
4.6 The left image shows an expression of “anger”, while the image to the right is expressing “sadness”. 53
4.7 Conceptualization of four data regions along with decision boundaries for each region [9]. 56
4.8 Conceptual image demonstrating an optimal hyperplane separating classes [9]. 60
4.9 Conceptual image demonstrating an optimal region separation [9]. 61

INTRODUCTION

Recognizing emotion from facial expressions represents a key element in human communication. Facial expressions constitute 55% of the effect of a communicated message [28]. The face provides visual feedback and aids in social interaction by expressing such things such as mood, speaking turn, and confusion. Studies suggest up to 93% of all human communication is done by interpreting emotion. In order for machines to communicate effectively with people, they must also have this ability. People often change their responses according to feedback received from others, and we often rely on facial expressions in many cases to give further meaning to a conversation. Other queues such as speech prosody and content give us some indication of the meaning behind spoken words, but in most cases it requires a multi-modal approach using both speech and visual feedback.

As society relies ever more on technology, the need to simplify and enhance human-computer interaction becomes clear. Machines operate by running programs that process instructions according to the design of the developer. If we could create machines that could understand the emotional element of communication, as well as recognize the individual to which they are communicating, we would find that such communication would be more effective and less frustrating to the user.

For ideal human-computer interfaces (HCI), we would desire that machines have the ability to not only recognize emotions, but have the ability to express them as well. Computer applications could better communicate by changing responses according to the emotional state of human users in various interactions. In order to work toward these capabilities, efforts have recently been devoted to integrating affect recognition into human-computer applications [46]. By creating machines that can understand emotion, we enhance the communication that exists between humans and computers. This would open a variety of possibilities in robotics and human-computer interfaces such as devices that warn a drowsy

driver, attempt to placate an angry customer, or improve security systems.

For two-way communication to exist, emotions are not only recognized, but also conveyed. Computer science has tackled the problem from both ends by not only creating systems that interpret expressions but also display emotion through various interfaces. We can see such systems at work in recent video games. Avatars are now being created that express emotions. Emotionally expressive agents in video games results in a more realistic feeling to the player. Microsoft is in the process of creating a better recognition console for gaming, known as "Natal". This system works similar to Nintendo's "WII" gaming console, but has been enhanced to include more gesture recognition. The evolution of these systems demonstrates the need to improve HCI devices in gaming, with the end result being a system fully capable of recognizing not only gestures but also expressions.

Facial expressions provide a key mechanism for understanding and conveying emotion. Even the term "interface" suggests the primary role of the face in communication between two entities. Studies have shown that interpreting facial expressions can significantly alter the interpretation of what is spoken as well as control the flow of a conversation [28]. Improving the communication between humans and computers has been one of the driving forces in computer science over the years. This research has involved coordination between several fields of research including: psychology, biology, computer science, and engineering. The key to creating a better communication between humans and machines involves developing more sophisticated techniques for interpreting responses.

Charles Darwin demonstrated in 1872 the universality of facial expressions and their continuity in man and animals. Biology approaches the topic from an evolutionary standpoint citing that the ability to recognize emotions stems back to the earliest days of man when sexual selection played key roles in survival. Before humans had the ability to vocalize thoughts, facial expressions allowed them to communicate feeling and intent. Because of this, we instinctively use facial queues during communication.

Two main challenges for engineering and computer science have been in the areas of facial feature extraction and classification. Ekman and Friesen have worked to develop a system for mapping the muscle movement of facial areas to different expressions [14]. Psychology has played an important part in this endeavor by providing valuable information regarding how people communicate as well as how we exhibit and respond to various emotions. Biology has shown us how people have used expressions as a means of survival even over vocal communication. Ekman and Friesen have been pioneers in this effort with the identification of six basic emotion categories, which include: anger, fear, disgust, joy, surprise, and sadness. Recently a seventh category "contempt" was added to this list. It has been found that this research is used by more researchers than other categorical models created for emotion research [1]. Ekman states that expressions are fairly static cross-culturally in expressiveness, which supports the idea of creating a coding system.

Criticisms have recently surfaced in response to the six categories proposed by Ekman and Friesen. Some psychologists and researchers believe that emotions cannot be categorized so easily and instead contain a higher dimension than once previously thought [39]. As a first approach it is valid to consider the simpler case first, and then branch off into the other flavors of emotion such as bordeum, satire, interest, and frustration. Several studies have moved away from template-based approaches to those of feature-based, where the individual features which drive expression are used. Fasel states that coding facial expressions directly into basic emotion categories has several drawbacks: 1) emotion categories can only describe a subset of all facial expressions 2) emotions constitute an interpretation of facial actions, rather than a description, and 3) the judgement of different coders can vary a great deal. [15]

1.1 Problem Statement

Emotion recognition using facial expressions is currently an active area of research. This study will investigate several different types of techniques that have been used along with their successes. This study will provide a simple and straightforward method for recog-

nizing emotional expressions using Active Appearance Models. In addition, three common classification schemes will be compared for effectiveness which will gauge the success of using Active Appearance Models. Two hypotheses are presented which will be addressed in the "Discussion" section of this paper once the study is complete.

1.1.1 Hypothesis I

Active Appearance Models can serve as an effective feature extraction method for expression recognition. This will be gauged by comparison with recent results by researchers in the area.

1.1.2 Hypothesis II

Certain expressions are more distinguishable than others. Results from several classifiers will be compared to see if certain emotion-based expressions are more or less difficult on average to identify.

BACKGROUND MATERIAL

Model-based classification methods rely on using models of known structures in order to create "plausible" interpretations of the structure they represent. Tim Cootes and his colleagues created such a system to be used in medical imaging [11]. Medical images often contain noisy and incomplete evidence. Model-based methods can solve this complexity problem by creating a model which best represents the given image or object using a set of predefined points.

One of the main issues with most classification problems is that of variability. If all objects in a given class of objects were identically shaped, deciding the class to which a given object belongs would be relatively simple. Variability in shape, texture, and size makes classification difficult. According to Cootes and Taylor [11], models should possess two main characteristics: First, they should be able to create any plausible example of the class to which they are assigned, and second they should only generate legal, "plausible" examples of the class. Cootes and Taylor call these types of models "deformable" in that they a range of examples described by training data.

Model based approaches are useful in a variety of applications, due to their generalization. The model created is specifically designed and suited for the given problem. Another reason models are advantageous is that models contain expert knowledge of the problem given. Rather than a random choosing of points for model creation, points can be chosen specifically for that problem. In emotion recognition, for example, one key feature that helps humans to differentiate between the classes of emotional expressions are the corners of the mouth. Using a non-model based method, this knowledge might not be encoded into the solution, making it difficult to create an effective solution for that problem. By determining which points best represent emotion in expressions, points can be chosen in advance, thus encoding expert knowledge for the given problem.

2.1 Eigenfaces

Matthew Turk and Alex Pentland introduced a concept of averaging using facial templates for the purposes of face recognition known as "Eigenfaces" [43]. Their goal was to create a complete model which would be fast, reasonably simple, and accurate in constrained environments. Using a model-based approach would free any such system from the need to use detailed geometry to detect small regions of the face, which is known to be computationally expensive as well as complex.

Their approach transforms face images into a small set of characteristic feature images, called "Eigenfaces", which are the principal components of the training set of facial images. Turk and Pentland coined the term "face space" which has been used extensively in this new area of research. This technique created the framework for learning new faces in an unsupervised way. Several techniques up to that time had been investigated, but all lacked the generalization that could come from a template-based method. Most worked by detecting individual features such as the eyes, nose, mouth and head outline, as well as the relationships that exist among those features. According to Turk and Pentland, such attempts proved difficult to extend to multiple views and were often quite fragile [43].

The goal of "Eigenfaces" was to find the principal components of the distribution of faces, also known as the eigenvectors of the covariance matrix of the set of face images. These eigenvectors provide information regarding the most prominent features that could be used to differentiate between face images. Because of this and the fact that the resulting face images are face-like in appearance, they are known as "Eigenfaces". Figure 2.1 demonstrates the visual result of this averaging of faces to find the most valuable features.

As seen in Figure 2.1 eigenfaces can be summed together to create an approximate grey-scale rendering of a human face. The eigenvectors derived from this method are the

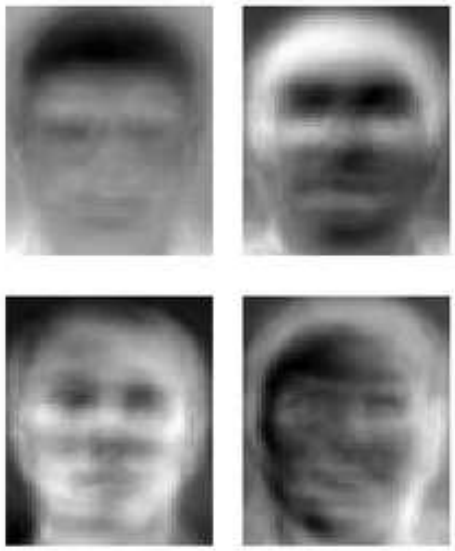

Figure 2.1: Eigenfaces provided by ATT Laboratories Cambridge

same as the eigenvectors found in principal components analysis. One of the strengths of this technique is the ability to create an "eigenface" from relatively few training images. The model requires little training data to be able to quickly identify a human face, but recognizing emotional expressions requires more training samples as the features used for those types of expressions are smaller and more difficult to determine.

2.2 Facial Action Coding System

The Facial Action Coding System (FACS), is a system developed by Paul Ekman and Wallace Friesen as a way to measure visible facial muscle movement [12]. This system, unlike AAMs, locates individual features while disregarding the face as a whole. Several studies in facial emotion recognition have subscribed to this system and use this technique for feature extraction [29]. Facial muscle movements have been coded into facial action units "AU". Facial movement can occur as a single unit, or more often occur in combination with other muscle movements. The muscles that relate to expressions work collectively, and as a result make it difficult to separate the areas of influence. One example of this is the

frontalis muscle (which covers the forehead). This muscle very seldom moves in alone, but in combination with other facial muscles. For example, the combination of the frontalis muscle and opening of the mouth are commonly associated with the "surprise" expression.

Individuals commonly use the intensity of an expression to better indicate emotion. Usually, the more intense the expression the more emotion is being conveyed. FACS uses this notion of intensity and has a scale from "A" to "E" which grades the intensity of an expression. "A" intensity indicates that not all of the muscle movement is present to meet the criteria for a specific AU action. "E" intensity represents the maximum possible movement of a particular AU action. The AU used in an expression in addition with an intensity scale represent the overall expression. In total there are 44 basic AUs which can be used independently or combination to describe almost any expression. These AUs are very suitable to be used in studies on human naturalistic facial behavior as thousands of anatomically possible facial expressions can be described by these 44 basic AUs. FACS tables were derived by anatomical and physiological studies of the human face, which aids in understanding how the bones and muscles affect facial movement. Once facial images are encoded using the appropriate AUs they are used to train a classification system. One major advantage of using this approach over Ekman's basic six expressions is the fact that any expression can be used. A study researching the effects of "bordeum", for example, could better benefit from a system not centered around a basic category.

The actual coding of AUs is performed by human FACS coders who are specially trained in identifying even the smallest changes in expression. Even details such as wrinkles or buldges can be identified and coded as needed. It has been shown in previous studies that age and skin texture play an important part in the consistency of expressions. Even though a person may use the same muscles for an expression, the skin surrounding those muscles change with age, which may introduce problems for facial emotion recognition systems.

FACS coders are trained by viewing videotaped facial behavior in slow motion [22]. Coders must pass a standardized test, which ensures uniform coding among international laboratories. As a general rule 15% to 20% of the data is comparison coded for consistency. To guard against coding drift restandarization is required periodically. A disadvantage to using this approach for feature extraction is the fact that this method takes over 100 hours of training to achieve minimal competency for a human expert [47].

The main goal of FACS is to create a method of feature extraction that can describe any expression needed for a particular study. Many studies have used FACS either directly or indirectly as a means of validating the extraction method chosen. Wong and Cho cites FACS as the leading method for measuring facial expression in behavioral science [47]. Rizon uses FACS for feature extraction and classifies expressions into one of the six basic expressions. Once the AUs are obtained from the training data Rizon uses this as input parameters into a SVM classifier. His results indicate an average success rate of 87% using FACS with this approach [36]. Feng-Jun reports using FACS with a Neural Network classification scheme [16]. Rather than using FACS coders specifically, Feng-Jun and his colleagues used the CMU-Pittsburgh AU-coded database. It can be expensive and time consuming to hire expert coders, so utilizing an existing AU coded database is preferred in most instances. Shang reports the use of FACS coding to train a Hidden Markov Model for expression in video [40]. This study reports the use of 58 facial points, which are extracted and classified into one Ekman's six basic expressions.

The ability of FACS to code specific movements has made it useful in many studies and has provided a way to study not only Ekmans's basic six emotions, but other emotions not considered mainstream. Building onto this framework Ekman, Friesen, and Rosenburg extended the idea of using AUs to the realm of emotion. The ever growing need for emotional feature identification and extraction led Ekman to develop the Emotion Facial Action Coding System (EMFACS). This framework builds upon FACS by targeting specifically those

areas of the face directly related to emotional expression.

EMFACS (Emotion Facial Action Coding System) allows a study to take an objective approach to emotion identification by using the existing FACs scoring system. Emotional expressions are fluid and the slightest movement can completely change the overall meaning of the expression. This system extends the ability of FACS specifically for recognizing emotions. The scoring is done without the aid of slowed motion viewing, and the location of an expression is simplified by requiring only a single locational point near the beginning of the event [6]. Since its conception FACS has been used and is now considered by many to be the best method of choice for facial feature identification and extraction.

2.3 Active Appearance Models

Active Appearance Models use a model-based approach and is built upon the framework provided by eigenfaces. Both eigenfaces and AAMs use the idea of determining the vectors that best represent change in expression over the entire face image. The system was first created by Tim Cootes and Taylor and involves building a statistical model of shape and appearance. The "Active Shape Model" manipulates a shape model to describe the location of structures in a target image [11]. The Appearance Model manipulates a model capable of creating new images similar to the one provided. Parameters are created from the AAM algorithm which represents the synthesized model. These parameters can then be used in a classification scheme in conjunction with parameters produced from the training images.

Shape modeling involves taking the points given and reducing the dimensionality to something more manageable. Points collected from training samples create a conglomeration of points in n-dimensional space. For dimensionality reduction Principle Components Analysis (PCA) are used. PCA operates by finding the "principal" components, or vectors which determine areas containing the most relevant information. Reduction occurs by removing the "noise" from the space by using these principal components as a representation

of the data. The data points are examined and the axes that best represent those points are chosen, as see in Figure 2.2.

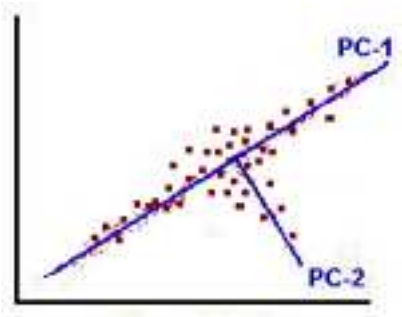

Figure 2.2: Conceptual image representing the principal components of a distribution.

Model approximation makes use of this dimensionality reduction by employing an algorithm that attempts to approximate a training set sample. AAMs can be used for either classification purposes or as a means of creating deformable models. Classification can be performed using the following method:

The mean of the training sample is computed:

$$mean(x) = 1/s \sum_{i=1}^{s} x_i$$

where s is the number of samples, and x is the training sample. Next, the covariance of the data is computed:

$$S = 1/(s-1) \sum_{i=1}^{s} (x_i - mean(x))(x_i - mean(x))^T$$

The final step is to compute eigenvectors and corresponding eigenvalues, which when combined with the algorithm for shape variation, will yield a vector which best approximates the training vector. The approximation vector (x) is computed using the following formula:

$$x \approx meanx + \Phi b$$

where $\Phi = (\phi_1 \mid \phi_2 \mid ... \mid \phi_t)$ and $\mathbf{b}$ is a t dimensional vector given by

$b = \Phi^t(x - mean(x))$

The vector $\mathbf{b}$ defines a set of parameters of a deformable model. Varying the elements of vector $\mathbf{b}$ will create different shapes of models similar to that given in $\mathbf{x}$. Shape variation is but one of two types of variations that needs to be modeled to create an effective system. Variation in texture also needs to be considered, as this will describe the lighting changes in the image. Models are generated by combining a model of shape variation with a model of the texture variation in a shape-normalized frame [11]. Cootes and Taylor describe "texture" as the pattern of intensities or colors across an image patch. Lighting is very important, and too much variation of color or lighting schemes will create an ineffective model. Similar to the process for reducing dimensionality in shape, PCA is also applied to the points. Each point in the texture model has a value which is a grey-scale value of the image at that point. Once the texture and shape models have been created, they are combined to produce a single vector which represents the image, both in color and shape. For each training sample the single vector is calculated by concatenating the texture and shape vectors, as follows:

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix} = \begin{pmatrix} W_s P_s^T (x - mean(x)) \\ P_g^T (g - mean(g)) \end{pmatrix} \tag{2.1}$$

where W_s is a diagonal matrix of weights for each shape parameter, allowing for the difference in units between the shape and grey models, and b being the parameters for both shape b_s and texture b_g. The result is the creation of a single vector combining both grey scale and texture variations. Once the overall vector has been created from the shape and texture parameters, PCA is performed again to reduce the dimensionality into the principal components. The result is a single vector which represents a particular face.

Several studies report using AAMs for facial and emotion recognition studies. Once such study performed by Yunus Saatci and Chirstopher Town use AAMs for the identification of emotional expressions as well as gender identification [38]. In their study four separate AAMs were created, which correspond to four basic emotions ("happy", "angry", "sad", and "neutral"). Saatci points out one of the weakness of using AAMs deals with lighting variations. AAMs work well in controlled light settings, but adding variation distorts the models resulting in inaccurate parameters. To compensate with this situation Saatci had to employ a histogram equalization scheme. Such systems are used to deal with these types of lighting problems commonly found among model-based approaches.

Datcu and Rothkrantz use AAMs for face recognition in still images and video [8]. Their system uses AAMs in conjunction with a Facial Characteristic Point (FCP) for determining the best points of interest. The resulting parameters are fed into a Support Vector Machine for classification. Their results conclude that AAMs were effective in the process of feature extraction for the purposes of their study. Along with their use in classification systems, AAMs are also widely used for their ability to model and segment deformable visual objects. This method makes use of linear subspaces, which allows for a compact representation of both shape and texture [34]. Using this approach Obaid and his colleagues used a basic Euclidean distance measure for classification. Results show a successful classification rate of 89%.

PREVIOUS APPROACHES

Facial expressions provide the building blocks with which to understand emotion. In order to effectively use facial expressions, it is necessary to understand how to interpret expressions, and it is also important to study what others have done in the past. Fasel and Luettin performed an in-depth study in an attempt to understand the sources that drive expressions [15] . Their results indicate that using FACS may incorporate other sources of emotional stimulus including non-emotional mental and physiological aspects used in generating emotional expressions.

Much of expression research to date has focused on understanding how underlying muscles move to create expressions [15][18]. For example, studies have shown that movement of the nasolabial furrow, in addition to movement of the eyes and eyebrows, is a primary indicator of an emotional expression [42]. FACS as well as EMFACS go into much detail regarding facial poses and have served as important references and a system of study for other work [12][13].

Though it was originally used for the analysis of facial movement by human observers, FACS has been adopted by the animation and recognition communities. Cohn et al. report that facial expressions can differ somewhat from culture to culture for a particular emotion, but the similarities in expression for an emotion are usually strong enough to overcome these cultural differences [5].

Much of previous work has used FACS as a framework for classification. In addition to this, previous studies have traditionally taken two approaches to emotion classification according to Fasel and Luettin [15]: a judgment based approach and a sign-based approach. The judgment approach develops the categories of emotion in advance such as Ekman's six universal emotions. The sign-based approach uses a FACS system, encoding action units

in order to categorize an expression based on its constituents. This approach assumes no categories, but rather assigns an emotional value to a face using a combination of the key action units that create the expression.

FACS laid the framework for understanding the relationship between muscle movement and expressions. Several studies have built upon this knowledge to create different types of classification systems for emotion recognition. These approaches can be divided into four main categories: a) expression databases b) data collection b) feature extraction methods c) classification techniques.

3.1 Expression Databases

Recent studies suggest that facial expression databases have been limited among the research community [35]. In order to create a successful mechanism for classifying expressions, robust and reliable databases are needed. In the past most studies have created their own expression database. Collecting emotional expressions can be problematic, and usually takes two different forms: a) Those whose expressions are elicited b) Those given by actors. Elicited expressions are preferable over those given by actors due to their spontaneous nature and genuineness. Emotional expressions given by actors have been used in several studies [31][35][50], but research suggests eliciting emotional responses results in a more genuine expression [41].

Several problems arise when trying to create a database of emotional expressions. First, a choice must be made to use either actors, or to try and elicit emotions. In order to evoke emotional responses, the subjects typically involved are not aware of the process. Emotions such as "joy" and "surprise" can be easily evoked with little to no harm done to the subject. Emotions having negative connotations such as "sadness" and "fear" are difficult to elicit without moral implications. In order to evoke ''sadness", for example, the subject would have to be given some information which would lead to a feeling of loss, and or depression. So the dilemma becomes how to collect the data without causing harm to the subjects involved.

Recently there has been a move toward a hybrid approach. With this type of approach the process is explained to the subject before the process begins. Subjects are not told to express certain emotions, only that they simply are aware that an emotional response is needed. Responses are then elicited through a series of video clips. Videos are shown based on the emotion which is needed for the study. A scary video, for example, may be shown to try to capture a fearful expression.

The advantage of using actors is having control over the experiment. With elicited responses there is no guarantee that the response needed will be the one given. In addition, without detailed knowledge of the experiment the subject would less likely to adhere to certain rules required for the experiment. By communicating the intent with the subject, the person in control of the experiment would be able to guide the subject into what is needed and expected[22].

Work continues in developing good databases for facial expressions containing emotional content. A wiki has been created to aid researchers in locating emotional expression databases: http://www.emotion-research.net/wiki/databases. Many of these databases consist of static facial expressions (images) with little work devoted to video sequences. As the debate continues among which type of system works best, I found the following databases to be referenced more than any others in studies of expression recognition.

3.1.1 JAFFE Database

The Japanese Female Facial Expression Database (JAFFE) consists of 213 images with seven different expressions given by the 10 female subjects. This database uses Ekman's six universal emotions, along with the additional category of "neutral". Each image in the database was graded by 60 different Japanese graders to ensure consistency and relation to the emotion being expressed. "Fear" was excluded from the ratings due to the inaccuracies and lack of genuine expression. This study indicates that there is some evidence in the scientific literature indicating that "fear" may be processed differently from other expressions [30].

The creation of the images involved the subjects being told which emotion to express. So rather than the emotion being evoked, the subjects simply provided the expressions needed. Several studies make reference to this database, including a study performed by Bin Hua and Ting Liu for the recognition of facial expressions based on Local Feature Bidi-

rectional 2DPCA [20]. Sun states that this database does not represent authentic facial expressions for corresponding emotional states [41]. Other studies also make reference to this database for facial expression recognition [35] [50] [37]. Figure 3.1 shows a sample of the faces available.



Figure 3.1: Some of the expressions available in the JAFFE database.

Even though this database is referenced in mainstream research, it still has issues in the fact that it is limited in both gender and ethnic background. In order to create a database that generalizes across cultures, gender, and ages, it is necessary to use subjects with varying age and ethnic backgrounds. In order for a system to correctly classify an unseen genuine expression it should be trained with databases containing the same types of facial images. The minor details that differentiate genuine expression from those given by an actor may cause a recognition system to return inaccurate results.

3.1.2 Cohn-Kanade Database

Referred to as one of the largest publicly available facial expression databases[15], this database was created by an interdisciplinary research group consisting of psychologists and computer scientists at the University of Pittsburgh. This is an AU-Coded database which means that FACS Action Units are applied to provide interested groups with the ability to use FACS encoding for classification. Facial behavior was recorded in 210 adults between the ages of 18 and 50 years. From this group 31% were male and 69% female. The ethnic

background includes Euro-American 81%, Afro-American 13%, and another category of 6%.

Subjects were recorded using two Panasonic WV3230 cameras connected to a Panasonic AG-7500 video recorder. One camera was positioned directly in front of the subject, while the other were positioned 30 degrees to the subjects right. The room was lit using three high-intensity lamps. One of the lamps was used to illuminate one-third of the room, while the other two lamps were fitted with reflective umbrellas to ensure uniform lighting within the room. Figure 3.2 shows a sample of images contained in this database.



Figure 3.2: Some of the expressions available in the Cohn-Kanade database.

Subjects were instructed to perform a series of expressions which included single action units as well as AU combinations. Each series begins and ends with a neutral expression. According to Ekman all expressions originate from a neutral position[14]. To date approximately 1917 image sequences from 182 subjects have been FACS coded for either the action units or the entire sequence. Approximately 15% of the subjects coded have been cross-validated by a certified FACS coder. AU 13 was not recorded due to the difficulties in expressing this movement. Studies have found that this AU can only be recorded through spontaneous involuntary behavior [22].

EMFACS are also applied allowing emotional expressions to be evaluated. The emotion specific expressions included are joy, surprise, sadness, disgust, anger and fear. These

expressions can be combined to form aggregates of positive and negative expressions [22]. One limitation of the database is the lack of spontaneous expressions, which seems to be a common problem among several expression databases.

In order to combat this problem the development group is currently reviewing videos taken in an attempt to identify spontaneous expressions and categorize them appropriately. Also, as an extension to the database the group is pursuing more diversity in age groups by including infants and children, as well as adults. Emotions in infants and children are more likely elicited rather than planned due to their limited ability to follow instructions. One challenge that the group currently faces is how to capture spontaneous emotions while controlling the scene. Head movement and occlusions can corrupt sequences making them difficult to use for classification, as was seen in this study.

The Cohn-Kanade database is one of the most widely used in the research community due to its uniqueness for being FACS encoded [33][16][4][34]. One of the main criticisms of this database is the use of actors rather using spontaneous expressions. Kanade and Cohn make reference to the difficulty of using AU 13 due to the inability of people to generate this expression voluntarily [22]. With this in mind, an elicited system would not have this problem since expressions would be genuine. One of the things that Kanade and Cohn want to pursue in furthering this database is to add more variability to the database as well as more images containing head movement and scene complexity. This study found that most systems tend to move away from introducing scene complexities in order to avoid feature extraction and classification problems. Kanade and Cohn indicate that for a system to be truly effective it must be able to handle noise and have the ability to filter out unwanted or irrelevant information. An advantage of using this database is being able to use the images without the worry of incorrect expressions. One common problem among those created from elicited expressions is that they will often convey an unintended expression, or an overall lack of intensity.

3.1.3 FG-NET Database

This publicly available database was created by Frank Walhoff from the Technical University of Munich. This image database was created in an attempt to aid researchers with studies dealing with emotion recognition, and was generated as part of the European Union project FG-NET (Face and Gesture Recognition Research Network)[45].

This database consists of 18 different subjects (male and female) between the ages of 23 and 38. Video is captured using a Sony SC-999P camera equipped with a 8mm COSMICAR 1:1.4 television lens. A BTTV 878 framegrabber card was used to grab the images with a size of 640x480 pixels, a color depth of 24 bits and a framerate of 25 frame per second. For capacity reasons, images were converted into a 8 Bit JPEG-compressed format with a size of 320x240. Approximately four video sequences were captured for each subject for each of Ekman's six basic emotions, which include (joy, surprise, disgust, anger, fear, sadness). A sample from this database can be seen in Figure 3.3.



Figure 3.3: Some of the expressions available in the FG-Net database.

Emotions are elicited by showing subjects various films designed to elicit particular emotions. The videos are augmented with some type of action to cause a spontaneous behavior in the subject. One example given is a gunshot given at just the right time in the video to create a surprise expression. One common problem in all emotional expression databases is how to elicit an emotion while maintaining control of the scene. Because of the way these

sequences were created, several of the images are unusable. Head movement, unexpected behavior, and occlusions require several images and video sequences to be removed from the set before training. In some instances one type of emotional response was expected and another was received.

A major limitation to this database is the lack of a wide-spread ethnic background. All of the subjects were of European-caucasian decent. As seen in the CMU database, providing images from a variety of cultural and ethnic backgrounds is preferred. "Fear" is one expression that suffers from a lack of data samples. The difficulties in eliciting "fear" can be seen in this database as the expression is either non-existent, or lacks intensity.

One major advantage this database has over CMU and JAFFE is the genuineness of the expressions. Even though several of the videos and still images are unusable, those that can be used do offer genuine expressions. In one instance, subject number 15 appears to be crying when viewing the video to elicit sadness. It is plain to see that by the coloration of the face and overall expression that a feeling of pain is being expressed.

Many studies use several different types of databases when conducting classifications to see if any one database performs better than another. Zhan references the FG-Net database for a study involving on-line game avatars [50]. Saatci and Town performed their study by combining data samples from the FG-Net database as well as two others. The hope was to add more variability to the training data, which could result in more generalization and better classification across ethnic backgrounds. Their results indicate that the combination of these databases in addition to using a SVM classifier resulted in successful recognition of up to 94.4% for happy, and a minimum recognition of 63.6% for neutral.

3.1.4 Kinetic Facial Expressions

The DaFEx database was originally created for the purposes of providing a valid benchmark for the evaluation of facial expressivity [1]. The DaFEx consists of 1008 short video

clips, lasting between 4 and 27 seconds, each showing a facial expression corresponding to of the Ekman's six emotions plus an additional category for neutral. The facial expressions were collected by using 8 Italian professional actors (4 males and 4 female). Each emotion was recorded, by every actor at 3 intensity levels (low, medium, and high). Recordings were made with both speech and non-speech sequences. The speech sequences were developed specifically for multi-modal techniques.

Actors were given a paper form containing the guidelines for the recording of facial expressions (which indicates that none of the expressions were elicited). The form contains a detailed explanation of the terminology used in the recordings, along with the specification of the recording sequence of the facial expressions. Actors were explained the purposes of the work as well as the importance of making expressions looking as natural as possible. Actors were also given a second form containing six emotionally expressive stories, whose purpose was to provide actors with identical scenarios. Providing identical scenarios ensures consistency during recording.

To reduce the actors' movements when in front of the camera, they were asked to act out emotions while sitting in a chair, which was placed in the same location for all subjects. The intensity level was also recorded separately for each actor. Low, medium, and high intensity scales were recorded for each emotional category for each subject.

Lighting was set by the use of two professional 20cm reflectors equipped with a 200W halogen lamp each. Lights are carefully placed in an attempt to seem as natural as possible. The background consisted of a light blue panel set at a particular distance from each actor. This approach is vastly different from that chosen by the CMU database. The content of the recorded images in this database is carefully controlled (in contrast to the FG-Net), whose images consist of subjects with facial occlusions and head movement. A sample of images from this database are provided in Figure 3.4.



Figure 3.4: Some of the expressions available in the DaFEx database.

The database was evaluated by human judges who were given 10 seconds per video to rate the emotion given the expression. Results indicate that "fear" had the lowest success rate of 69.2%, with "neutral" scoring the highest with 87.9%.

3.2 Data Collection Methods

The data collection part of the process attempts to find the best way to identify which aspects of expression provides the best source of information, and then create a mechanism by which to code this information for feature extraction. Over the years data collection methods have experienced many problems. Still images provide the easiest platform for identifying and collecting data as those areas best representing expression are static. Collecting data from video sequences introduces problems such as face detection and tracking. Several different types of face tracking systems exist and are used to track faces in video, as this is necessary since head movement often occurs as a consequence of expression. The face tracking system proposed by Tao and Huang is known as the Piecewise Bèzier Deformation tracker (PBVD) [21].

This face tracker uses a model-based approach where an explicit wire frame model of the face is constructed. The model changes with the face from one frame to the next, and the points are auto-located within the image. Lighting affects these types of systems as edge detection plays a role in locating points of interest. If the lighting changes then the tracker will have a difficult time trying to find the appropriate points. Another system known as "Kalman filters" was used by Zhan to predict landmark positions in successive frames, which also corrects the localization results in the current frame [50]. The Kalman filter is a recursive procedure consisting of two stages: prediction and correction. During each iteration the system provides an estimate of the current state, and produces an estimate of the future state using a state model. Taking into account movement, the filter is used to predict the location of the landmarked points in the upcoming frame. This types of face tracking is required in order for a system to be used in real-time emotion recognition. Without the ability to track the face, a system would be unable to locate and extract expression points.

Feature location in an image is one of the principle responsibilities of the data collection process. Recent studies have used both static images as well as motion dependent. Motion dependent is a fully dynamic approach using several frames of video sequences of facial expression, typically lasting 0.5 to 4 seconds. This technique has evolved to include data analysis in the form of optical flow energy estimation and templates, as well as region tracking and deformation estimation.

In addition to locating the face in an image, points within the face that best represent change in expression must be identified. FACS is commonly used to auto detect and locate features without manual pre-coding of expression points. Template-based methods such as the those commonly used in conjunction with the active appearance models, identify expression points in an supervised way by manually selecting those points of interest. Karpouzis studies show that the eyes proved to be the most valuable and dependable source of expression information [23].

Data collection of facial expressions is a common problem in image processing. Low resolution, poor lighting, and the consideration of color versus gray-scale commonly plague this process. Lyons mentions the need for coloration in images as it has been shown that facial coloration is a key element in interpreting emotion [26]. As seen from the FG-Net database facial coloration plays an important part in interpreting emotions. Humans naturally use multiple sources when interpreting emotion. Not only does the movement of the face provide a good source of information, but the color also holds valuable information. Facial coloration due to embarrassment, sadness, and anger all provide valuable information. For example, when people get angry their face typically turns a shade of red, which is a queue often used in human communication.

Other data collection problems such as facial occlusion commonly plague facial feature data collection methods. Several studies report the need to prune out images and sometimes even entire subjects because of this problem. Facial occlusions such as eyeglasses, facial hair,

and feature distortions (damage to certain areas of the face), have caused several studies to simply filter out these images in pre-screening. This removes the burdon from the extraction process. More recently there has been a moving trend to purposefully incorporate such images under the idea that an effective data collection and extraction system would need to have the ability to deal with such problems. Cohn and Kanade suggest that the filtration of subjects should not occur before feature extraction, but instead should be made to deal with by the data collection and extraction process [5]. Cohn and Kanade also recognize that there are not only individual differences in appearance, but also in expressiveness, referring to the degree of plasticity, morphology, and frequency of intense expression, as well as overall rate of expression.

Another issue facing data collection methods are the variances of facial texture given from people of different ages and ethnic backgrounds. Infants have smoother, less textured skin, and often lack facial hair in the brows or scalp. Ethnic background complicates this as seen in the structural differences in the face. The contrast between the iris and sclera is different in Asians as compared to Northern Europeans. Lyons states that fear has been proven to be a problematic expression for Japanese subjects [26], and by excluding "fear" the classification improved. Ekman pointed out that basic emotions are expressed in relatively the same way, but cultural and structural differences may affect the overall expression.

FACS coding provides a way of automatically locating features of importance in an image, and has been used by several studies for the auto-location of expression points [15] [23] [49] [47]. Das claims that Gabor filters, which are based on Gaussian transfer functions, provide a more efficient way to encode natural images [7]. According to his work, natural images demonstrate that they have an amplitude spectra that fall off at approximately 1/f. He proposed that functions having the extended tails should be able to encode natural images more efficiently by better representing the higher frequency components.

Gabor filters have received considerable attention because of the fact that characteristics of certain cells in the visual cortex of some mammals can be approximated by these filters. They are also unaffected by illumination changes and noise. Ramanathan indicates that FACS and Fisher spaces have been used for face recognition by using class specific linear projection and have found to be insensitive to large variations in illumination and facial expressions [35]. Since illumination is a key factor is auto-detection of expression points, Gabor filters presents a viable alternative for data collection.

Features can also be identified in a supervised way by having the coder manually choose the landmarks before extraction. The AAM allows the coder to choose landmarked points for each image. Obaid uses the AAM for model creation and manually coded their models with 37 facial feature points, which include the corners of the mouth, eye lids and brow as well as the cheeks [34]. One common problem with this approach is the potential for not choosing the correct points which best describe expression. Another problem is accidentally coding the wrong points, which may have severe consequences to the classification as well as difficult to backtrack.

3.3 Feature Extraction Methods

Feature extraction refers to the problem of matching features extracted in either video or still images. The choice of the feature depends on a variety of factors and is often problem specific. A template-based method is preferred in some instances where the overall object needs to be modeled for shape and texture variation. Illumination changes and complex movements within an object tend to lead to the choice of a feature-based approach.

Feature extraction is used to obtain the parameters needed for classification. Once the facial expression points have been identified and coded, the feature extraction method will extract the needed information. Feature extraction in video must be fast so that processing can be performed in real-time. Often this phase of the classification system involves much computation as parameters are extracted, and is often reduced to simplify the representation and computation.

Studies have categorized feature extraction into two main categories: template-based methods and feature-based methods [23][50][21]. Template-based methods view the face as a single unit and extract features as a whole, whereas feature-based methods extract features from specific areas of the face, such as the mouth and eyes. It has been shown that a combination of methods is effective in facial expression recognition [25]. Lee uses a combined approach where template matching is used to locate the face and then optical flow used on individual regions. In his study, once the face is located it is divided into two distinct parts: one containing the mouth, and the other the nose and eyes. Once template-matching and feature reduction through PCA have occurred, an optical flow method is used to track the features within each region.

3.3.1 Template-based Methods

These types of methods are less concerned with the individual aspects of movement that correspond to a particular region, and are more concerned with the face as a whole.

Template prototypes are commonly used to extract feature information. These methods use correlation as a basic tool for comparing the template with the part of the image to be identified. The idea of the appearance or template-based method is to project face images, or sub-images into a low-dimensional feature space.

Eigenfaces and AAMs make use of PCA to extract features. Any face can be represented by using a linear-combination of eigenfaces. PCA is a common template-based technique [21]. The parameters collected by PCA operate on the entire face rather than individual facial features. Another algorithm commonly used for the template-based method is the k-nearest-neighbor (k-NN). One common problem among feature extraction methods is how to handle low resolution images. Karpouzis found that template-matching algorithms tend to work successfully in low resolution, and that templates of size 36x36 proved sufficient for his study [23].

Cohn refers to the distinction between the template and feature based methods as sign and measure judgement [5]. The sign based approach views the face as a whole. These methods common use a set of pre-defined emotional categories for classification. Ekman's basic six emotional categories are commonly used. Using emotional categories simplifies the training process as each face can be categorized into one of these basic six categories.

Template based methods are not without disadvantages. The "masking" smile, for example, introduces problems for correlation based approaches, and is commonly misclassified as sadness as the corners of the mouth turn down. It is only with a combination of movements such as the raising of the eyebrows in conjunction with the turned-down mouth corners that an appropriate emotion may be discerned. Feature-based methods would be able to code the movement of particular areas of the face to define this expression. In addition, template-based methods need to allocate large amounts of memory and cannot handle face with different poses well. [21].

Hua divides the face up into regions where each region is given a weight corresponding to their emotional relevance [20] . Each of the three sections of the face (divided horizontally) are used in the training. 2DPCA extracts the information and a similarity measure is used between the training images, which is used in classification of unseen facial images. Shang and his colleagues use AAMs for feature location and tracking. Their shape model consists of 58 facial points from Ekman's six basic expressions. Landmark locations are chosen based on FACS information. Feature reduction resulted in a 52-dimensional feature vector, which was then used an input into a Hidden Markov Model (HMM) for real-time expression recognition. [40].

3.3.2 Feature-based Methods

Feature-based methods use a modular approach to feature extraction. By examining specific areas rather than the face as a whole decisions can be made regarding the expression. Template based methods are concerned only with the overall shape and appearance of the model rather than specific features. One advantage this method has over template-based methods is the fact that feature-based methods have a smaller memory requirement and a higher recognition speed [50].

FACS is one of the main proponents of this technique. Action Units encode certain movements with those muscles responsible for creating the expression. Using FACS and viewing video-recorded facial behavior at frame rate and slow motion, coders manually code nearly all possible facial expressions. Action units may occur singly or in combinations. Better decisions can be made regarding the variation in expressions by combining action units. Making combinations additive yields a more accurate result and can even completely change the overall classification. Michel and Kaliouby make use of FACS for their feature tracking and extraction process [31]. In their system, 22 facial features are located and tracked in their chosen model. Using a Euclidean distance measure for point movement in successive frames, feature points are extracted.

Feature based methods makes use of feature specific knowledge such as the location of the eyes and mouth, as well as textural features such as wrinkles, bulges, and furrows. Examples of such feature-based methods can be seen in the work performed by Shang [40], who used a shape model defined by 58 facial landmarks. Valenti used FACS for feature identification and extraction. Their system tracks 12 facial motion units in the following categories: vertical movement of the lips, horizontal movement of the mouth corners, vertical movement of the mouth corners, vertical movement of the eyebrows, lifting of the cheeks, and blinking of the eyes [44].

Other studies such as Ramanathan [35] and Rose [37] uses Gabor filters for feature extraction. Using different frequencies and orientations these wavelets are useful for extracting different pieces of information from an image. Wavelets are commonly used in feature based extraction [16][50]. Juns method employs a wavelet energy distribution [16]. Wavelet energy is extracted from the regions where facial expressions reside. A similar study was conducted by Zhan who utilized Gabor filters for his recognition study. According to Zhan, Gabor filters are most discriminative for facial expressions and provide robustness against various types of noise. By applying these filters only to a set of facial landmark positions, rather than the whole face, the computational cost is lowered, as well as the sensitivity to illumination variations.

Other types of feature specific methods include the study performed by Rizon [36], where a genetic algorithm was used to find optimal ellipse fitting equations for fitting the eyes and mouth. Once the eyes and mouth have been properly fitted, the parameters were fed into a neural network for classification. Huang uses an optical flow method for feature tracking, which is a crucial part of the extraction process [21]. Optical flow is a commonly used method for tracking movement of feature points by finding all the pixels in an image. The process of tracking in sequential frames occurs in a three step process: a) detect the feature points in the first image according to the face model b) match the corresponding points in

another image using the optical flow parameters obtained. c) calculate the displacement vectors of all the feature points (Euclidean distance).

3.4 Classification Techniques

Several classification schemes have been used thus far, including Support Vector Machines [8][35][31][50][25][21][7][19], fuzzy-logic systems [32], Neural Networks [41] [15] [16], Linear Discriminant Analysis [17] [24], Gaussian Mixture Models [47] [29], Distance distance measures [50] [34], Genetic Algorithms [36], Bayesian Networks [44] [4], and HMMs for expression recognition using video [40].

For instance, Eckschlager used an Artificial Neural Network known as NEmESys (Neural Emotion Eliciting System) to identify a user's emotional state based on certain pre-defined criteria [10]. This system attempts to predict the emotional state of a user by obtaining certain knowledge about things that commonly cause changes in behavior. By giving the computer prior information such as eating habits, stress levels, and sleep habits, the ANN predicts the emotional state of the user and can change its responses accordingly. One problem with this approach is the requirement made by the system for the user to fill out a questionnaire providing the system with the needed information. While this system is unique, it does not incorporate any interpretation of facial expression, which has been identified as one of the key sources of emotional content [14] [18] [6].

Support Vector Machine (SVM) is a common approach in classification and has been used in applications ranging from online game avatars to emotion care service systems [25] [50]. Zhan, Li, Safaei, and Ogunbona proposed a system for online gaming which uses facial recognition of features to control emotional states in avatars [50]. Zhan originally created a system for real-time expression recognition, which would also control the emotional states of game avatars. The result of this study found that the system failed to recognize players expressions accurately when the spatial resolution of input face regions is lower than 100 x 100. Low resolution in images has been a problem among many emotion recognition systems, and continues to be a problem for image processing systems in general. AAMs for data collection and modeling are criticized for this very reason. As the image resolution

drops the entire system becomes unable to resolve the discrepancies.

In response to this problem Zhan and his colleagues attempted to improve upon their existing system. Using Ekman's basic six emotional categories - happy, sad, surprise, fear, disgust, and anger, a Gabor filtering and SVM classification approach was proposed. Using the JAFFE and FG-Net databases, Kalman filters are applied to each labeled image (used in tracking landmark locations across several frames). Once the features have been identified and extracted they are fed into a set of SVMs for classification. Rather than employing a single SVM, this study uses a cascade approach where each classifier improves upon the previous, resulting in a better overall classification. The overall recognition resulted in the lowest success rate reported as "sadness" at 52 percent, with the highest being "happiness' at 85 percent. Interesting enough these results are consistent across several studies, who also report "happiness" and "surprise" as being the emotion most recognized, and having either "sadness" or "fear" reported as being the least [31][50][34].

Michel and Rana Kaliouby used an SVM classification scheme for their study [31]. Using their own database, an automatic facial feature tracking system was used. A total of 22 landmark positions were calculated using the FACS system as a guide. The tracker employed collected the features from the training images using a displacement algorithm. Once collected a SVM classifier was used for the classification of the expressions into Ekman's six basic emotions. Michel states that for the SVM the kernel choice was the important customization that was made during the study. Experimenting with a range of polynomial, Gaussian radial basis function, and sigmoid kernels, the end result proved that the radial basis function (RBF) outperformed the others, boosting the recognition rate on the still images to 87.9%. The confusion matrix provided by their results from video indicated once again that "happiness" performed the best with a recognition rate of 91.7%, with "sadness" coming in last at 62.5%.

Lee, Chun, and Park reported the use of SVMs for an Emotion Care Service System [25].Their system infers emotions by recognizing facial expressions for input video in real-time. PCA and optical flow were used for feature extraction. Once extracted they are fed into a SVM for classification. The support vector machine used by Huang used a radial basis function for its kernel selection, which is known to be well suited for problems which deal with nonlinear relationships [21]. In their study optical flow is used for feature separation, which is then used as input into a SVM. The SVM is used to classify the feature point information such as location, distance, and angle, whose results indicate an overall accuracy of 81.5% using Ekman's six basic emotions.

Das, Horlings, and Ramanathan also utilize a SVM classification scheme for their study. Horlings takes a unique perspective on recognizing expressions by studying the EEG signals of brainwave patterns for the emotional states of "valence" and "arousal" [19]. Even though, most research studies subscribe to the effectiveness of Paul Ekman's six basic emotion categories, the field of psychology debates whether or not these are indeed modular as Ekman has suggested. Many in the field of psychology states that the emotional categories can only be categorized by two dimensions which include "valence" and "arousal".

For data collection, Horlings and his colleagues use video recordings. The subjects watch the emotionally driven recordings, which is designed to evoke the needed emotion. Rather than examining the expression, Horlings uses the signals received by the EEG readings to infer the emotion. Results indicate a classification rate of 32% for valence and 37% for arousal. This study shows how SVMs are a general purpose tool that can be used in a variety of circumstances.

Another study explored the use of a linear SVM for expression recognition. Das and his colleagues report success rates of 88% [7]. Using their own dataset along with Ekman's six basic emotions, a SVM classifier is trained on the parameters provided from the mouth, eye, and eyebrow regions. Restricting their attention to only those areas required the use of a

feature-based data collection mechanism. Attention was given specifically to the collection of data from the mouth-opening, eye-opening, and eyebrow constriction. Using MATLAB's statistical library, feature parameters were extracted and fed into "svmtrain" for training the classifier. The function "svmclassify" was used to classify the observation using the trained classifier. A linear function was chosen for the SVM kernel. Das indicates that the reason for this selection was driven by the fact that nonlinear operations yield results with lesser accuracy, as compared to linear kernel functions. Results indicate that using a custom built dataset with a linear SVM classifier yielded an overall success rate of 90.14%. Emotional categories "disgust" and "surprise" performed the best with success rates of 100% for the given test. "sadness" performed the worse at 77.71%. Other results indicate that removing the mouth opening from the classification decreased the recognition by 8%. Removing both mouth opening and eye opening reduced overall recognition by 13%, and removing all three reduced the overall rate by 55%. From this information it's clear to see how important these three features are in expression recognition.

Ramanathan uses a RBF kernel function to train a SVM classifier. The JAFFE database was chosen for the source images in this study, and Ekman's six basic categories used for classification. The open source tool "LibSVM" was chosen to create the SVM model. Overall results from the classification indicate that "anger" was the least recognized, having a success rate of 80.33%, with "neutral" performing the best at 84%.

In contrast to the SVM approach, many studies have chosen more of an unsupervised method for classification. Feng-jun, Sun, and Fasel all use a Neural network scheme for expression recognition [16] [41] [15].

Feng-jun and his colleagues proposed a system using wavelet energy distributions for feature extraction and a Neural network ensemble for classification. Using the CMU-Pittsburgh FACS coded database provided by Cohn [5], features are classified into one of the six basic emotion categories, described by Ekman. Wavelet energy distribution is a

form of feature-based extraction, as opposed to it's template-based counterpart. Once the data is extracted it is fed into a neural network ensemble for classification. Jun states that the choice for the neural network ensemble was driven by the fact that they have shown good generalization when the neural network results were combined [16]. Results from the classification report the best average recognition rate for the ensemble was 75.9%.

Sun and his colleagues reported an overall success rate of 97%, for their system [41]. Sun claims the CMU-Pittsburgh and JAFFE do not represent authentic facial expressions (since they are all created by actors). Sun and his colleagues expand this database by adding new video, most of which are recorded by news reporters in China. Results from their study indicate that "joy" performed the lowest at 95.33%, with "surprise" having the highest recognition of 99.3%. It is important to note though that "sadness", "anger", and "fear" were not part of this study. These particular emotions have shown low recognition rates as reported by previous approaches.

Rizon uses a combination of a Genetic Algorithm and a Neural Network for feature extraction and classification. The Genetic Algorithm (GA) is used to identify and extract the features. The parameters are then passed to a neural network for classification. The GA operates by using a fitting equation to find the location of three features: the top lip, bottom lip, and eye. These features are identified through successive generations of training the genetic algorithm. The optimized value of the top lip, bottom lip, and eye are fed as input parameters to the network. A FFNN is used and classified into one of the Ekman's six basic emotion categories. Overall success rates were reported at 83.57%.

Zhan and Obaid use a distance measure as of the classification techniques for their study. This is a good first approach into a problem simply because the distance measure is understandable and can be used for a proof-of-concept for the study. Applying the technique to multiplayer online games, Zhan and his colleagues decide on using this as a first technique in addition to k-nearest-neighbor, and a decision tree classifier [50]. This study uses a

face tracking algorithm for feature detection and Gabor filters for extraction. Detecting landmark locations automatically presents problems for systems that cannot work in a supervised manner. Multiplayer online Games (MOG) requires a fast technique for finding and extracting landmarked points. Once the landmark points have been identified the classification occurs by passing the parameters into the three different types of classification schemes. Malanobis distance classifiers are of the simplest classification schemes in that they classify patterns based on a Malanobis distance. Euclidean distance operates by taking the direct distance from one location to the next. Malanobis extends this by taking into account the distribution of the data. The mean vector μ and covariance matrix Σ are passed in as parameters to the classifier. A more accurate measure can be obtained by taking into account the distribution of the data. Operating on the JAFFE database, classification achieved successful recognition of 83% across all of the six basic emotional categories. No information was given regarding the break-down of each individual category.

Using the Euclidean distance classifier, Obaid reported an overall success rate of 88.9%. The breakdown of the individual categories reports "fear" being the least recognized at 66.7% and "surprise" achieving 100% recognition. Another approach commonly used in classification is the use of Gaussian Mixture Models (GMM). Gaussians represent data distributions by locating the distribution mean and variance of the data. Metallinou, Lee, and Narayanan make use of GMMs for classification into one of the six basic emotional categories [29]. In their approach a GMM is trained for each of the emotional states that are examined. A total of 64 mixtures are used because it was found empirically that this number of mixtures achieves good performance for the training dataset. A Bayesian classifier was used on the GMMs to find the class with the highest probability of membership. Results indicate that "neutral" was the least recognized at 54.64%, with "happy" the most recognized at 71.97%.

Mufti uses a fuzzy rule-based system to match facial expressions [32]. Facial action points are used to locate features using a customized database. Overall results indicate that "joy" has the best success rate, with "disgust" resulting in the lowest recognition. "Sadness" was left out of the study, but no reasons were given as to the exclusion.

Several different approaches have been used to deal with expression recognition, with most of research done up to this point using SVMs for classification. The nature of the SVM makes it a good choice for the problem of expression identification. Facial expression data has been shown to be highly dispersed in "face space". SVMs have shown great success with such problems, so it is only logical to consider them for further research in this area.

PROPOSED SYSTEM

The idea of using Active Appearance Models for emotion recognition is a fairly new concept. The face holds most of the information regarding emotions, so its only logical to explore face modeling when approaching the problem of recognizing emotional expressions. The system being proposed makes use of AAMs for feature location and extraction, and then employs three distinct classification schemes to prove the hypotheses presented in this paper. In all three experiments I use a free publicly available database of faces[45]. This database consists of a series of still images depicting various emotions from Ekman's six basic emotional categories [14]. Ekman's six basic emotional categories are well supported in the research community, which is one of the main reasons his system was chosen for this study [26] [27] [50] [7] [25] [31] [47] [36].

The FG-Net database was chosen for two main reasons: a) it contains elicited expressions rather than those given by actors b) the availability of the database for use in similar studies.

Using the Active Appearance Model (AAM), expressions from the database of images are evaluated for emotional content and labeled appropriately. The AAM is trained using a set of still images from the previously mentioned database. Points representing those areas of the face having the most intensity are then labeled on each training image. Once all of the training images are labeled the AAM is constructed. Model parameters are then extracted and used to classify unseen expressions into their appropriate category of emotion.

In this study three types of classification techniques are proposed: a) Euclidean distance measure b) Gaussian Mixture Model c) Support Vector Machine. The first technique employs a simple Euclidean distance measure from the observation to the mean of each cluster of training faces. Once this is proven as a successful technique, the GMM will then be used to wrap the data into a model which should yield better results as the data

distribution is taken into account. The Support Vector Machine (SVM) will be used to see if an optimal hyperplane can be found for separating the classes into clusters. Results from each experiment will be compared for effectiveness.

4.1 Architecture

The basic architecture of the system consists of data collection, analysis and segmentation, and model creation. The data collection phase consists of locating the appropriate landmarks for the face. Data analysis and segmentation involve filtering out unwanted images from the dataset to be used in the training process. Model creation involves creating the model from the training data which is to be used as input into a classification scheme.

4.1.1 Data Collection

Using the FG-Net dataset, images are evaluated using a set of predefined landmark locations. The facial landmarks chosen for this experiment can be seen in Figure 4.1.



Figure 4.1: Landmarked location points used in training the AAM.

The choice of the landmark locations were decided based on the underlying muscles structure of the face. FACS help provide insight into which parts of the face correspond to certain emotional expressions [12]. This guided the landmark selection process with a total of 113 points. Key areas were chosen to capture the movement of the brow, eyes, mouth, and nasio-labial region as formed by the underlying muscles expected for expression of the face.

4.1.2 Data Analysis and Segmentation

This phase consists of taking the given database and prunning out those undesirable images, creating a clean set of images used for training the Active Appearance Model. Undesirable images are identified by those having occlusions of the face, emotion clarity, and overall sincerity. Occlusions such as eye glasses and facial hair present complications in model creation as well as classification. Subjects with such features are eliminated from the dataset.

The use of professional actors may fail to capture expression that is completely and accurately representational of underlying emotional content. There is likely a difference between artificially posed expressions and those based on true underlying emotion. One feature that suggests this is the lack of constriction in the orbicularis oculi during artificial smiles [5]. Fasel and Luettin recognized that the use of posed facial expressions tends to be exaggerated and easier to recognize as opposed to those found in spontaneous expressions [15]. Walhoff however, has developed and released a database constructed in an effort to elicit genuine responses [45]. Actual comparisons between results on specific databases has also been somewhat limited, and it would be useful to encourage comparisons of techniques on the same data sets. Figure 4.2 show samples of some of the faces from the FG-Net database used for training.

Subjects 7, 14, and 17 were removed from the data set due to facial occlusions, such as eyeglasses and hair as well as other inconsistencies. Overall, the database stills were evaluated, and each image given an overall score as shown in Table 4.1. A benchmark was set, which marked the minimum score required for inclusion in this initial experiment. Once the scoring process was complete, subjects with very low scores were omitted from the training and testing set for this initial experiment. Table 4.1 provides a sample of the table used to evaluate the images to be used in the study.



Figure 4.2: Sample of faces in the FG-Net database.

The three biggest contributing factors for inclusion into the training dataset involved the sincerity or genuineness of the expression, the clarity of the expression, and the amount of head movement in the sequence. The genuineness of the expression is one of the key factors that has been cited by other studies for the creation of an effective system [41]. If a system is to be developed for use in a real-world scenario it must be trained using real-world expressions. It would be difficult to create a system using actors, and then apply the scheme in a real-world situation where the movement is expressed in a different way.

The clarity of the expression was another contributing factor. In many cases the emotion being expressed was either not representative of the expected emotion, or there was simply not enough intensity in the expression to register as an emotion. Subject 17 was completely removed from the training and test sets due to a lack of expression. For this subject, "fear", "joy", and "sadness" were all given with the same expression.

A total of 15 subjects were used for the training and testing of the three experiments presented. 4 images are taken for each subject for each emotion. A total of 60 images were needed for each subject, but certain images were prunned from the database due to facial

Image evaluation			
Image	*Sincerity*	*Clarity*	*Movement*	Score
1	3 of10	3 of10	No	7.0
2	4 of10	7 of10	No	7.5
3	8 of10	4 of10	Yes	6.5
4	7 of10	9 of10	No	6.0
5	9 of10	5 of10	No	3.5
⋮	⋮	⋱	⋮	
500	9 of10	9 of10	No	9.5

Table 4.1: Images were evaluated for their use in the study.

occlusions and other related factors. Table 4.2 lists the total number of images collected for each emotion.

Emotion	Total Number
Fear	52
Joy	60
Surprise	59
Anger	36
Disgust	52
Sad	20
Neutral	60
Total	339

Table 4.2: Total number of images collected for the study.

4.1.3 Model Creation

During model creation texture and shape variation samples are taken from each chosen image. The AAM software creates parameters for the generated model. These parameters are then used an input into the three classification systems discussed. By taking shape and texture variation in each face in the training set, a single model is created. This model represents an average of all faces in the training set. Figure 4.3 shows a "happy" face being labeled during model creation.



Figure 4.3: Face being labeled during model creation.

Points are labeled given the landmark selection scheme. Using the built-in tools provided by the AAM software these points are manually linked, resulting in a type of wire frame deformable model. Each image in the training set is labeled this way. Face images in successive frames are able to use the points labeled in the previous image. By simply dragging the points, the wire frame moves and allows the user to label new faces by simply dragging the wire frame to the desired location to fit the new expression. This is useful for images who movement varies little from one frame to the next. It also helps to create a more consistent method for labeling faces. One problem with human intervention in creating models is the possibility of labeling points in the incorrect positions. The AAM software does provide functionality for the auto-labeling of predefined points, but this technique was not within the scope of my study.

Once all of the images have been labeled the model is then created. Model creation involves deciding upon the number of principal components you wish to use. For this study I chose 30 principal components (for more information regarding principal components see the background section of this paper). The AAM software then creates shape and texture models for the training data and combines these model to create a single model whose data is represented by the number of principal components chosen. Figure 4.4 shows the resulting model created from the FG-Net training data.

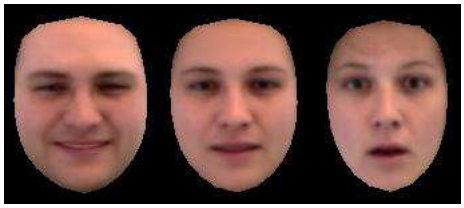

Figure 4.4: Model created from training data.

The two faces shown in figure 4.4 on either side represent variation from the mean within the model. The center face represents the average face created using the combinations of all images for all emotions. The variability of the data is adjusted by modifying certain parameters in the AAM to make sure that a high percentage of the data is represented by the model to deal with subtle changes in facial features. It should be noted that none of the images in this figure come from any one person, but instead are generated by the combined subjects used in the training data.

4.2 Classification

Several classification schemes were considered for this study. The three most commonly used types of classifiers were chosen based on previous work. As a first approach and proof-of-concept the Euclidean Distance measure was chosen. Euclidean distance is a common first classifier as it gives a basic distance measure in feature space which can be used as a benchmarking tool [31][34]. As a second scheme the Gaussian Mixture Model was chosen, which should produce better results than a simple distance measure due to the nature of the distribution. The Gaussian distribution works by fitting the data into a normal distribution where the mean of the data's distribution is taken into account. Euclidean distance measures have no knowledge of the data distribution. By having knowledge of the distribution better class assignments can be made.

The third approach taken in this study involved using a Support Vector Machine to separate classes for better classification. Support vector machines work by finding the best non-linear separation between datasets that may not have clear boundaries. Non-linear separation is achieved by choosing a mapping function that best separates the data classes. Once a kernel function is decided upon, the data is then passed through the function and projected into a higher dimension, which should improve class separation.

In general, classification occurs once a model has been created. Models are created using the data provided from the data analysis and segmentation phases of the study. Usually, multiple classifiers are used and then compared for effectiveness [36][19][50].

4.2.1 Euclidean Distance

The Euclidean distance classifier was chosen as a first approach due to the simplistic nature of the algorithm and its capability to produce a benchmark with which to test other more sophisticated techniques. Euclidean distance as it is used in classification measures the distance between datasets which represent what is being modeled. A distance measure

is taken from two distinct points located in different areas of n-dimensional space.

For n dimensions, where $P = (p_1, p_2, ..., p_n)$ and $Q = (q_1, q_2, ..., q_n)$ the distance is computed as

$$d(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} \tag{4.1}$$

Straight-line distance measure is one of the simplest classification techniques, and is commonly used as a first approach when beginning a new problem. The simplicity of the calculation provides insight into how well the data is separated. This measure establishes a benchmark for researchers. Malanobis distance, for example, takes distance checking a step further by measuring against the contours created by the variance of the data.

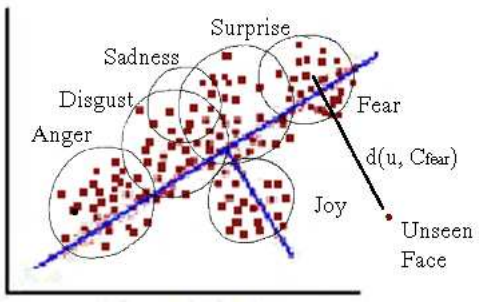

Figure 4.5: Conceptual diagram demonstrating euclidean distance measure.

Still images from fifteen different subjects were used for training and testing in this experiment. The implementation of this classifier was coded in Matlab. In order to make the extraction of the face image data easier to process I created a set of functions that would read in the needed data, which could be used by all three classification schemes.

I chose to use a leave-one-out approach for training the classifier. This is commonly used to improve testing methods with those studies having relatively few test subjects. Stills from each of the fifteen subjects were used for testing data after an AAM, and class parameter-vector means were found using the other subject stills as training data. The Euclidean distance measure operates on face data by taking a straight-line measurement from the mean of the test data to the mean for each cluster of data for a given emotion category. Table 4.3 contains the results from this classification.

Subject	Percentage Correct
Subject 1	80.7%
Subject 2	74.9%
Subject 3	89.5%
Subject 4	90.2%
Subject 5	88.1%
Subject 6	70.8%
Subject 8	80.7%
Subject 9	100%
Subject 10	60.4%
Subject 11	100%
Subject 12	57.2%
Subject 13	100%
Subject 15	76.2%
Subject 16	89.7%
Subject 18	100%
Total Average Correct	83.9%

Table 4.3: Euclidean classification results by subject.

An analysis of the results show that the system correctly classified anywhere between 60% and 100% for each individual, except for subject 12 whose fear expression was misclassified as "anger" for each expression. Most subjects were classified within the 80% to 90% range, but a few subjects showed poor recognition. It is difficult with these early results to say whether that is due to the subject's expression, feature method, or classification method. It may be that with more sophisticated classifiers we can better speculate as to why certain subjects performed worse that others. Since the Euclidean distance is a base-line classifier, the next two classification schemes should produce better results. A confusion matrix of the classification results is provided in Table 4.4.

	Anger	Disgust	Fear	Joy	Sad	Surprise	Neutral
Anger	0.59	0.06	0	0.03	0.28	0	0.04
Disgust	0.09	0.91	0	0	0	0	0
Fear	0	0	0.88	0	0	0.08	0.05
Joy	0	0.06	0	0.8	0	0	0.14
Sad	0.25	0.12	0.06	0	0.45	0	0.12
Surprise	0	0	0.13	0	0	0.84	0.03
Neutral	0	0	0	0	0.07	0	0.93

Table 4.4: Confusion matrix for Euclidean classification.

The confusion matrix shows "sad" and "anger" performing the worst. "Sad" was only correctly classified 44% of the time, with "anger" having a classification success rate of 59%. It is interesting to see that "sad" was most often misclassified as "anger" and vice versa. Both of these emotions are usually conveyed with similar movement of expressions, having the corners of the mouth turned down. With sad expressions the eyebrows usually turn up, rather than down, when expressing anger. Figure 4.6 shows two separate expressions being given by the same subject. The image on the left is expressing "anger" while the one

on the right is expressing "sadness". Even with the human eye it is difficult to distinguish between these two expressions.



Figure 4.6: The left image shows an expression of "anger", while the image to the right is expressing "sadness".

Perhaps the problem with these results were simply due to the lack of good training data. "Sadness" was an expression that seemed to be the most difficult to obtain from the dataset. Limited training data and as well as the quality of the expressions provide evidence into the possible cause of poor classification in some instances. Based on the scoring scheme mentioned earlier, an evaluation of the database also suggests that the subjects had difficulty expressing negative emotions. The overall success in this classification approach leaves room for future development.

4.2.2 Gaussian Mixture Model

As a next step in the study the GMM was chosen. By modeling the data in a Gaussian distribution I hoped to improve upon the Euclidean classification results. A Gaussian Mixture Model (GMM) is a probabilistic model for density estimation. A GMM is based on the concept of data distributions, using the variance and mean of the data sample. With the Gaussian distribution, intelligent decisions can be made having only a small training set. Data distributions are "fitted" with Gaussians (normal distributions) by generating

data samples centered around the mean and distribution of the data available, and the Gaussian density function describes the way a particular Gaussian is distributed. Gaussian distributions can take the form of a continuous univariate form where the mean and variance are scalar values, or can take the form of a multivariate normal distribution. Data drawn from facial expressions take the form of a multivariate distribution due to the variation in the way emotions are expressed. Knowing this, the multivariate normal distribution seemed a likely candidate for this experiment.

GMMs are often used in situations where the data distribution is unknown, which is the case in most pattern recognition problems. The multivariate normal distribution is the case that fits the study. In multivariate distributions, several types of distributions are present, each having it's own variance and mean. Classification uses this information to make a better decision regarding which class, or distribution, a particular sample belongs. The mean vector μ and the covariance matrix Σ describe this distribution. The general multivariate normal density in d dimensions is written as

$$p(x) = \frac{1}{2\pi^{d/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(\mathbf{x}-\mu)^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)) \tag{4.2}$$

where $\mathbf{x}$ is a d-component column vector, μ is the d-component mean vector, Σ is the d-by-d covariance matrix, and —Σ— and $\mathbf{\Sigma}^{-1}$ are its determinant and inverse respectively. For simplicity reasons equation 4.2 is often written as

$$p(x) \sim \mathcal{N}(\mu, \mathbf{\Sigma}) \tag{4.3}$$

.

The mean is a vector of means rather than a scaler as in the univariate case, and is represented as

$$\mu \equiv \varepsilon[\mathbf{x}] = \int \mathbf{x} p(x)\, dx \tag{4.4}$$

and the covariance represented as,

$$\mathbf{\Sigma} \equiv \varepsilon[(x-\mu)(x-\mu)\exp t] = \int (x-\mu)(x-\mu)^t p(\mathbf{x})\, d\mathbf{x} \tag{4.5}$$

where the expected value of a vector or a matrix is found by taking the expected value of its components. A vector in n-dimensional space would represent a point in n-dimensions, and would take the form of a vector with length n. The overall mean of the n-dimensional vector is represented as a vector of means. Samples from normal distributions tend to cluster about the mean and spread related to the standard deviation σ. Due to the difficulty in finding and collecting facial data, the Gaussian seemed to be a logical classification scheme since they tend to work well with sparse datasets. Gaussians are used for data clustering and mixture models can be used in situations where the data distribution is unknown and a distinct of number of classes are involved in creating the distribution. Figure 4.7 shows examples of four normal distributions each having a different mean and covariance. Classification takes the form of decision boundaries used to separate the data regions.

Decisions for class assignment as seen in figure 4.7 can be determined using a variety of methods. One such method, known as Maximum Likelihood Estimation, was chosen for this study. Maximum Likelihood Estimation (MLE) works by finding the sample with the greatest likelihood of generating the given sample. Class assignment occurs as a result of this determination.

For this study I decided to use a library provided originally by Professor Charles Bouman of Purdue University [2]. Given a set of multidimensional training vectors, the program

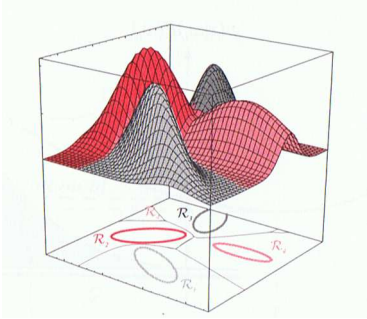

Figure 4.7: Conceptualization of four data regions along with decision boundaries for each region [9].

models the data as a Gaussian mixture distribution, estimates the order of the mixture by the minimum description length (MDL) criterion, and estimates the parameters of the Gaussian mixture by the expectation-maximization (EM) algorithm.

The training data for each class was passed as a parameter into a Matlab function *GaussianMixture()*, which created a mixture model of the data distribution for the given class. This was used for each of the 7 emotion classes, resulting in 7 mixture models being created. Once the models were created, the log likelihood for each of the classes was calculated. *GMClassLikelihood()* is then called for each of the classes. The function calculates the log likelihood of observation given a particular Gaussian mixture. The mixture containing the greatest likelihood is the one chosen for the observation. Using the "leave-one-out" method described in the previous experiment, subjects are rotated being used as a test set in one instance and part of the training set in another instance. Table 4.5 shows the results on a per subject basis.

Results report 87.1% correct classification across all subjects. This is an improvement over the 83.66% classification results seen in the Euclidean distance experiment. The confusion matrix for this classification can be seen in Table 4.6.

Subject	Percentage Correct
Subject 1	78.9%
Subject 2	81.5%
Subject 3	89.5%
Subject 4	90.2%
Subject 5	100%
Subject 6	80.6%
Subject 8	80.1%
Subject 9	100%
Subject 10	60.3%
Subject 11	100%
Subject 12	71.4%
Subject 13	100%
Subject 15	82.2%
Subject 16	93.1%
Subject 18	100%
Total Average Correct	87.1%

Table 4.5: GMM classification results by subject.

From the results list in table 4.6, the two emotions that reported the most misclassifications were "sadness" 40%, and "anger" 62%. This is consistent with results obtained from the Euclidean distance classification seen in Table 4.4. Subject 10 reported the lowest classification score of 60%, and subjects 5, 9, 11, 13, 18 reported the highest recognition of 100% correct classification of all images for each subject.

	Anger	Disgust	Fear	Joy	Sad	Surprise	Neutral
Anger	0.62	0.03	0.07	0	0.28	0	0
Disgust	0.08	0.92	0	0	0	0	0
Fear	0	0	0.94	0	0	0.06	0
Joy	0	0.05	0	0.93	0.02	0	0
Sad	0.20	0.10	0.20	0	0.40	0	0.10
Surprise	0.06	0	0.04	0	0	0.90	0
Neutral	0	0.03	0	0	0.03	0	0.94

Table 4.6: Confusion matrix for GMM MLE classification.

4.2.3 Support Vector Machine

Support Vector Machines (SVMs) have been used in many studies where data classes need discrimination, specifically in the realm of expression recognition where distribution patterns warrant separation [31] [35] [7]. SVMs are a set of statistically based supervised learning algorithms used for classification and regression, and are closely related to neural networks. The motivation behind the use of these algorithms is the need to separate data distributions into distinct classes.

SVMs rely on preprocessing the data to represent patterns in a high dimensional space - typically much higher than the original feature space [9]. A non-linear mapping function is used to project data into a higher dimension. The goal of the projection is to create a hyperplane that can be used to separate classes into distinct clusters. The vectors that reside between this separating hyperplane and the distribution is known as "support vectors". The training process occurs as a result of the need to find a separating hyperplane containing the largest margin for the given data. The hyperplane chosen must be equadistance from the determined support vectors.

A non-linear mapping function φ is used to project the data into a higher dimension. The training data, represented as $\mathbf{x}_k$ is transformed into $\mathbf{y}_k = \varphi(\mathbf{x}_k)$. For each of the n patterns, k=1,2,...,n let $z_k = \pm 1$, where k is the number of classes involved.

A linear discriminant in an augmented y space is

$$g(\mathbf{y}) = \mathbf{a}^t\mathbf{y} \tag{4.6}$$

A separating hyperplane ensures that

$$z_k g(\mathbf{y}_k) \geq 1, k = 1, ..., n \tag{4.7}$$

Using this information the discriminant described in equation 4.6 can be used to create a separation between the various classes k. It has been shown that the distance from any hyperplane to a (transformed) pattern $\mathbf{y}$ is $|g(\mathbf{y}|/||a||$, and assuming that a margin exists, the following equation can be used to help find the values needed for determining the optimal hyperplane:

$$\frac{z_t g(\mathbf{y}_k))}{||\mathbf{a}||} \geq b, k = 1, ..., n \tag{4.8}$$

From equation 4.8, the goal is now to find the value of $\mathbf{a}$ that maximizes b, which is the distance from the support vectors to the hyperplane described in equation 4.8. Figure 4.8 shows the optimal hyperplane for a given problem. The mapping function projects the data into a higher dimension, and the value chosen for $\mathbf{a}$ is such that b is optimal and equa-distance to the support vectors.

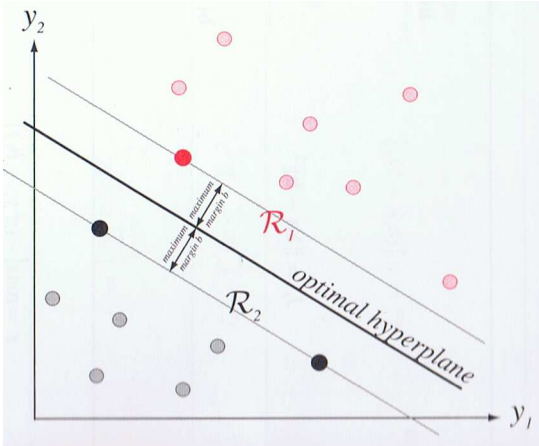

Figure 4.8: Conceptual image demonstrating an optimal hyperplane separating classes [9].

Training involves the choice of a kernel function. Kernel functions that are commonly used include: a) Polynomials b) Gaussians c) other radial basis functions. For more information regarding the types of basis functions available as well as an indepth discussion on the topic see the tutorial given by Christopher Burges [3].

In most real-world scenarios data clusters are not modularized, but instead overlap making classification difficult. Ideally, the best scenario would be for each cluster to be completely separated from all other clusters, creating distinct classes. Figure 4.9 demonstrates this type of scenario.

For this study the SVM library chosen is the statistical toolbox provided by Vojtech Franc and Vaclav Hlavac. This library was first introduced in February of 2000 at the Czech Technical University in Prague. The toolbox provides methods for Bayesian classification, linear discriminant analysis, and support vector machines. Several functions are available for SVM analysis and classification. A multi-class SVM trainer and classification system is also provided.

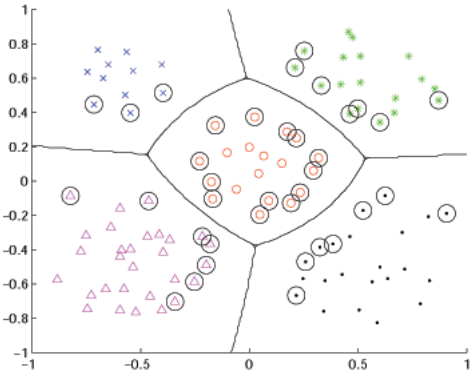

Figure 4.9: Conceptual image demonstrating an optimal region separation [9].

The training aspect of the experiment involved taking the existing model and test data from the previous experiments, and creating a model for each class. I created a set of scripts to read in the data, format according to the needs of the model, and pass it to the function for model creation. Several Matlab files are available to use for training and classification. For this study I chose to use "oaosvm.m" for training and model creation, and "mvsmclass.m" for classification. This function takes the model data as input and outputs a multi-class majority voting classifier structure. Using "mvsmclass.m", each of the models created was used as input into this function along with the observation. Within this function you can specifiy the kernel choice for the training. In this experiment I chose to use the Gaussian radial basis function. The classification function provided returns voting results for each model/observation combination. I found the voting output to be very useful in situations where the class assignment was incorrect and needed further investigation. Using the "leave-one-out" method outlined in the previous sections, each subject was rotated in, and classified accordingly. The results of this classification are available in Table 4.8 and Table 4.7.

From the results in Table 4.7 it can be shown that "sadness" performed the worst only successfully recognizing 10% of the observed images. "Anger" was most improved going

	Anger	Disgust	Fear	Joy	Sad	Surprise	Neutral
Anger	0.90	0	0.10	0	0	0	0
Disgust	0.20	0.72	.07	0	0	0	0.01
Fear	.09	0	0.73	0	0.01	0.13	0.03
Joy	0.06	0.01	0.06	0.86	0	0	0
Sad	0.45	0.20	0.35	0	0.10	0	0
Surprise	0.06	0.04	0.05	0.02	0	0.8	0.03
Neutral	0.13	0	0	0	0	0	0.87

Table 4.7: Confusion matrix for SVM classification.

from 62% in the GMM classification to 90%. The results on a per subject basis are displayed in Table 4.8.

Overall, the SVM performed well resulting in an overall success rate of 91.3% across all subjects. This is an improvement over GMM which reported an overall recognition of 87.1%. Subject 12 reported as being the least recognized. This was also true in the previous two classification schemes as well. Figure 4.6 shows an "anger" and "sad" image taken from subject 12. It can be seen that both images contain very similar features which may have attributed to the poor recognition results for this subject.

Subject	Percentage Correct
Subject 1	89.5%
Subject 2	93%
Subject 3	100%
Subject 4	100%
Subject 5	92.5%
Subject 6	95.8%
Subject 8	100%
Subject 9	100%
Subject 10	95%
Subject 11	100%
Subject 12	57.1%
Subject 13	89.5%
Subject 15	70.8%
Subject 16	86.2%
Subject 18	100%
Total Average Correct	91.3%

Table 4.8: SVM Classification results by subject.

DISCUSSION

Throughout this study several different types of emotion classification techniques have been presented, in addition to an extensive background review of previous studies. Classification schemes ranging from Neural Networks to Support Vector Machines have been used in an attempt to create an effective emotion recognition system.

One of the goals of this research was to determine the effectiveness of this approach in comparison to what has been presented thus far. In order to validate that this study was successful, a closer examination of the results reported by previous studies is required. Table 5.1 displays a list of recent studies in the area of emotion recognition using facial expressions. In each case the database, model, and classifier is reported, as well as the overall recognition results.

In the introduction of this paper two hypothesis were given. The first of which poses a question regarding the effectiveness of AAMs for the recognition of emotions using facial expressions. Table 5.1 lists recent studies performed in this area of research. The results given demonstrate the effectiveness of other techniques for recognizing emotional expressions. In order to validate the claim that AAMs are useful in this type of research, results from studies using AAMs should compare to or surpass those reported by previous studies. The three experiments conducted in this study resulted in an overall average recognition of rate of 87.4%. The Euclidean distance classifier reported an overall recognition of 83.9% which surpassed the results obtained from 6 of the 13 studies. The GMM classification scheme reported even better results outperforming 9 of the 13 studies. Finally, the SVM reported the best recogntion of 91.3% outperforming 11 of the 13 studies. Individually, each classification scheme reported results comparable to those given in the previous studies mentioned in table 5.1.

Study	*Database*	*Model*	*Classifier*	Overall
Cohen [4]	$Cohn-Kanade$	$Bayesian$	$BayesianNetwork$	83%
Yafei [41]	$Cohn-Kanade$	BVD	NN	97%
Wong [47]	$JAFFE$	$GaborFilters$	NN	86%
Ramanathan [35]	$JAFFE$	$GaborFilters$	SVM	83%
Feng-Jun [16]	$Cohn-Kanade$	$Wavelets$	NN	76%
Zhan [50]	$FG-Net$	$GaborFilters$	SVM	72%
Michel [31]	$Cohn-Kanade$	$Facetemplate$	SVM	86%
Zhan [50]	$JAFFE$	$GaborFilters$	$MalanobisDistance$	76%
This Study	$FG-Net$	AAM	$EuclideanDistance$	83.9%
This Study	$FG-Net$	AAM	GMM	87.1%
This Study	$FG-Net$	AAM	SVM	91.3%

Table 5.1: Comparison of results from previous studies.

Once I had completed the three experiments I began to record my results. During this time I discovered two other studies that were using AAMs for their emotion recognition research. Results from those studies in addition to the results reported from the experiments in this study can be seen in table 5.2.

Both Obaid [34] and Datcu [8] conducted similar studies at about the same time that this work was being conducted. Obaid employed an Euclidean distance classifier, similar to the classification scheme performed for the first experiment in this study. His results indicate an overall recognition rate of 89%. AAMs were used as a method of facial feature tracking (AAM fitting). This technique works by finding the model parameters which best fit an AAM to an image. FACS was used as a guide for determining the best location for modeling facial behavior. Using FACS as a guide, 37 Facial Action Points (FAPs) were

Study	*Database*	*Model*	*Classifier*	Overall
Obaid [34]	$Cohn-Kanade$	AAM	$EuclideanDistance$	89%
Datcu [8]	$Cohn-Kanade$	AAM	SVM	80%
This Study	$FG-Net$	AAM	$EuclideanDistance$	83.9%
This Study	$FG-Net$	AAM	GMM	87.1%
This Study	$FG-Net$	AAM	SVM	91.3%

Table 5.2: Comparison of results from previous studies.

chosen. Sixteen different regions of the face were chosen for representing facial deformation, and a rubber-sheet transformation was applied to determine the expression deformation parameters. Key facial features used include: the eyes, eyebrows, cheeks, and mouth. Once the deformation parameters were extracted a deformation-table was constructed for each of the six emotion categories defined by Ekman [14]. A Euclidean distance measure was then used across all FAPs to determine the expression which best matches a certain known expression.

The results obtained from the Euclidean distance from this study are comparable to those obtained from the study performed by Obaid. The results reported by Obaid as well as this study are comparable to what has been reported thus far in the research community (as seen in table 5.1).

Datcu and Rothkrantz utilized AAMs for model creation and tracking in their study using both still images and video. Using the Cohn-Kanade database AAMs were used as a method of feature extraction for shape and texture variation of the chosen images. AAMs consisted of 19 shape modes, 24 texture modes, and 22 appearance modes. A total of 17 facial parameters were collected from each face which include 6 parameters from the eyebrows, 6 from the eyes, and 5 from the mouth. A two-fold cross validation method was used for testing the images. Parameters were then fed into a SVM classifier which chose the

appropriate emotion based on the given parameters for the test set. Table 5.2 indicates an overall recognition rate of 80% for their study, which is significantly less than that reported by the SVM experiment in this paper. Overall, the AAMs chosen for this study resulted in comparable results to previous studies and even exceed results in some instances.

One possible reason for the success reported by Obaid using the Euclidean distance measure over the successes reported by this study may be attributed to the differences in the size of the training set. For example, Obaid collected a total of 432 images for "sadness", which is significantly more than the 20 chosen for this study. Table 5.3 lists the number of samples collected by each study for each emotion.

Emotion	Obaid [34]	Datcu [8]	This Study
Joy	532	107	52
Sad	423	92	60
Anger	448	30	20
Fear	346	84	36
Disgust	296	56	52
Surprise	387	105	60

Table 5.3: Comparison for number chosen samples

As shown in table 5.3, both Obaid and Datcu collected more images for each emotion. Typically the larger the dataset the more accurate the results. The small size of the dataset in this study may have contributed to the poor recognition reported by certain emotions, which directly impacts the second hypothesis posed in this paper.

The second hypothesis posed the question of whether certain expressions are more distinguishable than others. According to the results from this study, "sadness" was the primary expression that ranked lowest in overall recognition. This was a common theme found in

all three experiments. In order to better understand the reasoning behind this, a comparative analysis to previous studies has been provided. Rankings for recognition by previous studies can be seen in table 5.4.

Study	Fear	Joy	Surprise	Anger	Disgust	Sad	Neutral
Shang [40]	75	95.8	100	87.5	91.7	100	-
Rose [37]	84.4	83.9	80	96.7	82.8	83.9	93.3
Obaid [34]	66.7	91.7	100	93.3	83.3	96.7	-
Michel [31]	66.7	91.7	83.3	66.7	64.3	62.5	-
Zhan [50]	74	85	82	69	63	52	80
Das [7]	98.7	99.4	100	100	100	77.7	89.7
Feng-Jun [16]	75.2	77.7	81.8	76.4	74.2	70.2	-
Ramanathan [35]	82	80.9	-	80.3	81.5	83	84
Wong [47]	60	100	100	70	100	80	100
Hua [20]	85.7	100	100	95.2	95.2	90.4	-
Overall	76.8	90.6	80.8	83.5	83.6	79.6	-

Table 5.4: Recognition results reported by previous studies for Ekman's six emotions.

The results from the previous studies provided in table 5.4 show that "fear" was the emotion least recognized followed closely by "sad". "Joy" performed the best and was the most recognized on average. Between this study and results given by previous studies, "fear" and "sad" reported the lowest recognition rates. Because of this fact I chose to examine the results from these emotions in closer detail. In order to better understand why "fear" and "sad" reported such low recognition rates a confidence interval analysis is performed. Tables 5.5, 5.7, and 5.6 have been provided to aid in this analysis. The CI for "joy" is included in this analysis since it was the emotion reported as having the highest

recognition in this study.

	90% CI	95% CI	99% CI
Euclidean Distance	.2420 - .5818	.2183 - .6140	.1782 - .6716
GMM	.5104 - .7778	.4873 - .7967	.4293 - .8296
SVM	.0252 - 0.2701	0157 - .3132	.0001 - .3992

Table 5.5: Confidence Interval for correct classification of "sad".

	90% CI	95% CI	99% CI
Euclidean Distance	.8566 - .9727	.8361 - .9784	.7915 - .9888
GMM	.8566 - .9727	.8361 - .9784	.7915 - .9888
SVM	.7768 - 9248	.7557 - .9335	.7117 - .9486

Table 5.6: Confidence Interval for correct classification of "joy".

	90% CI	95% CI	99% CI
Euclidean Distance	.8361 - .9682	.8132 - .9747	.7640 - .9864
GMM	.8602 - .9806	.8375 - .9362	.7880 - .9966
SVM	.8602 - .9806	.8375 - .9362	.7880 - .9966

Table 5.7: Confidence Interval for correct classification of "fear".

The confidence intervals provided in table 5.5 show a 90% CI for "sad" being recognized only 13% to 58% of the time. "Fear" reported a 90% CI between 84% and 96%. The range for successful recognition for "fear" was somewhat higher for this study in comparison to other studies. "Joy" reported a 90% CI between 78% and 97%. Previous studies report "joy" being one of the emotions that was most recognized resulting in an overall rate of

90.6% on average. "Sad" was only recognized on average 79.6% of the time. Table 5.8 reports a confidence interval for each of the six emotional categories gathered from this study.

	90% CI
Fear	.8602 - .9806%
Joy	.8566 - .9727%
Surprise	.8130 - .9487%
Anger	.5104 - .7778%
Disgust	.8361 - .9682%
Sad	.2420 - .5818%
Neutral	.8566 - .9270%

Table 5.8: Confidence Intervals across all emotions for the GMM classification scheme.

The CI provided in table 5.8 shows that the 90% CI for "sad" falls below the CI ranges for the other emotions. However, in order to make the conclusion that "sad" is more difficult to recognize that other emotions a more thorough investigation is warranted.

As mentioned earlier, one possible reason for the poor performance of "sad" may be attributed to the lack of good "sad" expressions. Only 6 of the 15 subjects used in this study contained quality "sad" expressions capable of being used for the purposes of this study, as opposed to 15 of the 15 subjects for the "joy" expression. As mentioned earlier, actors are commonly used to generate expressions that cannot be easily evoked, and an analysis of the FG-Net database revealed a lack of good "sad" expressions.

In summary, the results from this study provide insight into the effectiveness of using AAMs for emotion recognition. These results show promise, indicating that AAMs should be further investigated. The idea that certain emotions are more easily identified than

others is still up for debate. Initially I had thought that "sadness" was more difficult to recognize, but upon inspection of the studies performed by Datcu [8] and Obaid [34] this may not be the case. More investigation is needed in order to determine if this is indeed the case that certain expressions are more easily recognized than others.

CONCLUSIONS

During this study several types of methods for data collection, feature extraction, and classification were reviewed as well as a study of the currently available facial expression databases. From this review and the results shown in this experiment I have been able to formulate some possible improvements that may result in a better recognition system.

First there is the question of using 3d models over 2d models. Studies have shown that 3d models are advantageous over 2d models [48]. People rarely move their head in 2-dimensions. Most of the time people must determine an expression based on limited information. In order to create a system capable of operating close to a human capability it must be able to infer expressions similar to the way a human does in human-human communication. 3d models in my opinion would be advantageous over 2d models in that they could create a system capable of recognition without the need for a full frontal pose of the face. However, one disadvantage to using the 3d model is the data representation. The study proposed by Yin, while advantageous due to the amount of detail, resulted in the creation a 84-dimensional feature vector. Processing for such large feature spaces takes valuable time and memory that would hamper real-time systems. A 3d model would be advantageous only if a way were found to better represent the data for computational purposes.

The second improvement would be to move away from the idea of using the entire face for expression recognition, but instead adopt a FACS based system where an individual feature, or a combination of features, are analyzed. FACS coding has the advantage of being able to ignore those parts of the face which doesn't provide information for expression differentiation. This also reduces overhead in processing as only those features that matter are processed, thus reducing "noise" in the model produced.

Another improvement needs to be made in the area of facial expression databases. There seem to be a wide variety of fairly incomplete databases. The database provided by FACS AU-coded CMU database provided by Cohn and Kanade seems to be the most advanced database currently available. Most databases examined in this study specifically created images under a constrained setting, with lighting, background, and head movement under control. The approach that Cohn and Kanade took when creating the CMU database was to create a database which would allow for the development of a complete system capable of handling pose, lighting and other common problems. When humans recognize expressions it is usually in an open environment where nothing is controlled. To train a system to behave in the same way, a database which represents the real-world must be used. Most recent databases have not taken into account age, size, ethnic backgrounds, and facial deformations. All of these things exist in the real-world and as such should be modeled in an expression database. Head movement should be encouraged rather than controlled. Images containing facial occlusions such as hair and glasses should not be pruned from the system, but integrated into the system in such a way as to be representative of a real-world scenario.

Head pose changes and posture form important cues which work together with facial expressions to provide more emotional information. People rarely express emotions without head movement. When you think of being surprised, the first thought that surfaces is an image of someone jumping from just being startled. Without the addition of robust facial feature tracking algorithms the feature becomes lost in the noise of the motion. Even the jump itself provides valuable information regarding the emotion which is about to be conveyed.

Previous studies have shown the effectiveness of combining multiple databases to introduce variability in age, gender, and ethnicity. Looking ahead being able to successfully combine data sources may actually be advantageous, as opposed to having all sources in

one location. For example, if a study could pull FACS coded face images from multiple sources and specify the ethnicity, age and gender needs for the study, then a database could be custom-built for the purposes of that study.

Ekman and Freisen introduced the six basic emotions as a way to categorize emotions, which could be used for emotion recognition. Ekman and Friesen's claim of the "basic" emotions: anger, disgust, fear, happiness, sadness and surprise has come under much debate recently. There has been much debate from several psychologists as to the validity of these categories. James Russell and colleagues have recently challenged the classic data and argues that emotion in general can best be characterized in terms of a multi-dimensional affect space, rather than discrete categories, such as suggested by Ekman. According to Russell's view, two dimensions of "pleasure" and "valence" are sufficient to characterize facial affect space [39]. Russell calls for new research on perception of facial affect using improved methods and multi-dimensional analysis. During this study I found a few studies which makes use of Russell's two dimensions of emotion [19], but the overwhelming consensus shows Ekman's basic six to be sufficient for the studies given. Further research may indicate that other dimensions should be more heavily weighted, such as bordeum and sarcasm, for example. Nevertheless, the basic six provided by Ekman and Freisen continue to be used with relative success.

One final thought for an improved system would involve using multiple inputs rather than being restricted to only facial expressions. Rarely do humans use only one mode of input when making a decision. Human beings are multi-modal pattern recognition machines capable of handling several inputs in parallel. In order to achieve effective human-computer interaction a model of the way humans interpret emotions must be created. This model should use every aspect of the way which humans interpret and express emotions. Typically during human-human communication we often use the context of the situation to help infer an expression. For example, if an individual is in a scenario where "joy" is expected, such

as a birthday party, then the probability of encountering a joyful expression is greater than that of "sadness". Also, it isn't just the expression that provides information, but also facial coloration. A model should be able to represent color as well as a change in expression. Embarrassment, anger, and sadness typically are typically associated with redness around the eyes and face.

REFERENCES

[1] Alberto Battocchi, Fabio Pianesi, and Dina Goren-Bar. A first evaluation study of a database of kinetic facial expressions (dafex). In *ICMI '05: Proceedings of the 7th international conference on Multimodal interfaces*, pages 214–221, New York, NY, USA, 2005. ACM.

[2] C. A. Bouman. Cluster: An unsupervised algorithm for modeling Gaussian mixtures. https://engineering.purdue.edu/ bouman/software/cluster/manual.pdf, April 1997.

[3] Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[4] Ira Cohen, Nicu Sebe, Fabio G. Cozman, and Thomas S. Huang. Semi-supervised learning for facial expression recognition. In *MIR '03: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 17–22, New York, NY, USA, 2003. ACM.

[5] Jeffrey F. Cohn. Foundations of human computing: facial expression and emotion. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 233–238, New York, NY, USA, 2006. ACM.

[6] Jeffrey F. Cohn, Karen Schmidt, Ralph Gross, and Paul Ekman. Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification. 2002.

[7] S. Das, A. Halder, P. Bhowmik, A. Chakraborty, A. Konar, and A. K. Nagar. Voice and facial expression based classification of emotion using linear support vector machine. *Developments in eSystems Engineering, International Conference on*, 0:377–384, 2009.

[8] Dragoş Datcu and Léon Rothkrantz. Facial expression recognition in still pictures and videos using active appearance models: a comparison approach. In *CompSysTech '07: Proceedings of the 2007 international conference on Computer systems and technologies*, pages 1–6, New York, NY, USA, 2007. ACM.

[9] Richard O. Duda, Peter E. Hart, and David G.Stork. *Pattern Classification 2nd ed.* Wiley-Interscience, 2001.

[10] Manfred Eckschlager, Regina Bernhaupt, and Manfred Tscheligi. Nemesys: neural emotion eliciting system. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 1347–1350, New York, NY, USA, 2005. ACM.

[11] G.J. Edwards, T.F. Cootes, and C.J. Taylor. Proceedings of the european conference on computer vision. In *Face Recogion Using Active Appearance Models*, 1998.

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

[13] P. Ekman and W. Friesen. *Emotional Facial Action Coding System.* Unpublished Manual, 1984. http://www.face-and-emotion.com/dataface/facs/emfacs.jsp.

[14] Paul Ekman. *Emotions Revealed: Recognizing Faces and Feeling to Improve Communication and Emotional Life.* Time Books, 2003.

[15] Beat Fasel, Florent Monay, and Daniel Gatica-Perez. Latent semantic analysis of facial action codes for automatic facial expression recognition. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 181–188, New York, NY, USA, 2004. ACM.

[16] Chen Feng-jun, Wang Zhi-liang, Xu Zheng-guang, and Xiao Jiang. Facial expression recognition based on wavelet energy distribution feature and neural network ensemble. *Intelligent Systems, WRI Global Congress on*, 2:122–126, 2009.

[17] T.F. Cootes G.J. Edwards and C.J. Taylor. Face recognition using active appearance models. In *Proc. European Conference on Computer Vision 1998*, pages 581–695. Springer, 1998.

[18] Rita T. Griesser, Douglas W. Cunningham, Christian Wallraven, and Heinrich H. Bülthoff. Psychophysical investigation of facial expressions using computer animated faces. In *APGV '07: Proceedings of the 4th symposium on Applied perception in graphics and visualization*, pages 11–18, New York, NY, USA, 2007. ACM.

[19] Robert Horlings, Dragos Datcu, and Leon J. M. Rothkrantz. Emotion recognition using brain activity. In *CompSysTech '08: Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing*, pages II.1–1, New York, NY, USA, 2008. ACM.

[20] Bin Hua and Ting Liu. Facial expression recognition based on local feature bidirectional 2dpca. *Information Technology and Computer Science, International Conference on*, 1:301–304, 2009.

[21] X. Huang and Y. Lin. A vision-based hybrid method for facial expression recognition. In *Ambi-Sys '08: Proceedings of the 1st international conference on Ambient media and systems*, pages 1–7, ICST, Brussels, Belgium, Belgium, 2008. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[22] Takeo Kanade, Yingli Tian, and Jeffrey F. Cohn. Comprehensive database for facial expression analysis. *Automatic Face and Gesture Recognition, IEEE International Conference on*, 0:46, 2000.

[23] Kostas Karpouzis, Nicolas Tsapatsoulis, Amaryllis Raouzaiou, George Moshovitis, and Stefanos Kollias. Enhancing nonverbal human computer interaction with expression recognition. *SIGCAPH Comput. Phys. Handicap.*, (67):1–9, 2000.

[24] Masood Mehmood Khan, Michael Ingleby, and Robert D. Ward. Automated facial expression classification and affect interpretation using infrared measurement of facial skin temperature variations. *ACM Trans. Auton. Adapt. Syst.*, 1(1):91–113, 2006.

[25] Byungsung Lee, Junchul Chun, and Peom Park. Classification of facial expression using svm for emotion care service system. *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, ACIS International Conference on*, 0:8–12, 2008.

[26] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. *Automatic Face and Gesture Recognition, IEEE International Conference on*, 0:200, 1998.

[27] Michael J. Lyons, Julien Budynek, and Shigeru Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:1357–1362, 1999.

[28] A. Mehrabian. Communication without words. *Psychology Today*, 1968.

[29] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. Audio-visual emotion recognition using gaussian mixture models for face and voice. *Multimedia, International Symposium on*, 0:250–257, 2008.

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

[31] Philipp Michel and Rana El Kaliouby. Real time facial expression recognition in video using support vector machines. In *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264, New York, NY, USA, 2003. ACM.

[32] Muid Mufti and Assia Khanam. Fuzzy rule based facial expression recognition. In *CIMCA '06: Proceedings of the International Conference on Computational Inteligence*

for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce, page 57, Washington, DC, USA, 2006. IEEE Computer Society.

[33] P. Nagesh and Baoxin Li. A compressive sensing approach for expression-invariant face recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1518–1525, 2009.

[34] M. Obaid, R. Mukundan, R. Goecke, M. Billinghurst, and H. Seichter. A quadratic deformation model for facial expression recognition. *Digital Image Computing: Techniques and Applications*, 0:264–270, 2009.

[35] R. Ramanathan, K.P. Soman, Arun. S. Nair, V. Vidhya Sagar, and N. Sriram. A support vector machines approach for efficient facial expression recognition. *Advances in Recent Technologies in Communication and Computing, International Conference on*, 0:850–854, 2009.

[36] M. Rizon, D. Hazry, M. Karthigayan, R. Nagarajan, N. Alajlan, and Y. Sazali. Personalized human emotion classification using genetic algorithm. *Visualisation, International Conference in*, 0:224–228, 2009.

[37] Nectarios Rose. Facial expression classification using gabor and log-gabor filters. *Automatic Face and Gesture Recognition, IEEE International Conference on*, 0:346–350, 2006.

[38] Yunus Saatci and Christopher Town. Cascaded classification of gender and facial expression using active appearance models. In *FGR '06: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 393–400, Washington, DC, USA, 2006. IEEE Computer Society.

[39] Diane J. Schiano, Sheryl M. Ehrlich, Krisnawan Rahardja, and Kyle Sheridan. Face to interface: facial affect in (hu)man and machine. In *CHI '00: Proceedings of the SIGCHI*

conference on Human factors in computing systems, pages 193–200, New York, NY, USA, 2000. ACM.

[40] Lifeng Shang and Kwok-Ping Chan. Nonparametric discriminant hmm and application to facial expression recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:2090–2096, 2009.

[41] Yafei Sun, Zhishu Li, Changjie Tang, Wangping Zhou, and Rong Jiang. An evolving neural network for authentic emotion classification. *International Conference on Natural Computation*, 2:109–113, 2009.

[42] Ying-Li Tian, Takeo Kanade, and Jeffrey Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(1):97 – 115, March 2001.

[43] Matthew A. Turk and Alex P. Pentland. Face recognition using eigenfaces. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 1991)*, 1991.

[44] Roberto Valenti, Alejandro Jaimes, and Nicu Sebe. Facial expression recognition as a creative interface. In *IUI '08: Proceedings of the 13th international conference on Intelligent user interfaces*, pages 433–434, New York, NY, USA, 2008. ACM.

[45] Frank Walhoff. Facial expression and emotion database, 2006.

[46] Christian Wallraven, Heinrich H. Bülthoff, Douglas W. Cunningham, Jan Fischer, and Dirk Bartz. Evaluation of real-world and computer-generated stylized facial expressions. *ACM Trans. Appl. Percept.*, 4(3):16, 2007.

[47] Jia-Jun Wong and Siu-Yeung Cho. Facial emotion recognition by adaptive processing of tree structures. In *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing*, pages 23–30, New York, NY, USA, 2006. ACM.

[48] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3d facial expression database for facial behavior research. *Automatic Face and Gesture Recognition, IEEE International Conference on*, 0:211–216, 2006.

[49] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:39–58, 2008.

[50] Ce Zhan, Wanqing Li, Philip Ogunbona, and Farzad Safaei. Facial expression recognition for multiplayer online games. In *IE '06: Proceedings of the 3rd Australasian conference on Interactive entertainment*, pages 52–58, Murdoch University, Australia, Australia, 2006. Murdoch University.