Automated Real Time Emotion Recognition using Facial Expression Analysis

by

Ashleigh Fratesi

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master

of

Computer Science

Carleton University
Ottawa, Ontario

Abstract

The focus of this study is on computer automated perception of human emotion. We explore the use of Fisherfaces for the recognition of human emotion in facial images. We train a multitude of Fisherface models and evaluate their performance against an independent test set. We build and test a compound hierarchical system that attempts to interpret human emotion in real time using face detection and tracking algorithms in conjunction with our facial expression analysis methodology.

Our results indicate that Fisherfaces can be useful in predicting emotion based on content retrieved from facial images. We note that, with this approach, some emotions are more easily predicted than others. We also suggest that a compound hierarchical model is more effective than a single stand-alone Fisherface model.

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my advisor Dr. Anthony Whitehead for his guidance, patience and breadth of knowledge and experience. His contributions were invaluable and critical to the success of this thesis.

I would also like to thank the team of graduate students working under Dr. Whitehead with whom I shared office space for their insightful comments and willingness to lend a helping hand.

I would like to extend a special thank you to Mr. Christopher Clarke for his assistance in getting my Linux environment up and running.

Lastly, I would like to thank my family and friends for their continued support and encouragement.

Table of Contents

List of Tables

List of Figures

1 Chapter: Introduction

Emotion recognition is a natural capability in human beings. However, if we are to ever create a humanoid robot that can interact and emote with its human companions, the difficult task of emotion recognition will have to be solved. The ability for a computer to recognize human emotion has many highly valuable real world applications. Consider the domain of therapy robots which are designed to provide care and comfort for infirm and disabled individuals. These machines could leverage information on a patient's current and evolving state of mind, in order to tailor personalized strategies for patient care and interaction. For example, when a patient is upset or unhappy, a more effective strategy may be to take a moment to recognize the emotion and offer sympathy. The patient, feeling heard and validated, may be more likely to be cooperative in subsequent requests issued by the machine. In the case of socially assistive robots (SARs), it is the emotional relationship with the robot itself that provides the therapeutic benefit. There has been a fair amount of research conducted in the area of emotion and its implications on human-robot interaction (HRI) and interface design (e.g. Norman [51], Picard [66]).

Even outside of the realm of robotics, working with computers that have the ability to sense and respond to emotional state can go a long way to improving the quality of human-computer interaction (HCI). By designing HCI to be more like human-human interaction, we have the ability to create more natural, fulfilling, and productive working relationships with our machines.

1.1 Problem Statement

This thesis explores the idea of applying the Fisherface method [6] to build an accurate and reliable classifier for human emotion as depicted in still images obtained from a video stream. We explore the efficacy of the Fisherface model in classifying human emotion based on content obtained from facial images. We use the prototypic emotions as defined by Ekman [22] - happiness (H), sadness (S), anger (A), disgust (D), fear (F), surprise (SU), and neutrality (N), as our seven classes.

1.2 Contributions

Our main contribution is a thorough systematic evaluation of the Fisherface model and its capability in predicting emotional expressions in facial images. The dataset we used to train our models came from one of the largest most inclusive databases [13] currently available. The datasets we used to test our models are derived from two additional distinct databases [11] [45]. As there is no overlap between our training and testing sets, our evaluation is person independent. Also, as our training and testing sets are database independent, our evaluation is more realistic than if we had used one database and performed cross-validation. Another significant contribution is the development of a compound hierarchical Fisherface model that outperforms individual stand-alone models.

2 Chapter: Background

2.1 Human Emotion Recognition

It is believed that humans interpret emotion primarily through facial expression analysis. It is therefore not surprising that a vast majority of automated emotion recognition systems attempt to model this approach. Research conducted by Mehrabian [49] reported that the facial expression of a speaker can contribute over half of the communicated message, followed by voice intonation at roughly 38% and trailed by the actual words spoken at a mere 7%. This suggests that facial expressions are a key modality in human communication.

Darwin [16] claimed the universality of facial expressions and their continuity in man and animals as early as 1872. In the early 1970's, Ekman and Friesen [20] suggested that the six primary emotions – happiness, sadness, anger, fear, disgust, and surprise – each elicit a unique facial expression. Ekman and Friesen subsequently developed the Facial Action Coding System (FACS) [21] as an objective measurement of facial activity. FACS provided a linguistic description of visually detectable facial changes caused by contractions of the facial muscles. A trained human FACS coder can decompose a shown expression into the specific Action Units (AUs) that describe the expression. The aim was to overcome biases imposed by individual observers by representing expressions and facial behaviors in terms of a fixed set of facial parameters. Many variants of FACS

followed, including the Emotional Facial Action Coding System (EMFACS), which was focused on detecting only those AUs with emotional significance.

In 1994, Russell [67] questioned the claims of universal emotion recognition from facial expressions. Later that same year, Ekman [23] and Izard [33] replied to Russell's critique, offering a point-by-point refutation of his claims. It is currently widely accepted that the recognition of emotions from facial expressions is universal and constant across cultures. It should be noted, however, that the expression of emotion through facial changes can differ based on culture and can be situationally dependent. For example, Japanese expressers often suppress their facial expressions when in the presence an authority figure. [24]

2.2 Computer Based Emotion Recognition

There are many things to consider when attempting to interpret emotion through facial expression analysis. One of the first items to note is that classification into six or seven prototypic categories certainly simplifies the problem space, but at what cost? Parrott [64] identified 136 emotional states recognizable through facial expression analysis. Additionally, it is most often the case that emotions elicited by human expressers are blended (e.g. happy and surprised). Furthermore, expression intensity can vary greatly from person to person.

Occlusions, or visual obstructions, are another consideration when designing a facial expression analyzer. The system must be robust against the presence of glasses, facial hair, scarves, hands, etc. Kotsia et al. [41] reports that the occlusion of the mouth can reduce classification results by more than 50%. There has been an ongoing debate on whether facial recognition practiced in humans is a holistic process or whether it is component based. As such, automated systems have been modelled on both schools of thought [26] [31] [29] [38] [47] [57] [77] [78].

It is also necessary to distinguish between posed and spontaneous expressions which can present very differently from one another. Posed expressions are easy to capture and recognize but are often highly exaggerated. Sebe et al. [69] attempted to compile a spontaneous expression database. In order to capture genuine expressions, the researchers designed a booth with a hidden camera which recorded participant's reactions to emotion eliciting videos. Participants were only informed post session that their facial expressions had been recorded. Only half of the participants agreed to allow the researchers to use their data. Sebe et al. discovered that it was very difficult to induce certain expressions (e.g. sadness or fear) and that participants reports of feeling a given emotion did not always match the observed facial expression. Databases containing spontaneous expressions continue to be developed. The interested reader can refer to Section 8 in the Face Expression Recognition and Analysis: The State of the Art [7] for further information on available databases.

Temporal dynamics, or the phases of an expression over time, must also be considered. The most effective systems will attempt classification only at the apex of expression, rather than during its onset or offset. An effective system must also be able to deal with additional noise (e.g. differences in lighting), and variations in size and orientation of faces in an input image.

2.3 Phases of Automated Emotion Recognition Systems

Real time emotion recognition using facial expression analysis requires three phases. In the first phase, a face is detected and subsequently tracked. In the second phase, facial features are extracted, and in the third phase, mathematical models are applied to the facial feature data in an attempt to interpret the content.

Face detection is the act of locating a face within an image, whereas face tracking is the process of following a located face in a video sequence from one frame to the next. Kanade and Schneiderman [68] used statistical methods to develop a robust face detection methodology. The Kanade-Lucas-Tomasi (KLT) tracker [44] [74] leveraged spatial intensity information from a localized face thereby reducing the search area for subsequent tracking. Tao and Huang [72] developed the Piecewise Bezier Volume Deformation (PBVD) tracker which employs a 3D wireframe mesh model to track faces and facial features. Viola and Jones [76] developed an object detection frame work (VJ-ODF) that was capable of rapid face detection.

The VJ-ODF capitalizes on the notion that all human faces share similar qualities and assumes a full frontal view of the face with minimal rotation. The potential Haar-like feature set relevant to face detection is identified using AdaBoost [75] as a feature selection mechanism. Weighted linear combinations of weak classifiers form a strong classification system [75]. Early stages of a cascade of classifiers reject large portions of an image which are unlikely to contain faces, in order to focus successive computations on promising regions. The use of an Integral Image, which can be computed from an image using only a few operations per pixel, further increases algorithm efficiency by allowing haar-like features to be computed in constant time at any scale or location [75].

There are many approaches to face detection and tracking and a plethora of literature on the same. The interested reader can refer to Section 6 of Face Expression Recognition and Analysis: The State of the Art [7] for further details.

There are also many approaches to feature extraction. A holistic methodology processes the face in its entirety, whereas a local approach focusses on areas of the face that are likely to change with a change in facial expression. Some features of the face are permanent, for example eyes, nose, mouth, wrinkles due to aging, etc. Other features of the face are transient, for example, lines or furrows that appear with changes in expression that are not present on a neutral face. Both transient and intransient facial features may change with changes in facial expression. Deformation extraction based systems [50] [60] [81] [73] can be applied to both single images and image sequences,

whereas motion extraction based systems [1] [15] [58] can only operate on image sequences, and require knowledge of the neutral face. Some systems take a hybrid approach to feature extraction [18] [29] [41] [59].

2.3.1 Feature Representation for Image Classification

In their raw pixel format, images lie in a very high dimensional feature space and can require a lot of memory for storage and retrieval. Additionally, the time and resources required to execute computations on this representation are often too high to be useful in practical settings.

Principal component analysis (PCA) has been employed in previous studies as a technique for dimensionality reduction [17] in images and was first introduced by Sirovich and Kirby in 1987 [70]. This method attempts to model the variance within the dataset, but does not take any class labels into consideration. Although a very powerful and useful technique, for example in cases where image reconstruction is necessary, using this approach for classification may lead to substandard results [6].

When the main task is classification, and we are working with a labelled dataset [42], leveraging class information can be useful in building a more reliable technique for reducing the dimensionality of the image space. For the purpose of reducing dimensionality in the image classification domain, ideally we are aiming to map same-

class images to a single point in the subspace while maximizing the distance to instances belonging to different classes. The techniques that are employed to achieve this goal are referred to as Discriminant Analysis [36].

One of the most well-known and widely used Discriminant Analysis techniques is Linear Discriminant Analysis (LDA) [43]. LDA was first introduced by R. A. Fisher [28] for the classification of flowers in his 1936 paper "The use of multiple measurements in taxonomic problems." The Iris dataset described in his paper is still available today in the UCI Machine Learning Repository [2]. When LDA is applied to a set of face images, the resulting basis vectors, which define the image subspace representation, are referred to as Fisherfaces.

2.3.2 Fisherfaces

Fisherfaces were first described by Belhumeur, Hespanha, and Kriegman [6] in their paper "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection." The authors note that with respect to a set of images, variation within classes lies in a linear subspace of the image space indicating that the classes are convex, and thus linearly separable.

In the Fisherfaces method, the task of classification is simplified by the use of Fisher's Linear Discriminant (FLD) which attains a greater between-class scatter than PCA. In

order to obtain tightly clustered well separated classes, LDA maximizes the ratio of the determinant of between-class to within-class scatter.

The Fisherfaces technique takes a pattern classification approach considering each pixel in an image as a coordinate in the high-dimensional image space. The algorithm begins by creating a matrix wherein each column vector (consisting of pixel intensities) represents an image. A corresponding class vector containing class labels is also created. The image matrix is projected into (n-c)-dimensional subspace (where n is the number of images and c is the number of classes). The between-class and within-class scatter of the projection is calculated and LDA is applied. For our purposes here, we leveraged functionality available within the libraries of OpenCV [52] to implement LDA using the Fisherface methodology.

This method requires that all images, both in the training and testing set, be equal in size. Method performance is highest when all images are full frontal head shots with major features aligned. This technique does not work on an image directly rather it converts images into greyscale vector matrices and works with the vector form. Ultimately, each image is represented by a weight vector which indicates the percentage of each Fisherface it contains. It is this weight vector representing unique image attributes that is used in a nearest neighbour search of the training set to predict the identity of an unknown face.

Written a bit more formally, we have an image I is represented by a weight vector w = $\{w_1, w_2, \ldots, w_k\}$ where the image is defined by:

$$I = \text{mean face} + w_1x_1 + w_2x_2 + \ldots + w_kx_k \qquad (1)$$

such that $x_1..x_k$ are Fisherface templates ($k <$ number of training images). Refer to Figure 1 for a graphical representation of (1).

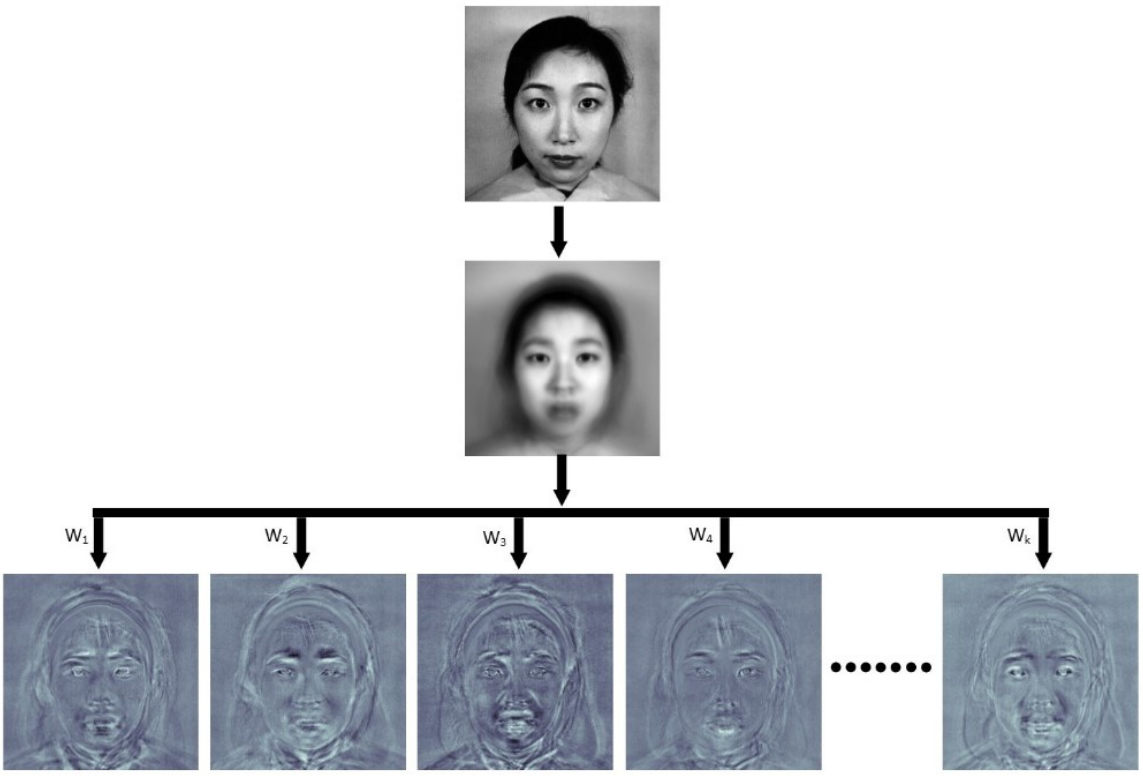

Figure 1 High-level visualization of how an image is represented in the Fisherface model. Each image is comprised of the average face and a weighted sum of the Fisherface templates. Images courtesy of the JAFFE database [34].

In one experiment reported on in the Belhumeur, Hespanha, and Kriegman paper, the researchers applied the Fisherfaces method to a set images depicting faces with or without glasses. Their results indicated that "PCA had recognition rates near chance, since, in most cases, it classified both images with and without glasses to the same class. On the other hand, the Fisherface methods can be viewed as deriving a template which is suited for finding glasses and ignoring other characteristics of the face." [6] They also commented that "it is expected that the same techniques could be applied to identifying facial expressions where the set of training images is divided into classes based on the facial expression." [6]

The two main approaches to content interpretation, or classification, are template based [3] [12] [73] and rule based [58] [59] [60]. In a template based approach, an unknown facial image is compared to a set of templates and a "best match" predicts the displayed emotion. One of the main drawbacks to a template based methodology lies in the variability of facial expressions and intensities both within and between different individuals. Another challenge for a template based methodology is quantification of a given expression. A rule based system is comprised of a set of premises that hold for each class. In this approach, an unknown facial image, described in terms of facial actions, is compared to the prototypical facial actions for a given emotion and a "best match" predicts the displayed emotion.

2.4 History of Automated Emotion Recognition Systems

In the following section, we describe the evolution of automated facial emotion analyzers by highlighting previous research. We offer a comprehensive discussion of many of the successes and failures in the area.

One of the earliest facial expression analysis systems was developed by Suwa et al. [71] in 1978. Their system for analyzing facial expressions included twenty tracking points manually placed on the faces present in a sequence of images. Although an important first step toward recognition of facial expressions by a machine, automating their process was not a possibility at that point in time. The robust face detection and tracking systems that would have made this possible, which relied on inexpensive computing power, did not begin becoming available until the late 1980s and early 1990s.

Beginning in 1994, Black and Yacoob [8] [9] [10] conducted research into analyzing facial expressions with local parameterized models of image motion representing rigid head motions, and non-rigid facial feature motions of the mouth, eyes, and eyebrows. Initial regions for the head and facial features were selected by hand and subsequently tracked automatically. They developed a model for each of the six basic emotional expressions and a set of rules for detecting the onset, apex, and offset of an expression. Their method did not deal with blended expressions. Experiments were conducted on forty subjects in the laboratory ranging in age and ethnicity. Seventy image sequences containing 145 expressions resulted in an average 88% correct recognition rate.

Additional testing with television and movie sequences resulted in a 60-100% correct recognition rate. The research presented by Black and Yacoob suggested that recognition of prototypical emotions was possible even in the presence of head motion and pose variation.

Zhao et al. [80] utilized a point-based frontal-view face model and employed a backpropagation neural network trained on 94 images for facial expression classification into one of the six prototypic emotions. Zhao et al. did not address automatic facial expression data extraction. Instead, 10 distances were manually measured on each of the 94 images, and the difference between a distance measured in an examined image and the same distance measured in the neutral face of the same individual was calculated. Each measure was normalized and mapped into a signaled interval to be used as input to the neural network. The neural network was tested on the same data used during training with a 100% recognition rate. Results for unknown subjects were not considered.

Padgett and Cottrell [57] presented an automatic facial expression analyzer that was capable of identifying six basic emotions. Rather than dealing with manual or automated feature extraction Padgett and Cottrell opted for a different approach. They digitized 97 images and scaled them so that the prominent facial features were roughly aligned, then extracted data from pixel blocks that were placed on the eyes and the mouth. They projected this data onto the top 15 Principal Component Analysis (PCA) eigenvectors and the normalized projections were fed into a back-propagation neural network. The hidden layer of the neural network contained 10 nodes and employed a nonlinear sigmoid

activation function. The output layer contained 7 units - one for each possible classification. They trained the network on images of 11 subjects and tested it on the images of the 12th subject. By using a leave-one-out approach, they were able to train 12 different networks with an average correct recognition rate of 86%.

Otsuka and Ohya [54] trained a HMM on 120 image sequences displayed by two male subjects. By applying an adapted gradient-based optical-flow algorithm, they estimated the motion in local facial areas of the mouth and the right eye. Their choice of excluding the motion of left eye based on facial symmetry simplifies calculations but consequently makes their method insensitive to unilateral appearance changes of the left eye. Once the optical-flow algorithm is applied, a 2D Fourier transform is used on vertical and horizontal velocity fields, and lower-frequency coefficients are extracted to form a 15D feature vector used for classification. The temporal sequence of the feature vector is matched to models of the prototypic emotions using a left-to-right Hidden Markov Model. Although Otsuka and Ohya claim high recognition performance for their system, the method was only tested on the two subjects used during the training phase, and no quantifiable results are available.

Bartlett et al. [3] proposed a system that combined three different methodologies – holistic difference-images motion extraction coupled with PCA, feature measurements along predefined intensity profiles for the estimation of wrinkles and holistic dense optical flow for whole-face motion extraction. Center points for the eyes and the mouth were manually located in the neutral face that began each image sequence. Rotation,

scaling and warping of aspect ratios were used to normalize facial images. A feed-forward neural network was used for classification. The input to the network consisted of 50 PCA component projections, 5 feature density measurements and six optical flow-based template matches. There were 10 hidden units and 6 output units – one for each prototypic expression. Correct recognition rates of 92-96% were reported during initial testing. Their fully automatic real-time emotion recognition system has been successfully deployed on Sony's Aibo pet robot, ATR's RoboVie and CU Animator with recognition rates exceeding 84%.

Yoneyama et al. [78] fit 8x10 quadratic grids to normalized face images and calculated the average optical flow between a neutral and an emotive face in each of these regions. The magnitude and direction of the optical flows were simplified to a ternary value magnitude in the vertical direction, excluding horizontal movement information. As a consequence, their method will not detect any change in facial features involving horizontal movement. Two neural networks were trained to identify one of four facial expressions (happiness, anger, sadness and surprise). One of the neural networks (NN1) was trained using 40 samples displayed by 10 participants. The other neural network (NN2) was trained on the best sample per expression category for a total training set of only 4 samples. Each test image was fed into NN1 and the Euclidean distance between its output, and that of each training sample, was calculated. Distances were averaged per expression and if the distance between the minimal average and the second minimal average was greater than 1, an expression prediction was made. If NN1 could not make a prediction, NN2 was used to make an expression prediction in a similar fashion. The

reported average recognition rate was 92% however testing images came directly from the training set.

Kobayashi and Hara [38] trained a 234x50x6 back-propagation neural network on 90 images of six basic facial expressions shown by 15 subjects. Normalization of input images began with an affine transformation which ensured that the distance between the irises of the left and right eyes became 20 pixels. Normalized brightness distributions of vertical lines crossing facial feature points were used as input for the neural network. Facial appearance changes in the horizontal direction were not considered. Their system was tested on 90 images from 15 novel subjects with an average recognition rate of 85%.

Huang and Huang [29] used a point distribution model (PDM) in conjunction with a mouth template in a hybrid feature/template based approach. The PDM is manually placed on the input image which helps in defining an appropriate search region for the mouth. A parabolic curve, found by identifying the darkest point in each vertical segment of the search region, is used to approximate the line where the upper and lower lip meet. Edges with the strongest gradient above and below the parabolic curve are used to define the upper and lower lips with two additional parabolic curves. The method for locating the mouth will fail if the point at which the upper and lower lips meet is not the darkest in the search region, for example, if the subject's teeth are visible. The PDM was generated from 90 facial feature points in 90 images of 15 Chinese subjects displaying six prototypic emotions. Huang and Huang also generated 10 Action Parameters (APs), used for classification, by calculating the difference between model feature parameters of a

neutral face and an expressive face of the same individual. They discovered that the first two eigenvalue terms were capable of representing over 90% of the APs variations. A minimum distance classifier was used to cluster the two principal action parameters of the training samples into six clusters representing the six prototypic expressions. Their system was tested on 90 images of the same 15 subjects used to train the model with a correct recognition rate of over 84%. Images from novel subjects were not tested.

Essa and Pentland [26] developed a system that was capable of automatic face detection and analysis in frontal-view facial image sequences. Faces and facial features were located and tracked in image sequences using a view-based and modular eigenspace method. Input faces were warped in order to match canonical face meshes. A two dimensional spatio-temporal energy representation of facial motion estimated from consecutive frames of training images was used to create dynamic face models. By learning typical 2D motion views for each expression category, Essa and Pentland generated the spatio-temporal templates for six different expressions – four emotional expressions (surprise, sadness, anger and disgust), and two facial actions (raised eyebrows and smile). Tested image motion energy was classified by comparing it to the six templates, and by using the Euclidean norms of the differences as a measure of similarity. An average correct recognition rate of 98% was achieved on 52 frontal-view image sequences of 8 participants.

Wang et al. [77] developed a three-class system used to identify anger, happiness and surprise. In this system, a labeled graph was used for face representation. The nodes of

the graph were comprised of facial feature points (FFPs) manually placed on the first frame of an image sequence and automatically tracked in subsequent frames. The links between nodes represented Euclidean distances between the FFPs. Links were weighted to account for impacting properties of facial features such as the potential for violent deformation. B-spline curves between the same nodes on consecutive frames, shown by five subjects in ten image sequences, were used to construct the expression models. Classification was based on the similarity of the trajectories of FFPs in a test image sequence as compared with the expression models. The system was tested on eight subjects who together provided 29 image sequences. The reported average recognition rate was 95%.

Hong et al. [31] developed an emotional facial expression recognition system that leveraged identity information. Personalized image galleries of 9 participants displaying the seven basic emotions were collected. The researchers were working under the assumption that people who looked alike would display similar emotional facial expressions. Elastic graph matching was employed to match an incoming image to the participant they most resembled, and then to which particular emotion in their personalized gallery the image best matched. Their method has been tested on 25 subjects with an average recognition rate of 89% for familiar subjects and 73% for novel subjects.

Edwards et al. [19] employed the use of an active appearance model (AAM) to match statistical models of the shape and appearance of the seven prototypic facial expressions

to novel images. 88 training images manually labelled with 122 facial feature landmark points were aligned to a common coordinate frame. Principal component analysis (PCA) was applied to the training images to extract shape and gray-level parameter information for the model. A multivariate multiple regression model derived from the training data was used during the recognition stage. The method, tested on 200 images of known subjects, achieved an average recognition rate of 74%.

Zhang et al. [79] trained a neural network for facial expression classification into one of the seven prototypic emotions. The input to the network consisted of the geometric position of 34 facial points manually placed on normalized frontal view images of 9 female Japanese subjects, along with 18 Gabor wavelet coefficients (3 spatial frequencies x 6 orientations) sampled at each point. The dataset of 213 images was partitioned into 10 segments, and a leave-one-out strategy was employed in order to train and test 10 neural networks. The average correct recognition rate over the networks was over 90% however it unknown how the networks will perform in the case of a novel subject.

Lyons et al. [47] also presented a Gabor wavelet-based facial analysis framework trained and tested on the same database used by Zhang et al. They also represented faces with 34 facial points manually placed on normalized frontal view images of 9 female Japanese subjects, but included 30 Gabor wavelet coefficients (5 spatial frequencies x 6 orientations) sampled at each point. They built six binary classifiers, one to test for the presence of each prototypic emotion, and combined them into a single facial expression classifier. In the case where the input image was positively classified into more than one

expression category, normalized distances to cluster centres were used as the deciding factor. An input image was classified as neutral if it did not test positively for any of the six binary classifiers. In person-dependent testing their method achieved a generalization rate of 92%. When the training and testing sets were partitioned to allow for person-independent testing, the generalization rate was 75%.

Pantic and Rothkrantz [60] [61] [62] [63] conducted research into the automatic coding of AUs defined by Ekman [20] and developed a 2D point-based model for both frontal and side views of emotionally expressive facial images. They employed multiple redundant feature detectors and leveraged facial anatomy information in order to localize facial features from both views. Their method was capable of automatic facial action coding of a dual-view input images achieving an average rate of 92% for the upper face AUs and 86% for the lower face AUs. A rule-based approach to classification was performed by comparing automated AU-coded descriptions of input images to AU-coded descriptions of the six prototypic expressions derived from Ekman [22]. The system was tested on 265 dual facial view expression images displayed by 8 participants with a correct recognition ratio of 91%. Tian et al. [73] also developed a face analysis system that could automatically identify six upper face AUs and 10 lower face AUs with an average recognition rate over 96%. They did not, however, address profile views, which may be necessary in many real-time settings.

Bourel et al. [12] conducted research into the recognition of facial expressions in the presence of occlusions. Their results suggested that local modular classifiers which could

be weighted and summed may be most effective in these situations. They also reported that occlusion of the mouth was most detrimental when the displayed expression was sadness.

Anderson and McOwen [1] used motion signatures from 253 emotive image sequence samples to build a fully automated, multistage facial emotion recognition system using support vector machines (SVMs) and multilayer perceptrons (MLPs). Their real-time system was able to operate efficiently in cluttered scenes with recognition rates around 80%.

Ji and Idrissi [35] built an automated system which evaluated appearance and spatial-temporal data on video sequences. These video sequences began with a neutral face and culminated at an expressive face displaying one of the six basic emotions. Facial features were tracked and movement was captured by analyzing texture and shape changes between frames. Many of the latter frames in the sequence were used to predict the emotion with initial frames assumed to be "neutral". They report between 94-100% accuracy rates on various versions of their system.

Majumder, Behera, and Subramanian [48] built an automated system for feature detection and emotion classification which used a feature vector containing 23 facial points extracted from prominent facial features such as the eyes, eyebrows, and mouth. Movement, or displacement of facial features, was calculated using the neutral expression as a point of reference, and the expression was labelled as one of the six basic emotion.

They tested their system on 81 video clips and reported an average recognition rate of almost 94%.

Happy and Routray [29] built an automatic emotion recognition system by analyzing salient facial patches surrounding major facial features (I.e. areas surrounding the eyebrows, eyes, nose and mouth). The deformation of these areas was used to classify an expression into one of the six basic emotions. They discovered that not all patches were equally important in discriminating between any two emotions. Areas that resulted in maximum levels of discrimination between emotions were used to build binary classifiers pairing all possible combinations of the basic emotions. They implemented a voting strategy using their 2-class classifiers and found that a surprised expression was most easily identifiable whereas an angry expression was one of the most difficult to detect. The researchers' experimental results indicate that their system performs well, with accuracy ratings between 85-90%, even in low resolution images.

Direct comparison between the various approaches outlined above is difficult as there are no standard training and testing sets. Furthermore, there exist many differences in evaluation protocols. Our technique is unparalleled in that it employs one of the largest and most inclusive databases available [13] for classifier training, as well as one of the strictest (i.e. person-independent) testing methodologies.

2.5 Metrics

The metrics with which we evaluate our models are precision, recall, and the F-measure.

Precision is the fraction of correctly labelled samples over the total number of labelled samples for a given class. Precision is calculated using the following formula:

$$\text{Precision} = TP / (TP + FP) \qquad (2)$$

where TP represents the number of true positives and FP represents the number of false positives for a given class. High precision rates are important in scenarios where we want to be certain of our prediction.

Recall is the fraction of all positive samples that were identified. Recall is calculated using the following formula:

$$\text{Recall} = TP / (TP + FN) \qquad (3)$$

where TP represents the number of true positives, FP represents the number of false positives, and FN represents the number of false negatives for a given class. High recall rates are important in scenarios where we do not want to miss any positive samples.

The F-measure provides an overall score by taking both recall and precision into consideration. The F-measure is calculated using the following formula:

$$\text{F-measure} = 2\,(\text{Precision})(\text{Recall}) \,/\, (\text{Precision} + \text{Recall}) \qquad (4)$$

A high F-measure is important in scenarios where we wish to find a balance between high precision and recall rates.

3 Chapter: Methodology

3.1 Overview

To make our model run in real time, we leverage the Viola-Jones Object Detection framework (VJ-ODF) [75] described in 2.3, as implemented in OpenCV [52], for the task of face detection and tracking. The algorithm is robust, and fast enough to run in real time (roughly fifteen frames per second when implemented on a conventional 700 MHz Intel Pentium III [76]).

The VJ-OD executes on live video stream and identifies segments of frames that are likely to contain faces. In our research, faces identified using the VJ-ODF are aligned with the help of OpenCV feature detectors, and are subsequently classified based on emotional content.

In phase 1 of our Emotion Classification System (ECS) we use the VJ-ODF as is. Our minor contributions occur in phase 2 with proper alignment and preprocessing of detected faces. Our main contributions for our ECS occur in phase 3 with a hierarchical Fisherface classification model.

3.2 Algorithm

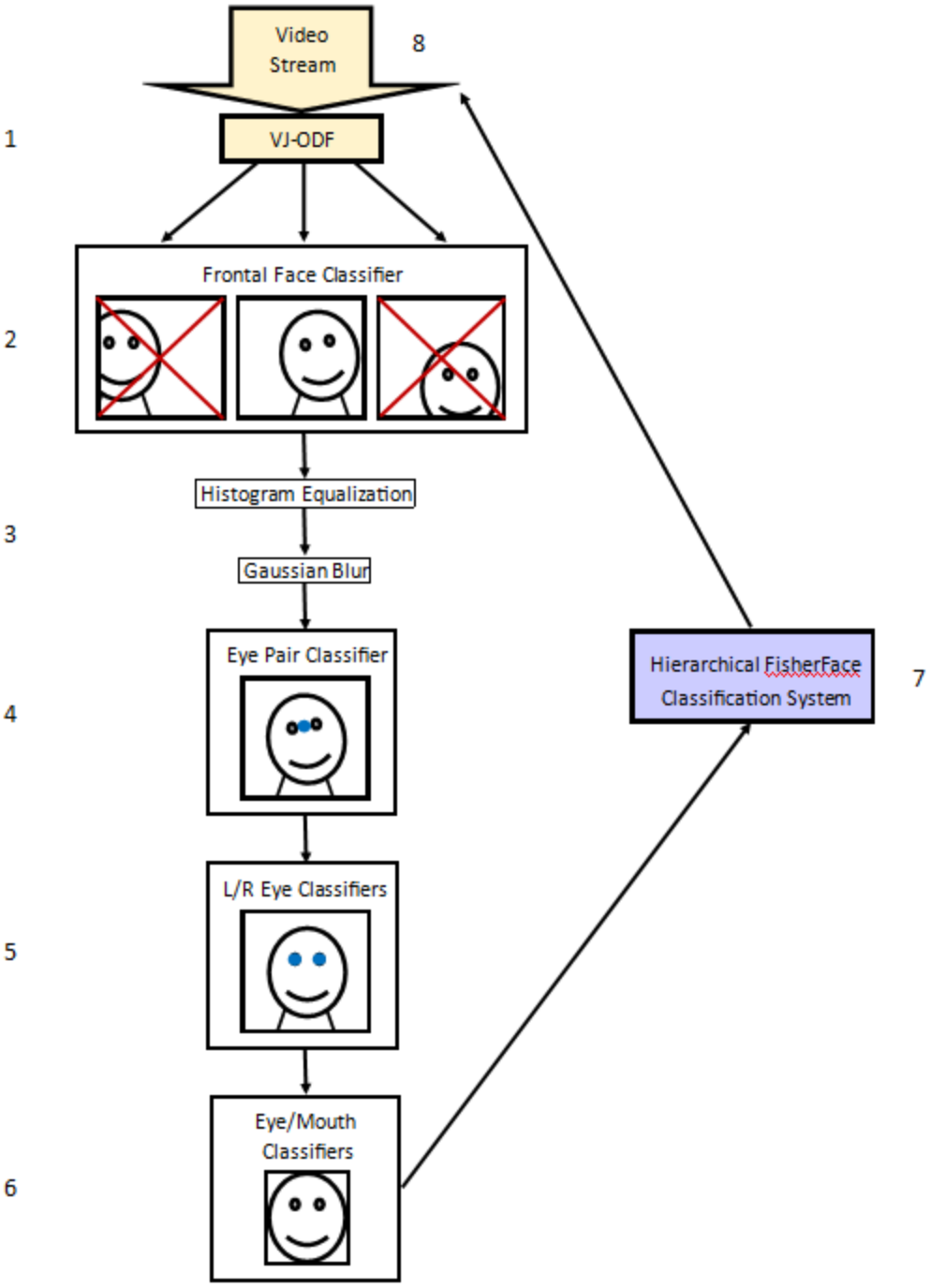

Figure 2 Real Time Emotion Classification System Methodology

3.2.1 Phase 1 - Face Detection and Tracking

1. Use VJ-ODF to detect faces in live video stream. Pass these frames on to step #2.

3.2.2 Phase 2 - Feature Extraction

2. Use OpenCV frontal face classifier to identify frames containing complete faces. For each frame containing a complete face, execute steps 3-8.
3. Apply histogram equalization for contrast enhancement [30] and Gaussian blur for noise reduction.
4. Use OpenCV eye pair classifier to locate the eyes. Calculate the midpoint between the eyes and use this data to vertically and horizontally align the face within the image.
5. Use OpenCV left and right eye detectors on bounded regions of interest. Use this location data to correct any facial rotation.
6. Use OpenCV eye pair and mouth classifiers. Use this location data to obtain a tightly cropped facial image that will align with the template images used to build the Fisherface models.

3.2.3 Phase 3 - Emotion Classification

7. Input the image into the hierarchical Fisherface Classification System for analysis.
8. Overlay prediction results of classification system to the live video stream.

3.3 Training Data

A total of 3430 (448 neutral, 448 sad, 802 happy, 284 angry, 422 disgusted, 423 fearful, and 603 surprised) images obtained from the Extended Cohn-Kanade dataset (CK+) [13] were used to train various Fisherface models. Images were aligned and cropped according to the frame capture and preprocessing algorithm discussed in section 3.2.2.

3.4 Model Descriptions

A series of 21 one-versus-one Fisherface models were built using a subset of the training data by pairing all possible combinations of the seven basic emotions.

A series of 7-class models (M1) were built using the complete set of training data representing all seven basic emotions. An initial probe into the effect of pixel dimensions

on model performance was conducted at 100x100, 50x50, and 25x25. Fisherface models were subsequently trained at pixel dimensions ranging from 6x6 pixels to 50x50 pixels.

A series of 6-class models (M2) were built at various pixel dimensions using a subset of the training data representing six of the seven basic emotions with the exclusion of happy. Happy was excluded as it experienced high recognition rates in our 7-class model and we were interested in focusing our efforts on other more poorly predicted emotions.

A series of 5-class models (M3) were built at various pixel dimensions using a subset of the training data representing five of the seven basic emotions with the exclusion of happy and surprised. Happy and surprised were excluded as these were the emotional expressions that experienced the highest recognition rates in our 7-class model. Again, we were interested in focusing our efforts on other more poorly predicted emotions.

Three 5-class models (M4) were built using five of the seven basic emotions with the exclusion of fearful and angry at pixel dimensions of 13x13 and 27x27 and 100x100.

A hierarchical model (H) was built by combining previously built useful models. This model is described in detail in section 4.6. Discussion of the construction of H has been deferred as knowledge of the results of individual models is necessary for its understanding.

3.5 Testing Data

Our research included testing from two distinct datasets. Test Set #1 (TS1) was obtained from the Karolinska Directed Emotional Faces Database [45]. TS1 was comprised of 140 images and contained 20 images for each of the seven basic emotions. Test Set #2 (TS2) was obtained from the Bosphorus Database [11]. TS2 was comprised of 100 images and contained 20 images of five of the basic emotions with the exclusion of fearful and angry.

Different databases were used for building and testing the models, making this a person-independent (PI) study. High accuracy rates for emotion classification have been achieved using a person-dependent test set (i.e. the test images are of individuals depicted in another image in the training set) [26] [77] [78] [79] but the accuracy rates for PI test sets are typically much lower [31] [47].

3.6 Implementation

All software was implemented in C++ and Python, and leveraged functionality available in the Open Source Computer Vision (OpenCV) libraries [52]. Three different systems were used during development and testing. Cameras used were the Logitech HD Pro Webcam, the RaspiCam and a built in laptop camera. Operating Systems used were Debian, Ubuntu and Windows 8. Hardware used was a Raspberry Pi, a Lenovo K450e Desktop PC and a Samsung ATIV Book 9 Plus.

4 Chapter: Results

4.1 One-versus-One

Some of the one-versus models achieved compartmentalized success on TS1. Interesting observations from this series of classifiers are discussed in the following paragraphs.

The angry versus fearful (AvF) model achieved a very high recall rate for fearful but a very low recall rate for angry, however, the precision for the angry class was higher than that of the fearful class. This classifier could prove useful for the detection of a fearful expression in the cases where we do not want to miss any positive samples. This classifier could also prove useful in the detection of an angry expression if we are willing to miss potential instances as a trade-off for precision. Details are presented in Table 1.

	A	F	RECALL	F-MEASURE
A	8	12	40.00%	53.33%
F	2	18	90.00%	72.00%
			65.00%	
PRECISION	80.00%	60.00%	70.00%	62.67%

Table 1 AvF built at 100x100 and Tested on 40 images (20 per emotion)

The TS1 images for the five remaining emotions (H, D, SU, N and S) were also tested against AvF. All of the happy and disgusted images were classified as fearful. 16 out of 20 surprised and 13 out of 20 sad images were also labelled as fearful. The most interesting finding with this model was observed within the neutral category where 15 out of 20 neutral images were classified as angry. This is a notable finding as our results suggest that this model has a higher likelihood than not to label an image as fearful. AvF may prove useful in helping to identify neutral images by functioning as a weak classifier for this expression category.

The angry versus surprised (AvSU) model achieved a high recall rate (90%) for surprised and a modest recall rate (70%) for angry. Details are presented in Table 2.

	A	SU	RECALL	F-MEASURE
A	14	6	70.00%	77.78%
SU	2	18	90.00%	81.82%
			80.00%	
PRECISION	87.50%	75.00%	81.25%	79.80%

Table 2 AvSU built at 100x100 and Tested on 40 images (20 per emotion)

Disgusted, neutral and sad images were more likely to be classified as angry (D = 16/20; N = 17/20; S = 13/20). Happy and fearful images were more likely to be classified as surprised (H = 14/20; F = 15/20). These results suggest that AvSU may prove useful in

distinguishing surprised, happy and fearful expressions from angry, disgusted, neutral and sad expressions.

The fearful versus surprised (FvSU) model achieved a very high recall rate (95%) for surprised but appeared to be ineffective at classifying fearful test images. Details are presented in Table 3.

	F	SU	RECALL	F-MEASURE
F	9	11	45.00%	60.00%
SU	1	19	95.00%	76.00%
			70.00%	
PRECISION	90.00%	63.33%	76.67%	68.00%

Table 3 FvSU built at 100x100 and Tested on 40 images (20 per emotion)

The most interesting observations for this model occurred when the remaining five emotion categories (H, D, N, S and A) were tested. A majority of these images were classified as fearful (H = 19/20; D = 19/20; N = 13/20; S = 15/20; A = 15/20). These results suggest that FvSU may prove useful as a detector for the surprised expression.

The disgusted versus surprised (DvSU) model achieved very high results (90%) for all of our metrics. Details are presented in Table 4.

	D	SU	RECALL	F-MEASURE
D	18	2	90.00%	90.00%
SU	2	18	90.00%	90.00%
			90.00%	
PRECISION	90.00%	90.00%	90.00%	90.00%

Table 4 DvSU built at 100x100 and Tested on 40 images (20 per emotion)

DvSU classified 15 out of 20 angry test images as disgusted and 14 out 20 fearful test images as surprised. These results suggest that DvSU may also prove useful as weak classifier for angry and fearful.

The neutral versus surprised (NvSU) model performed well overall. Details are presented in Table 5.

	N	SU	RECALL	F-MEASURE
N	17	3	85.00%	85.00%
SU	3	17	85.00%	85.00%
			85.00%	
PRECISION	85.00%	85.00%	85.00%	85.00%

Table 5 NvSU built at 100x100 and Tested on 40 images (20 per emotion)

NvSU also classified sad (17/20), disgusted (17/20) and angry (16/20) as neutral. This classifier could be used to separate surprised from a group of multiple different emotions.

The neutral versus disgusted (NvD) model also performed well during testing. Details are presented in Table 6.

	N	D	RECALL	F-MEASURE
N	20		100.00%	90.91%
D	4	16	80.00%	88.89%
			90.00%	
PRECISION	83.33%	100.00%	91.67%	89.90%

Table 6 NvD built at 100x100 and Tested on 40 images (20 per emotion)

NvD was also more likely to classify surprised (18/20), happy (17/20), sad (17/20), and fearful (16/20) as neutral suggesting that this model could be used to distinguish disgusted from a number of different emotional expressions.

The neutral versus fearful (NvF) model displayed a high recall rate for neutral but underperformed for fearful. Details are presented in Table 7.

	N	F	RECALL	F-MEASURE
N	18	2	90.00%	76.60%
F	9	11	55.00%	66.67%
			72.50%	
PRECISION	66.67%	84.62%	75.64%	71.63%

Table 7 NvF built at 100x100 and Tested on 40 images (20 per emotion)

NvF also classified all (20/20) disgusted test images as neutral and most (18/20) happy test images as fearful. The results suggest that NvF may prove useful in distinguishing a happy expression from a neutral or disgusted one.

The sad versus disgusted (SvD) model performed quite well overall on the KDEF test set. Details are presented in Table 8.

	S	D	RECALL	F-MEASURE
S	15	5	75.00%	78.95%
D	3	17	85.00%	80.95%
			80.00%	
PRECISION	83.33%	77.27%	80.30%	79.95%

Table 8 SvD built at 100x100 and Tested on 40 images (20 per emotion)

SvD classified 14/20 surprised and 14/20 neutral test images as sad. This model also classified 16/20 happy images as disgusted. These results suggest that SvD may prove useful in differentiating disgusted and happy from sad, surprised and neutral.

The sad versus surprised (SvSU) model displayed promising results for both emotions. Details are presented in Table 9.

	S	SU	RECALL	F-MEASURE
S	16	4	80.00%	80.00%
SU	4	16	80.00%	80.00%
			80.00%	
PRECISION	80.00%	80.00%	80.00%	80.00%

Table 9 SvSU built at 100x100 and Tested on 40 images (20 per emotion)

SvSU classified 16/20 disgusted images and 15/20 neutral images as sad. This model also labelled 18/20 happy images as surprised. The results suggest that SvSU may be helpful in distinguishing happy and surprised expressions from sad, neutral and disgusted ones.

The sad versus fearful (SvF) model achieved a score of 70% on the KDEF test set across the board. Details are presented in Table 10.

	S	F	RECALL	F-MEASURE
S	14	6	70.00%	70.00%
F	6	14	70.00%	70.00%
			70.00%	
PRECISION	70.00%	70.00%	70.00%	70.00%

Table 10 SvF built at 100x100 and Tested on 40 images (20 per emotion)

This model classified most of the other test images as fearful (H = 20/20; D = 20/20; SU = 16/20; A = 14/20). SvF also labelled 15/20 neutral images as sad. These results suggest that SvF may be useful in distinguishing sad and neutral images from other basic emotional expressions.

The sad versus happy (SvH) model achieved a recall rate of 100% for happy test images and 70% for sad test images. Details are presented in Table 11.

	S	H	RECALL	F-MEASURE
S	14	6	70.00%	82.35%
H		20	100.00%	86.96%
			85.00%	
PRECISION	100.00%	76.92%	88.46%	84.65%

Table 11 SvH built at 100x100 and Tested on 40 images (20 per emotion)

SvH also labelled 15/20 neutral test images as sad. These results suggest that this model may prove useful in distinguishing happy expressions from sad and neutral ones.

4.2 M1 @ 100x100, 50x50, and 25x25

TS1 was used in order to evaluate M1's performance at 100x100, 50x50, and 25x25. M1's overall performance was highest at 25x25 suggesting that this model is most effective at smaller pixel dimensions. Consequently, a thorough investigation into the impact of pixel dimensions from 6x6 to 50x50 was completed in order to ascertain the optimal pixel dimensions for this Fisherface model. Details are presented in Tables 12, 13 and 14.

	H	SU	S	N	A	F	D	RECALL	F-MEASURE
H	15		1		1	1	2	75.00%	68.18%
SU		3	5	5	3	1	3	15.00%	18.18%
S	1	1	7	2	5	1	3	35.00%	25.93%
N	2	2	9	5		1	1	25.00%	28.57%
A	2	4	2	3	3	2	4	15.00%	17.14%
F		3	8		1	3	5	15.00%	18.75%
D	4		2		2	3	9	45.00%	38.30%
								32.14%	
PRECISION	62.50%	23.08%	20.59%	33.33%	20.00%	25.00%	33.33%	31.12%	30.72%

Table 12 M1 built at 100x100, tested against TS1

	H	SU	S	N	A	F	D	RECALL	F-MEASURE
H	15		2	2		1		75.00%	56.60%
SU	1	7	4	7	1			35.00%	38.89%
S	5	1	7	2	2	2	1	35.00%	29.17%
N	1	1	8	8	1	1		40.00%	36.36%
A	6	2		3	4	3	2	20.00%	25.00%
F	3	4	4	1		7	1	35.00%	40.00%
D	2	1	3	1	4	1	8	40.00%	50.00%
								40.00%	
PRECISION	45.45%	43.75%	25.00%	33.33%	33.33%	46.67%	66.67%	42.03%	39.43%

Table 13 M1 built at 50x50, tested against TS1

	H	SU	S	N	A	F	D	RECALL	F-MEASURE
H	20							100.00%	85.11%
SU		12	4	3		1		60.00%	54.55%
S	1	1	9	4	2	3		45.00%	40.91%
N	1	1	4	11	1	1	1	55.00%	53.66%
A	1	2	4	1	5	4	3	25.00%	34.48%
F	2	7	3	1		5	2	25.00%	25.64%
D	2	1		1	1	5	10	50.00%	55.56%
								51.43%	
PRECISION	74.07%	50.00%	37.50%	52.38%	55.56%	26.32%	62.50%	51.19%	49.99%

Table 14 M1 built at 25x25, tested against TS1

4.3 M1 from 6x6 to 50x50

TS1 was used in order to test the efficacy of M1 ranging in pixel dimensions from 6x6 to 50x50. Overall this model displayed the highest performance at 13x13 pixels. Details are presented in Figure 3 for M1's overall f-measure performance and Figure 4 for emotion-specific performance.

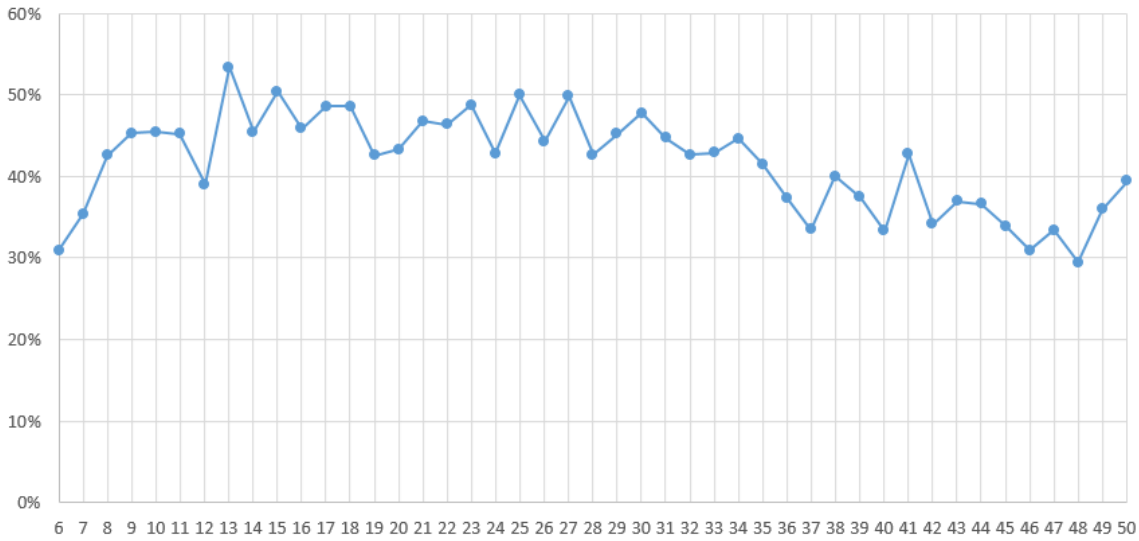

Figure 3 M1's Overall F-Measure Performance (y-axis) by Pixel Dimensions (x-axis)

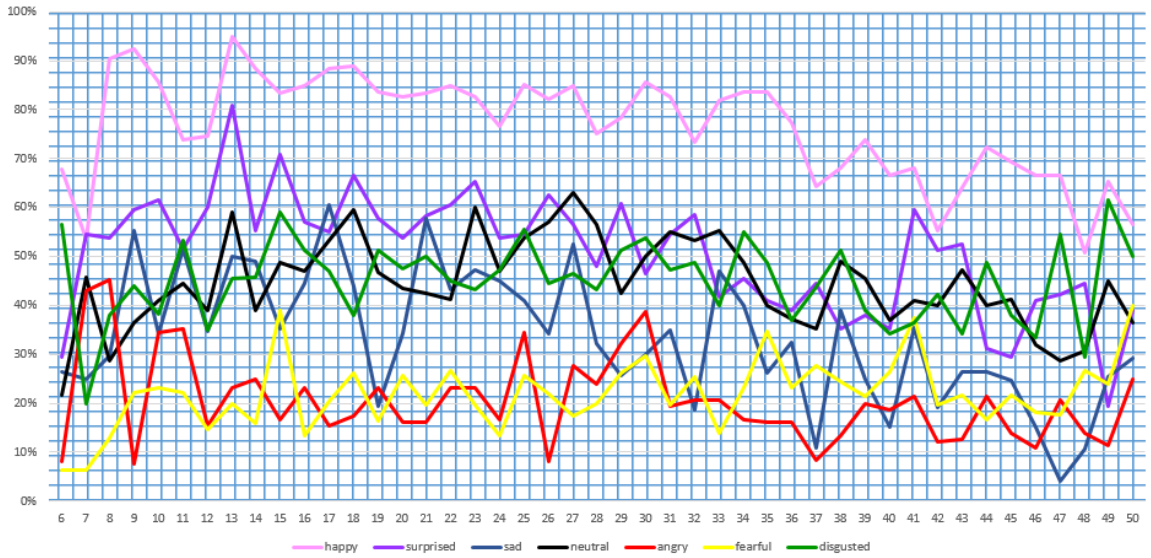

Figure 4 M1's Emotion Specific F-Measure Performance (y-axis) by Pixel Dimensions (x-axis)

The best classification results were observed for "happy" with an F-measure peaking at 95% for this model. This particular model also performed quite well for "surprised" with an F-measure reaching almost 81%. M1 showed some capability for classifying "sad", "neutral", and "disgusted" but was not very adept at detecting "angry" or "fearful" regardless of pixel dimensions. Details are presented in Table 15.

	H	SU	S	N	A	F	D	RECALL	F-MEASURE
H	19					1		95.00%	95.00%
SU		17		1		1	1	85.00%	80.95%
S		1	11	2	1	3	2	55.00%	50.00%
N		1	2	13	1	1	2	65.00%	59.09%
A		2	3	4	3	3	5	15.00%	23.08%
F	1	1	5	4	1	4	4	20.00%	20.00%
D			3			7	10	50.00%	45.45%
								55.00%	
PRECISION	95.00%	77.27%	45.83%	54.17%	50.00%	20.00%	41.67%	54.85%	53.37%

Table 15 M1 built at 13x13, tested against TS1

4.4 M2 and M3

A second (M2) and third (M3) series of classifiers were built at various pixel dimensions using a subset of potential training images. The second series excluded "happy" images, while the third series excluded both "happy" and "surprised" images. This strategy was used as an attempt to simplify the problem space and focus model performance on those emotions that were poorly detected by M1. M2 and M3 performed very similarly to M1. Minor performance improvements were observed in certain areas however minor diminishments in performance were observed in other areas. Overall, M2 and M3 performed no better or no worse than M1.

4.5 M4

Three 5-class classifiers were built using all potential training images with the exclusion of fearful and angry at pixel dimensions of 100x100, 27x27 and 13x13. Overall performance was lowest at the highest pixel dimensions and highest and the lowest pixel dimensions. We are unable to comment with any type of certainly why this occurred, however, one potential cause could be noise within the model. Perhaps it is the case that our 13x13 model contains sufficient data for classification, and that we are adding more noise than information when we increase pixel dimensions past that point. Details are presented in Tables 16, 17 and 18.

	H	SU	S	N	D	RECALL	F-MEASURE
H	19		1			95.00%	67.86%
SU		5	11	3	1	25.00%	38.46%
S	4		11	4	1	55.00%	39.29%
N	1	1	11	7		35.00%	38.89%
D	12		2	2	4	20.00%	30.77%
						46.00%	
PRECISION	52.78%	83.33%	30.56%	43.75%	66.67%	55.42%	43.05%

Table 16 M4 built at 100x100, tested against TS1

	H	SU	S	N	D	RECALL	F-MEASURE
H	18	1			1	90.00%	90.00%
SU		13	5	1	1	65.00%	70.27%
S			12	8		60.00%	54.55%
N		1	6	11	2	55.00%	52.38%
D	2	2	1	2	13	65.00%	70.27%
						67.00%	
PRECISION	90.00%	76.47%	50.00%	50.00%	76.47%	68.59%	67.49%

Table 17 M4 built at 27x27, tested against TS1

	H	SU	S	N	D	RECALL	F-MEASURE
H	19				1	95.00%	90.48%
SU	1	15	1	2	1	75.00%	78.95%
S		2	10	5	3	50.00%	55.56%
N		1	2	12	5	60.00%	61.54%
D	2		3		15	75.00%	66.67%
						71.00%	
PRECISION	86.36%	83.33%	62.50%	63.16%	60.00%	71.07%	70.64%

Table 18 M4 built at 13x13, tested against TS1

4.6 Hierarchical Model Description

By capitalizing on the strengths of previously built models, we were able to build a higher performing hierarchical model. This hierarchical model focused on five classes – happy, neutral, sad, disgusted and surprised. The fearful and angry classes were omitted as very little success was achieved with those classes in previous testing.

We offer two versions of the model – Model A and Model B. Model A classifies an input image into one of four prototypic emotions plus "neutral". Model B attempts to classify an input image into one of four prototypic emotions plus "neutral" but also includes an "undecided" classification.

Hierarchical Model methodology:

1. Test input image against M1 at 13x13. If the model predicts "happy" or "surprised", then label input image as such, else go to Step #2.
2. Test input image against AvF, DvA, DvSU, SvD, NvD, NvSU and M1 @ 49x49. If more than four of the models predict "disgusted", then label input image as such, else go to Step #3.
3. Test input image against SvF. If SvF predicts "sad", go to Step #4, else go to Step #5.

4. Test input image against AvF, M1 at 17x17, 21x21, 27x27, and M2 at 18x18 and 23x23. M1 @ 27x27 vote is weighted by a factor of 2 with the remaining classifiers getting a single vote. The classifiers vote on whether the input image is "sad" or "neutral". In the case of a tie, the input image is labelled as neutral.
5. Model A: Test input image against M4 at 13x13. Label input image with M4's prediction. Model B: Label input image as "undecided".

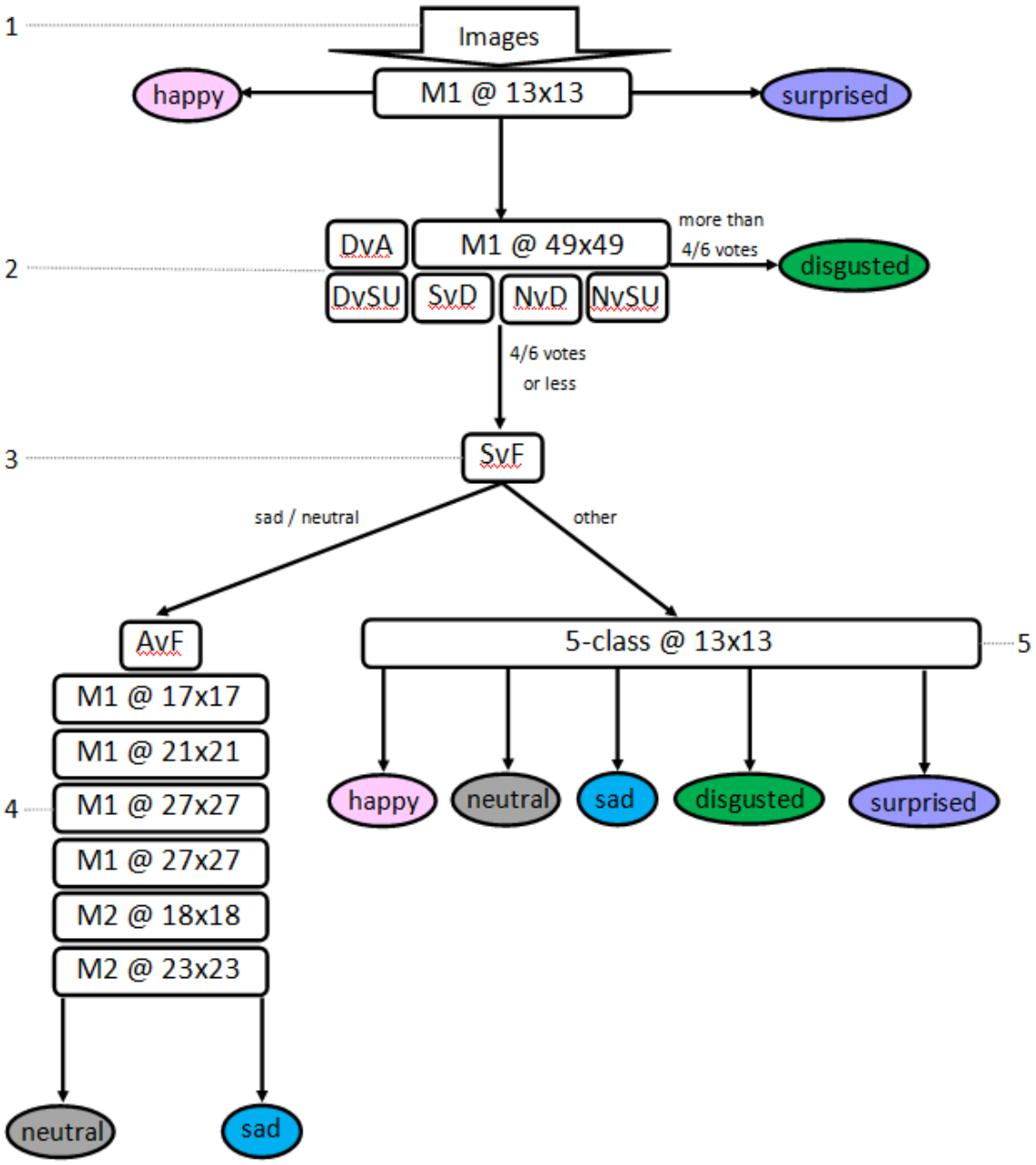

Figure 5 Hierarchical Model A

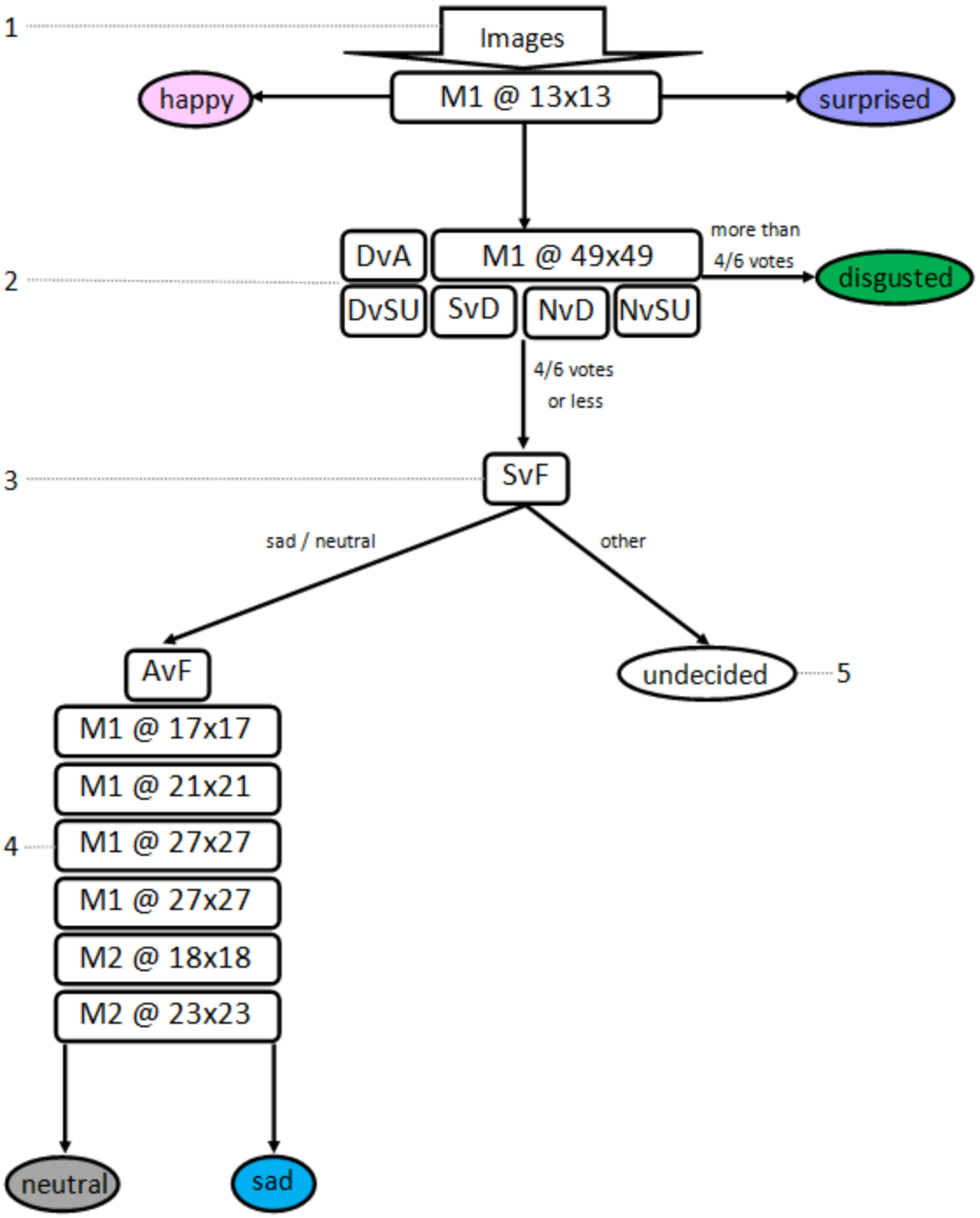

Figure 6 Hierarchical Model B

4.7 Hierarchical Model Results

Hierarchical Model A was tested against TS1 and TS2. Model performance was very high for the "happy" class and high for the "neutral" class for both test sets. The model performed well for "surprised" and "disgusted" against TS1 but displayed less promising results for TS2. The precision for the "sad" class dropped dramatically from TS1 to TS2. Details are presented in Tables 19 and 20.

	H	SU	S	N	D	RECALL	F-MEASURE
H	20					100.00%	100.00%
SU		17	1	2		85.00%	82.93%
S		1	12	5	2	60.00%	66.67%
N		1	3	16		80.00%	72.73%
D		2		1	17	85.00%	87.18%
						82.00%	
PRECISION	100.00%	80.95%	75.00%	66.67%	89.47%	82.42%	81.90%

Table 19 Hierarchical Model A, tested against TS1

	H	SU	S	N	D	RECALL	F-MEASURE
H	18	2				90.00%	83.72%
SU		10	4	6		50.00%	55.56%
S	2	2	11	5		55.00%	51.16%
N		1	3	15	1	75.00%	62.50%
D	3	1	5	2	9	45.00%	60.00%
						63.00%	
PRECISION	78.26%	62.50%	47.83%	53.57%	90.00%	66.43%	62.59%

Table 20 Hierarchical Model A, tested against TS2

5 Chapter: Discussion

Our results indicated that a hierarchical classification system for emotion recognition, comprised of Fisherface models with differing strengths, can outperform a single stand-alone Fisherface classifier. By leveraging strengths of individual models and combining these models strategically, we were able to improve classification performance.

The happy and neutral classes performed well in the hierarchical model with recall remaining high across both test sets. A machine with the ability to recognize and distinguish between the happy and neutral emotional states has many useful applications. One such application comes from embedding this capability within a child's toy. The toys behaviors (e.g. movements, sounds, light patterns and visualizations) could be geared toward each individual child through a semi-supervised learning reinforcement schedule based on what the child finds titillating and amusing. In short, the toy could learn which actions to take in order to elicit happiness in the child. This could be particularly useful in the case of someone with special needs such as an autistic individual. Having a tool that knows how to reliably elicit happiness could prove invaluable in the situation where more traditional calming and soothing approaches fail. It would also offer an alternative coping strategy to exhausted and over-taxed family members and health care professionals. Another potential less altruistic application is in market research and advertising. Knowing what makes an audience happy could go a long way to promoting products and services and captivating target demographics.

The surprised and disgusted classes showed high performance in the hierarchical model for TS1 but overall performance dropped in both of these classes for TS2. It was unclear via visual inspection of the test sets what caused the discrepancy in performance. Potential causal factors for the disparity are subtle unobservable differences in lighting, expression intensity, or subject demographics. The one exception to the drop in performance is in the precision metric for disgusted. The hierarchical model achieved roughly 90% precision against both of our test sets for the disgusted classification. If we can rely on the model to be correct in its prediction of disgusted with such high precision, this knowledge could have many applications. One example is in marketing and advertising where the goal is to reduce the number of images or messages that viewers find disgusting or offensive. Conversely, this knowledge could also be used to increase the number of offensive images in order to persuade viewers to change negative behaviors such as drinking and driving, smoking and consuming drugs for recreational purposes. Another less desirable application could be in political campaigns or religious movements. If a given party used our hierarchical model to understand what actions, images, or topics elicited disgust in the voters and/or followers, that party could use this information in their propaganda.

Predictions for a sad expression in the hierarchical model were not as reliable as most of the other classes. Sad's highest metric was observed when the model was tested against TS1 with a precision of 75%, however precision dropped on TS2 to under 45%. Like the surprised and disgusted classes, it was unclear via visual inspection of the test sets what caused the discrepancy in performance. The hierarchical model could prove to be a good

starting point for deciphering a sad expression. If we were able to build other models that were also somewhat reliable at perceiving the sad class, we could employ a voting strategy for more robust interpretation of this emotional expression.

Promising results were achieved on both test sets with the hierarchical model, however those results could only be observed in the real time application under optimal lighting conditions. "Happy" was the highest performing class across both test sets, and appeared to be one of the most reliably predicted classes during real time testing sessions. It is difficult to report with any certainty the performance of the model in the real time setting. Lighting seemed to be the biggest variable but other factors such as background scenery might also have contributed to variations in performance. Another issue observed during real-time testing was the inability of the OpenCV eye detectors to detect the individual eyes when the participant was wearing glasses. Without proper detection of the left and right eye, certain portions of our alignment algorithms fail.

In our model and system evaluation we examined the most stringent form of validation in that our training and testing sets were comprised of distinct samples. Our overall results for Hierarchical Model A tested against TS1 rivaled results for person-dependent testing reported by previous researchers in the field [19] [29] [38]. Our overall results for Hierarchical Model A tested against TS1 surpassed success rates reported for person-independent testing [31] [47]. Our results for "happy" persisted when tested against TS2 suggesting that Fisherfaces is a very capable methodology for predicting this emotion in facial images. Additionally, the persistence of high precision rates for the prediction of

"disgusted" suggests that Fisherfaces may also be useful for the recognition of other emotions as well.

6 Chapter: Future Work

One direction for future work is to build the 1v1 models at different pixel dimensions such as those that proved successful (e.g. 13x13) in the multiclass models.

One could also investigate a potential correlation between the numbers of facial muscles involved in each expression class and the pixel dimensions that display the highest model performance for a given expression. For example, it may be the case that smaller pixel dimensions best represent emotions that require less facial muscle activity, and conversely larger pixel dimensions may best represent emotions that require more facial muscle activity.

An investigation into the impact of range and intensity of emotional expression may also prove fruitful. It may be the case that happy was better predicted because most people express happy very similarly to one another. It may also be the case that angry and fearful expressions were poorly predicted as these may differ more from person to person.

7 Chapter: Conclusion

We conducted a thorough investigation into the efficacy of the Fisherface model in predicting emotional expressions as depicted in facial images. We built and tested a variety of different Fisherface models. We considered the impact of pixel dimensions and problem complexity (i.e. number of classes), and evaluated stand-alone models as well as a compound hierarchical model. Our results suggest that the Fisherface model can be successful in recognizing human emotions in facial images. Our results also supported the concept of synergy suggesting that a compound hierarchical model that leverages the strengths of individual stand-alone models may be the most effective strategy.

8 Chapter: References

[1] K. Anderson, and P.W. McOwan, "A Real-Time Automated System for Recognition of Human Facial Expressions," *IEEE Trans. Systems, Man, and Cybernetics Part B,* vol. 36, no. 1, pp. 96-105, 2006.

[2] K. Bache, and M. Lichman, "UCI Machine Learning Repository", Irvine California, 2013.

[3] M.S. Bartlett, G. Littlewort, I. Fasel, and R. Movellan, "Real Time Face Detection and Facial Expression Recognition: Development and Application to Human Computer Interaction," *Proc. Workshop Computer Vision and Pattern Recognition for Human-Computer Interaction,* vol. 5, 2003.

[4] M. Bartlett, P. Viola, T, Sejnowski, B. Golomb, J. Larsen, J. Hager, and P. Ekman, "Classifying Facial Action," *Advances in Neural Information Processing Systems,* vol. 8, 1996.

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

[6] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence.* vol. 19, no. 7, pp. 711-720, 1997.

[7] V. Bettadapura, "Face Expression Recognition and Analysis: The State of the Art," *Tech Report,* 2012.

[8] M.J. Black, D. Fleet, and Y. Yacoob, "A Framework for Modeling Appearance Change in Image Sequences," *IEEE Comp. Society Press,* 1998.

[9] M.J. Black and Y. Yacoob, "Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion," *Int'l J. Computer Vision,* vol. 25, no. 1, pp. 23-48, 1997.

[10] M.J. Black and Y. Yacoob, "Tracking and Recognizing Rigid and Non-Rigid Facial Motions Using Local Parametric Models of Image Motions," *Proc. Int'l Conf. Computer Vision,* pp. 374-381, 1985.

[11] Bosphorus 3D Face Database. Available online: http://bosphorus.ee.boun.edu.tr/Home.aspx

[12] F. Bourel, C.C. Chibelushi, and A.A. Low, "Recognition of Facial Expressions in the Presence of Occlusion," *Proc. Conf. Machine Vision,* vol. 1, pp. 213-222, 2001.

[13] Cohn-Kanade AU-Coded Expression Database. Available online: http://www.pitt.edu/~emotion/ck-spread.htm

[14] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T. Huang, "Learning Bayesian Network Classifiers for Facial Expression Recognition Using Both Labeled and Unlabeled Data," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. I-595-I-604, 2003.

[15] I. Cohen, N. Sebe, A. Garg, L.S. Chen, and T.S Huang, "Facial Expression Recognition From Video Sequences Temporal and Static Modeling," *Computer Vision and Image Understanding,* vol. 91, pp. 160-187, 2003.

[16] C. Darwin, "The Expression of the Emotions in Man and Animals," *J. Murray,* London, 1872.

[17] Dimensionality Reduction. Available online: https://en.wikipedia.org/wiki/Dimensionality_reduction

[18] F. Dornaika, and F. Davoine, "Simultaneous Facial Action Tracking and Expression Recognition in the Presence of Head Motion," *Int. J. Computer Vision,* vol. 76, no. 3, pp. 257-281, 2008.

[19] G. J. Edwards, T.F. Cootes, and C.J. Taylor, "Face Recognition Using Active Appearance Models," *Proc. European Conf. Computer Vision,* vol. 2, pp. 581-695, 1998.

[20] P. Ekman, and W.V. Friesen," Constants across Cultures in The Face and Emotion," *J. Personality Social Psychol.,* vol. 17, no. 2, pp. 124-129, 1971.

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

[22] P. Ekman, and W.V. Friesen, *Facial Action Coding System (FACS): Manual.* Palo Alto: Consulting Psychologists Press, 1978.

[23] P. Ekman, *Emotion in the Human Face.* Cambridge Univ. Press, 1982.

[24] P. Ekman, "Strong Evidence for Universals in Facial Expressions: A Reply to Russell's Mistaken Critique," *Psychology Bulletin,* vol. 115, no. 2, pp. 268-287, 1994.

[25] I. Essa, and A. Pentland, "Coding, Analysis, Interpretation and Recognition of Facial Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence.* vol. 19, no. 7, pp. 757-763, 1997.

[26] I. Essa, and A. Pentland, "Facial Expression Recognition Using A Dynamic Model and Motion Energy," *IEEE Proc. Int'l Conf. Computer Vision,* pp. 360-367, 1995.

[27] B. Fasel, and J. Luettin, "Automatic Facial Expression Analysis: A Survey," *J. Pattern Recognition,* vol. 36, pp. 259-275, 2003.

[28] Sir R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics,* vol. 7, pp. 179-188, 1936.

[29] S.L. Happy, and A. Routray, "Automatic Facial Expression Recognition Using Features of Salient Facial Patches," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 1-12, 2015.

[30] Histogram Equalization. Available online: www.math.uci.edu/icamp/courses/math77c/demos/hist_eq.pdf

[31] H. Hong, H. Neven, and C. von der Malsberg, "Online Facial Expression Recognition Based on Personalized Galleries," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 354-359, 1998.

[32] C.L. Huang and Y.M Huang, "Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters," *J. Visual Comm. And Image Representation,* vol. 8, no. 3, pp. 278-290, 1997.
[33] C.E Izard, "Innate and Universal Facial Expressions: Evidence from Developmental and Cross-Cultural Research," *Psychology Bulletin,* vol. 115, no. 2, pp. 288-299, 1994.
[34] JAFFE Face Database. Available online: http://www.kasrl.org/jaffe.html
[35] Y. Ji, and K. Idrissi, "Automatic Facial Expression Recognition Based on Spatiotemporal Descriptors," *Pattern Recognition Letters,* vol. 33, pp. 1373-1380, 2012.
[36] W.R. Klecka, "Discriminant analysis," No. 19, Sage, 1980.
[37] H. Kobayashi and F. Hara, "Facial Interaction between Animated 3D Face Robot and Human Beings," *Proc. Int'l Conf. Systems, Man, Cybernetics,* pp. 3,732-3,737, 1997.
[38] H. Kobayashi and F. Hara, "Recognition of Six Basic Facial Expressions and Their Strength by Neural Network," *Proc. Int'l Workshop Robot and Human Comm.,* pp. 381-386, 1992.
[39] H. Kobayashi and F. Hara, "Recognition of Mixed Facial Expressions by Neural Network," *Proc. Int'l Workshop Robot and Human Comm.,* pp. 387-391, 1992.
[40] I. Kotsia, I. Buciu, and I. Pitas, "An Analysis of Facial Recognition under Partial Facial Image Occlusion," *Image and Vision Computing,* vol. 26, no. 7, pp. 1052-1067, 2008.
[41] I. Kotsia, and J. Pitas, "Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines," *IEEE Trans. Image Processing,* vol. 16, no.1, pp. 172-187, 2007.
[42] Labeled Dataset. Available online: www.stackoverflow.com/questions/19170603/what-is-the-difference-between-labeled-and-unlabeled-data
[43] Linear Discriminant Analysis. Available online: https://en.wikipedia.org/wiki/Linear_discriminant_analysis
[44] B.D. Lucas, and T. Kanade, "An Iterative Image Registration Technique with and Application to Stereo Vision," *Proc. Int'l Conf. Artificial Intelligence,* pp. 674-679, 1981.
[45] D, Lundqvist, A. Flykt, and A. Ohman, "The Karolinska Directed Emotional Faces," 1998.
[46] M.J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 200-205, 1998.
[47] M.J. Lyons, J. Budynek, and S. Akamatsu, "Automatic Classification of Single Facial Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 12, pp. 1,357-1,362, 1999.

[48] A. Majumder, B. Laxmidhar, and V.K. Subramanian, "Emotion Recognition From Geometric Facial Features Using Self-Organizing Map," *Pattern Recognition,* vol. 47, pp. 1282-1293, 2014.

[49] A. Mehrabian, "Communication without Words," *Psychology Today,* vol. 2, no. 4, pp. 53-56, 1968.

[50] P. Michel, and R. Kaliouby, "Real Time Facial Expression Recognition in Video Using Support Vector Machines," *Proc. Conf. Multimodal Interfaces,* pp. 258-264, 2003.

[51] D.A Norman, "Emotion & Attractive," *Interactions,* vol. 9, no. 4, pp. 36-42, 2002.

[52] Open Source Computer Vision Library. Available online: http://opencv.org 2013

[53] T. Otsuka and J. Ohya, "Extracting Facial Motion Parameters by Tracking Feature Points," *Proc. Int'l Conf. Adv. Multimedia Content Processing,* pp. 442-453, 1998.

[54] T. Otsuka and J. Ohya, "Recognition of Facial Expressions Using HMM with Continuous Output Probabilities," *Proc. Int'l Workshop Robot and Human Comm.,* pp.323-328, 1996.

[55] T. Otsuka and J. Ohya, "Spotting Segments Displaying Facial Expression from Image Sequences Using HMM," *Proc. Int'l Conf. Automation Face and Gesture Recognition,* pp. 442-447, 1998.

[56] C. Padgett, G.W. Cottrell, and R. Adolphs, "Categorical Perception in Facial Emotion Classification," *Proc. Conf. Cog. Sci. Society,* 1996.

[57] C. Padgett and G.W. Cottrell, "Representing Face Images for Emotion Classification," *Proc. Conf. Advances in Neural Information Processing Systems,* pp. 894-900, 1996.

[58] M. Pantic, and I. Patras, "Detecting Facial Actions and Their Temporal Segments in Nearly Frontal-View Image Sequences," *Proc. IEEE Conf. Systems, Man, and Cybernetics,* vol. 4, pp. 3358-3363, 2005.

[59] M. Pantic, and I. Patras, "Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments Form Face Profile Image Sequences," *IEEE Trans. Systems, Man, and Cybernetics Part B,* vol. 36, no. 2, pp. 443-449, 2006.

[60] M. Pantic, and J.M Rothkrantz, "Facial Action Recognition for Facial Expression Analysis From Static Face Images," *IEEE Trans. Systems, Man arid Cybernetics Part B,* vol. 34, no. 3, pp. 1449-1461, 2004.

[61] M. Pantic, and L.J.M Rothkrantz, "Automatic Analysis of Facial Expressions: The State of The Art," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1424-1445, 2000.

[62] M. Pantic, and L.J.M. Rothkrantz, "Expert System for Automatic Analysis of Facial Expression," *Image and Vision Computing J.,* vol. 18, no. 11, pp. 881-905, 2000.

[63] M. Pantic, and L.J.M. Rothkrantz, "An Expert System for Multiple Emotional Classification of Facial Expressions," *Proc. Int'l Conf. Tools with Artificial Intelligence,* pp. 113-120, 1999.
[64] W.G. Parrott, "*Emotions in Social Psychlogy*," Psychology Press, 2000.
[65] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," *Proc. Computer Vision and Pattern Recognition,* pp. 84-91, 1994.
[66] R.W. Picard, "Toward Machines With Emotional Intelligence," *MIT Media Laboratory,* 2001.
[67] J.A. Russell, "Is There Universal Recognition of Emotion From Facial Expressions? A Review of the Cross-Cultural Studies," *Psychology Bulletin,* vol. 115, no. 1, pp. 102-141, 1994.
[68] H. Schneiderman, and T. Kanade, "Probalistic Modeling of Local Appearance and Spatial Relationships for Object Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp.45-51, 1998.
[69] N. Sebe, M.S. Lew, Y. Sun, I. Cohen, T. Gevers, and T.S Huang, "Authentic Facial Expression Analysis," *Image and Vision Computing,* vol. 25, pp. 1856-1863, 2007.
[70] L. Sirovich, and M. Kirby, "Low-Dimensional Procedure for the Characterization of Human Faces," *J. Opt. Soc. Am. A.,* vol. 4, no. 3, 1987.
[71] M. Suwa, N. Sugie, and K. Fujimora, "A Preliminary Note on Pattern Recognition of Human Emotional Expression," *Proc. Int'l Conf. Pattern Recognition,* pp. 408-410, 1978.
[72] H. Tao, and T.S Huang, "A Piecewise Bezier Deformation Model and Its Application in Facial Motion Capture," *Advances in Image Processing and Understanding,* 2002.
[73] Y. Tian, T. Kanade, and J. Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 2, pp. 97-115, 2001.
[74] C. Tomasi, and T. Kanade, "Detection and Tracking Point Features," *Carnegie Mellon University Technical Report CMU-CS-91-132,* 1991.
[75] P. Viola, and M.J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," *Computer Vision and Pattern Recognition,* vol. 1, pp. I-511-I-518, 2001.
[76] P. Viola, and M.J. Jones, "Robust Real-Time Object Detection," *Int. J. Computer Vision,* vol. 57, no. 2, pp. 137-154, 2004.
[77] M. Wang, Y. Iwai, and M. Yachida, "Expression recognition from Time-Sequential Facial Images by Use of Expression Change Model," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 324-329, 1998.
[78] M. Yoneyama, Y. Iwano, A. Ohtake, and K. Shirai, "Facial Expressions Recognition Using Discrete Hopfield Neural Networks," *Proc. Int'l Conf. Information Processing,* vol.3, pp. 117-120, 1997.

[79] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between Geometry-Based and Gabor Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron." *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 454-459, 1998.

[80] J. Zhao, and G. Kerney, "Classifying Facial Emotions by Backpropagation Neural Networks with Fuzzy Inputs," *Proc. Int'l Conf. Neural Information Processing,* vol. 1, pp. 454-457, 1996.

[81] W. Zheng, X. Zhou, C. Zou, and L. Zhao, "Facial Expression Recognition Using Kernel Correlation Analysis (KCCA)," *IEEE Trans. Neural Networks,* vol. 17, no. 1, pp. 233-238, 2006.