Experiments in Expression Recognition

by

James P. Skelley

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 16, 2005

Author . . . .
Department of Electrical Engineering and Computer Science
August 16, 2005

Certified by . . . .
Tomaso Poggio
Eugene McDermott Professor in the Brain Sciences, CBCL
Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **16 AUG 2005**	2. REPORT TYPE	3. DATES COVERED **00-08-2005 to 00-08-2005**

4. TITLE AND SUBTITLE **Experiments in Expression Recognition**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Department of Electrical Engineering and Computer Science,Cambridge,MA,02139**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**41**	

Experiments in Expression Recognition

by

James P. Skelley

Submitted to the Department of Electrical Engineering and Computer Science
on August 16, 2005, in partial fulfillment of the
requirements for the degree of
Master of Engineering

Abstract

Despite the significant effort devoted to methods for expression recognition, suitable training and test databases designed explicitly for expression research have been largely neglected. Additionally, possible techniques for expression recognition within an Man-Machine-Interface (MMI) domain are numerous, but it remains unclear what methods are most effective for expression recognition. In response, this thesis describes the means by which an appropriate expression database has been generated and then enumerates the results of five different recognition methods as applied to that database. An analysis of the results of these experiments is given, and conclusions for future research based upon these results is put forth.

Thesis Supervisor: Tomaso Poggio
Title: Eugene McDermott Professor in the Brain Sciences, CBCL

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Man-Machine Interfaces (MMI) seek to eliminate the alienation users frequently experience when interacting with automated devices. To accomplish this, MMIs utilize information regarding the state of the client to govern the course of the interaction. This information may take to the form of physical cues, verbal cues, and information recorded during previous interactions with the user. As MMIs approach conversational levels on par with typical human-human interactions, however, it will be necessary that they also incorporate emotional cues into their repertoire. In particular, just as humans direct the course of a conversation via emotional cues, so will it be necessary for MMIs to recognize these cues during their own interactions with the user [1].

While emotional cues are expressed in a variety of forms (body temperature, blood acidity, breath rate), they are most frequently and most readily recognized via facial expressions. Providing MMIs with the ability to correlate facial expressions with emotional states would first require that a system be trained to readily acknowledge facial expressions at least as successfully as humans do. Only when these primitive recognitions are available will higher-level processing be possible.

1.1 Aims of Research

This thesis addresses the challenge posed by expression recognition in an MMI domain by first creating a database of images which may be used to train an MMI to recognize facial expressions. Then, building upon existing facial recognition techniques, this thesis investigates various approaches to facial expression recognition and ascertains the relative efficacy of each method. Ultimately, this research will provide a framework to direct future efforts using these techniques.

1.2 Thesis Organization

This thesis begins by describing the rationale for the creation and structure of the expression database utilized in this research. This includes a description of the processing methods involved in generating appropriate images. Subsequently, a discussion of the various methods utilized for facial extraction from the database are given. A description of each of the five expression recognition algorithms to be compared follows, after which the experimental results for each of the methods are described. The thesis concludes with a discussion of the relative success rates of each of the methods and recommendations for their improvement.

Chapter 2

Database Creation and Description

In [12], generating a comprehensive and labeled training database of expressions has been stated as one of the main challenges posed before automatic facial expression recognition. Because of the lack of appropriate databases, researchers in expression recognition frequently use databases developed for face recognition [14, 15] and person identification [17]. Unfortunately, these databases usually do not contain video sequences and also lack appropriate labeling information for the expressions. A second set of widely used expression databases stems from behavioral science [7]. These databases often contain specific facial muscle contractions but lack natural expressions. In any of the expression databases evaluated for use as a training set, it was evident that at least one of the following deficiencies was present:

- Poor image quality
- Too few expression samples were provided for each class
- The database did not include video sequences
- The database did not include natural expressions
- The labeling schema used was insufficient

These considerations prompted the construction of a database suitable for expression recognition research which specifically addressed the issues listed above.

2.1 Existing Expression Databases

Table 2.1 compares some of the most commonly used facial expression databases with the database generated for this thesis.[1] Because of their wide-usage, two of these databases will be discussed in more detail: the *Cohn-Kanade* database, as an example of a FACS-based database and the *Human Identification at Distance* database (HID) as an example for a database which was developed for person identification.

Name	**Images**	**Sequences**	**Subjects**	**Expressions**	**Remarks**
Thesis Database	original and processed	≥ 1400 RGB	12	played and natural	natural head movements
HID [17]	9 mug-shots per subj.	RGB	284	natural	expr. often subtle or mixed
Cohn-Kanade [7]	None	329 BW/RGB	100	played	FACS coded, no head movement
CMU-PIE [15]	≥ 40,000 RGB	includes some talking clips	68	played	large variations in pose and illum.
FERET [14]	≥ 14,000 BW	None	1009	played	designed for identification
JAFFE [11]	213 BW	None	10	played	only Japanese women

Table 2.1: Existing Expression Databases.

2.1.1 Cohn-Kanade Database

The Cohn-Kanade database can be considered today's de-facto standard for comparative studies in facial expression analysis. The subjects were instructed by the experimenter to activate a specific action unit (AU) or combinations of AUs; the recorded sequences were annotated using the FACS. The length of the sequences varies between 9 and 60 frames. Each of the sequences starts with the neutral expression and ends when the maximum intensity of the performed AU is reached. Two cameras recorded the face in frontal and half profile view. For now, only the frontal views have been made available to the public.

[1] Completeness is not claimed in this comparison - e.g. commercial databases have not been considered at all.

Using this database to train and test an expression recognizer for an MMI application involves a number of drawbacks:

- No natural expressions
- No head motion
- The clips don't show the full expression over time
- Many clips show expressions, or rather activities of AUs, which usually don't occur in real life

2.1.2 HID Database

A relatively new database has been recorded within the HID DARPA project [17]. This database was designed to test algorithms for person identification and for human motion analysis. It includes ten mug shots per subject and video sequences of the subjects speaking and walking. Of interest for expression analysis are the video clips which show the natural expressions of the subjects while watching a ten minute video. This video contained scenes from various movies and television programs intended to elicit different emotions. Based on the judgment of the experimenter, facial expressions where cut into five second clips and assigned one of the following eleven labels: happiness, sadness, fear, disgust, anger, puzzlement, laughter, surprise, boredom, disbelief or blank stare.

At the time of writing, no publicly available database which includes a fairly large number of annotated video clips of natural expressions has been encountered by the author. In the future, databases of this kind might be extremely helpful, perhaps essential, for evaluation algorithms with respect to their real-world applicability. Still, this database by itself is not sufficient for training and evaluating expression recognition systems for several reasons:

- Most of the natural expressions are very subtle and far beyond what is currently recognizable by state-of-the-art systems

- Some clips contain mixtures of expressions (e.g. puzzlement which turns into surprise) which are not reflected in the labels
- There are rarely more than two clips per subject and expression class.

2.2 Database Considerations

Several considerations must be made before creating an expression database. In particular, automated facial expression recognition within an MMI setting poses a number of problems that must be addressed by the database:

- (a) Pose invariance. Slight changes in the pose of the face are usually handled by applying an alignment phase prior to extracting the features for classification. Aligning a face to a canonical view requires the computation of correspondences between a reference image and the image to be aligned. The correspondences can be computed sparsely, i.e. by matching or tracking a few facial features (see e.g. [16]), or across large parts of the face (see e.g. [21]). This thesis makes use of an alignment algorithm which computes a dense correspondence map based on optical flow field (see Figure 2-2).
- (b) Exploiting temporal information. Important temporal context information is lost when single images are classified independently [13]. This problem can be dealt with on various levels of complexity. For example, the results of the independently classified images can be merged to classify a whole image sequence. A more sophisticated approach is to model state transitions in expressions by Hidden Markov Models (see e.g. [10]). This thesis performs each recognition experiment only on single images which were processed (see Figure 2-2) from their original temporal sequence.
- (c) The generalization problem - classifying the expressions of a person who has not been part of the training set. The expressions between people might vary due to differences in the shape and texture of their faces and differences in the

way they perform a facial expression. Texture dependencies can be removed by using optical flow as input to the classifier [3]. One method to remove the shape dependency involves performing the warp to a synthetic reference face (see Section 3.1) [18].

- (d) Subtleness of natural expressions. Often, natural expression show a degree of subtleness which is not present in databases of played expressions. This thesis provides a preliminary investigation of the effect of training a system on played expressions, and then testing on natural expressions. This mimics the requirements imposed upon an MMI system trained in an artificial environment and then deployed in a real-world setting.

2.2.1 Labeling Methodology

The expression labeling stratagem used in this thesis was chosen to address the particular aspects of the MMI domain described above. For both played and natural expressions labels used include: neutral, happy, surprise, fear, sadness, disgust, and anger. Within the natural expressions, particular expressions were expanded depending on the form they took. That is, happiness became smile and laugh, and the categories shock, puzzlement, dislike were added. The suitability of these expression categories within an MMI application may vary depending on the domain–additional expressions such as annoyed, confused or bored may be useful in some circumstances and should be considered in the creation of future databases.

2.3 Expression Recording Methodology

The facial expression database was specifically intended as a training and test database for expression recognition systems within an MMI setting. In this setting, the person will be close to the camera, either looking straight into the camera or at an object close by, e.g. a computer screen, or kiosk terminal. It is likely that head motion occurs in and out of the image plane when the user naturally interacts with a computer.

Thus, the head motion of each of the subjects in the database was not constrained during the recording of their expressions.

Robert Fischer [5] performed the video capture using two digital video cameras. Recordings were taken from both frontal and half-profile views (about 30° from center) of each of the subjects. The scene was illuminated by a ceiling light and a photographic lamp with a diffuser, which was placed behind the camera at a height of approximately 2 meters. Table 2.2 lists the specifications for the video-capture equipment used.

Camera	Sony DCR-VX2000
Video standard	NTSC
Color format	RGB
Frame rate	29.97 fps, non drop-frame
Frame size	720 × 480

Table 2.2: Specifications for Recording Apparatus.

2.3.1 Played and Natural Recordings

As mentioned previously, the database consists of two parts: played and natural expressions. In the first part of their recording session, subjects were asked to display six basic emotions: happiness, sadness, fear, disgust, surprise and anger. Data from these categories were each labeled as such and as being played, not natural. Intending to create a database demonstrating the widest gamut of expressions, this labeling schema was believed to be more suitable than a more granular degree or valence-based labeling. Furthermore, degree-based labeling can be ambiguous, considering the range of values, and unique responses provided by each of the subjects.

Most of the subjects had prior experience in acting. Each subject was asked to repeat each expression between 10 and 20 times in order to get a sufficiently large number of examples.

In order to obtain natural expressions, the subjects watched a twenty minute movie that consisted of short video clips. The clips covered a large variety of topics, such as funny commercials, scenes of surgery and war documentaries. Similar to the HID database, the labeling of the recorded natural expressions was done by the judgment of the experimenter (the current video clip watched by the subject likewise suggesting which emotions were most likely to arise). Whenever possible, single expressions were isolated and cut from neutral state to neutral state.

2.4 Statistical Overview of Database

Of the total of twelve subjects eight were female and four male. The age of the subjects ranged from eighteen to thirty years; three were of African, one of Asian and one of Indian ethnicity - the remainder were Caucasian. Altogether, 1407 sequences were captured from video, averaging 117 sequences per subject and 176 sequences per class. Figure 2-1 shows sample images of each of the subjects.

subject	100	101	102	103	104	105	106	107	500	501	502	503
Sex	f	f	f	f	f	f	f	f	m	m	m	m
Age	20	19	21	22	25	25	29	18	22	30	19	21
Ethnicity	Cauc	Cauc	African	Indian	Cauc	Cauc	Cauc	Asian	African	Cauc	Cauc	African

Table 2.3: Subject Background Information.

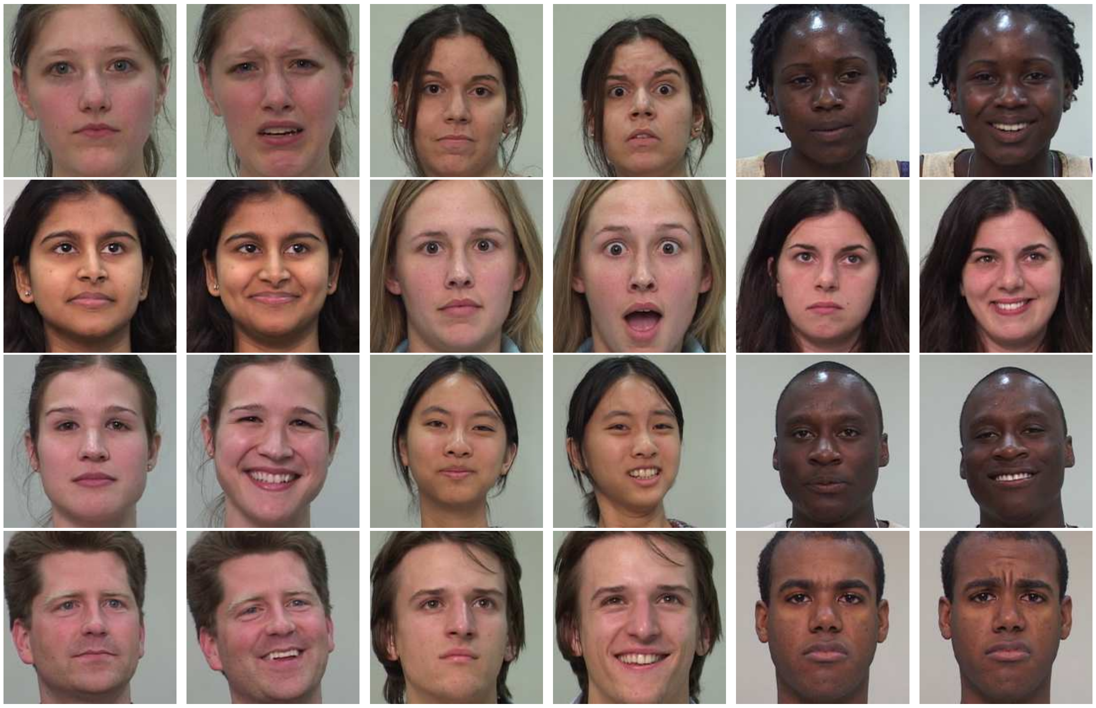
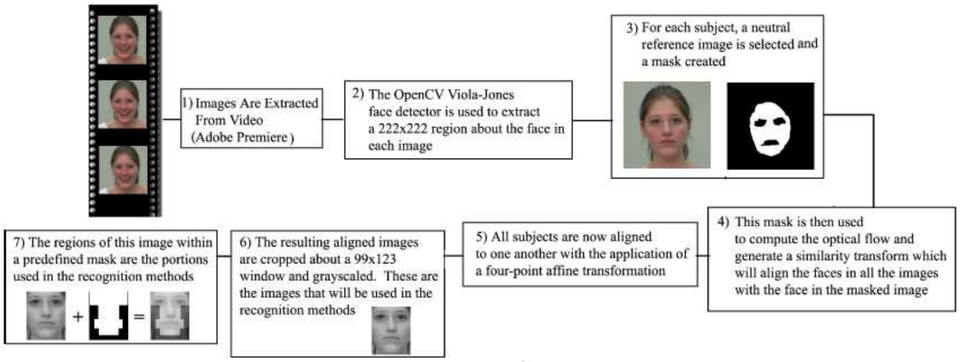


Figure 2-1: Sample Images of Each Subject.

Subjects in Figure 2-1 are given in pairs, from left to right, top to bottom, in the following order: 100, 101, 102, 103, 104, 105, 106, 107, 500, 501, 502, 503.

2.5 Alignment and Preprocessing

The initial labeling and video acquisition was accomplished in previous work by Robert Fischer [5]. Using these video sequences, images were generated and then processed independently for use in this thesis. Once the videos had been separated into image sequences it was necessary to extract only that portion of the image relevant to recognition in a controlled manner. The occasional error in cropping and alignment made it necessary to manually remove several sequences. Additionally, played expression sequences were edited by hand so that only those images demonstrating the expression under investigation were included in the database. This prevented intermediary neutral expressions in the video sequence from interfering with the training of the recognition methods. Figure 2-2 outlines the processing methods employed:



1) Images Are Extracted From Video (Adobe Premiere)
2) The OpenCV Viola-Jones face detector is used to extract a 222x222 region about the face in each image
3) For each subject, a neutral reference image is selected and a mask created
4) This mask is then used to compute the optical flow and generate a similarity transform which will align the faces in all the images with the face in the masked image
5) All subjects are now aligned to one another with the application of a four-point affine transformation
6) The resulting aligned images are cropped about a 99x123 window and grayscaled. These are the images that will be used in the recognition methods
7) The regions of this image within a predefined mask are the portions used in the recognition methods
+
=

Figure 2-2: Generalized Image Processing Method.

In (1) each of the video sequences was processed to generate a sequence of images. These are the images tabulated in “Single Images Before Processing” of Tables 2.4 and 2.5. In (2) the Viola-Jones face detector [2,19,20] was used to extract a 222x222 region around the face in each image[2]. As shown in (3), a neutral reference image

[2] To account for rotations in the subject’s head, each of the images was rotated between -30 and +30 degrees. The software then returned a 222 × 222 pixel cropping of the image at the location and rotation for which the OpenCV recognition metric returned the greatest value.

was then chosen for each of the twelve subjects and a binary mask was drawn on this image. The region inside this mask was used in the first of two alignment steps. In this first step, the mapping returned by the optical flow between the masked region of the reference image and one of the subject's corpus images was used to generate a similarity transform. This transform was then applied to the corpus image to align it with the subject's reference image. The process was repeated for every image in the corpus. Thus, by stage (4) each of the subjects has had the entirety of the images in his or her corpus aligned to a reference image particular to that subject. It was still necessary to align each of the subject's corpora to each other so that **all** images in the database were aligned to one another. In (5), this was accomplished by manually selecting four points[3] in the reference image. By selecting the equivalent four points in the reference images of the other subjects a second similarity transformation which would minimize the error between those points was computed. Each of the twelve transformations for each of the twelve subjects was then applied to the images of their respective corpora so that all images of the database were now aligned with one another. All the images in this globally aligned database were then gray-scaled and cropped to a 99 × 123 region about the face (6). Before each classification method was performed, the images were further masked as shown in (7) to extract only the expression-relevant portions of the face.

Tables 2.4 and 2.5 delineate the final state of the database. Each subject's corpus of natural images was used as it was returned by the final stage of processing in Figure 2-2. Played expressions, however, were manually inspected to ensure that only those images of the sequence which demonstrated the labeled emotion were present. This explains the significant difference in the numbers of "Single Images Before Processing" and "Processed Images" for the played expressions in Table 2.4.

Finally, it should be noted that while subject 103 is made available in the final processed images of the database and included in the tables below, her images were ultimately not used during the experimentation phase of this thesis. Several of her

[3] These points were the center of each eye, the midpoint between the nostrils, and the center meeting point between the lips.

image sequences contain expressions which were ambiguous and did not agree with the labels assigned to the sequence. Removing these sequences would have significantly diminished the size of her corpus relative to the other subjects, so she was instead excluded from the trials.

Original Number of Played Video Clips								
subject	**neutral**	**happiness**	**surprise**	**fear**	**sadness**	**disgust**	**anger**	**natural**
100	6	9	11	16	9	13	14	33
101	10	11	7	8	5	11	8	26
102	7	17	11	14	15	14	9	20
103	11	14	23	25	11	15	12	25
104	6	14	19	18	14	9	19	27
105	9	14	24	16	10	9	12	31
106	7	16	15	12	16	9	18	26
107	7	24	43	33	18	27	11	19
500	5	12	8	12	14	11	17	23
501	6	13	20	10	12	8	8	21
502	2	12	14	14	7	9	13	28
503	6	12	19	10	11	13	16	30
Single Images Before Processing								
subject	**neutral**	**happiness**	**surprise**	**fear**	**sadness**	**disgust**	**anger**	**natural**
100	1200	896	984	1414	1500	1136	1243	4369
101	1800	1550	558	661	1176	1075	963	5253
102	1260	2627	1018	1791	2532	1562	796	3469
103	774	1301	1234	1713	1464	1693	1553	3107
104	1080	1621	1024	1111	1632	637	1711	6063
105	617	1412	1485	1171	1985	688	1968	3158
106	1260	1842	1232	995	2615	724	1673	4857
107	1260	1787	2854	2380	1860	2265	1031	2762
500	900	1612	771	1211	2444	1427	1985	4503
501	538	1160	1670	1066	1469	640	561	2332
502	200	1182	1159	1051	1094	740	1060	3168
503	1080	1362	1593	1633	1720	1502	1880	5701
Processed Images								
subject	**neutral**	**happiness**	**surprise**	**fear**	**sadness**	**disgust**	**anger**	**natural**
100	1200	609	324	929	1363	861	958	4369
101	1800	1242	321	418	521	687	702	5253
102	1260	1491	626	1180	1989	949	503	3467
103	774	1301	1234	1713	1464	1693	1553	3107
104	1080	963	507	927	1590	582	1204	6054
105	617	1027	749	748	1822	483	1145	3158
106	1260	1231	391	330	1904	328	627	4857
107	1260	1046	1017	1006	717	931	436	2762
500	900	1065	418	620	1253	874	1478	4306
501	538	979	498	336	1045	400	442	2332
502	200	775	658	556	647	456	612	3168
503	1080	817	853	1238	895	953	1176	5701

Table 2.4: Enumeration of Available Played Data.

Natural Clips

subject	neutral	smile	laugh	surprise	fear	shock	disgust	dislike	puzzlement
100	6	5	5	2	4	1	3	4	6
101	10	5	6	1	1	0	3	1	5
102	7	2	11	0	0	0	2	3	1
103	11	3	5	6	0	0	3	5	1
104	6	1	11	5	0	0	6	0	1
105	9	1	11	1	0	2	6	5	2
106	7	2	6	2	1	1	2	4	5
107	7	3	4	0	0	4	2	3	1
500	5	2	8	0	0	0	2	1	3
501	6	2	10	2	0	0	2	2	0
502	2	3	6	0	0	0	3	6	1
503	6	2	6	3	0	0	3	4	5

Single Images Before Processing

subject	neutral	smile	laugh	surprise	fear	shock	disgust	dislike	puzzlement
100	1200	553	513	81	431	34	281	245	636
101	1800	670	853	69	245	0	331	129	549
102	1260	185	1410	0	0	0	105	386	81
103	774	200	856	340	0	0	153	520	70
104	1080	72	2232	826	0	0	1222	0	86
105	617	88	980	87	0	135	440	333	123
106	1260	202	870	203	171	125	281	535	540
107	1260	308	360	0	0	270	126	249	49
500	900	288	1565	0	0	0	318	51	374
501	538	182	1026	115	0	0	95	126	0
502	200	287	790	0	0	0	296	574	100
503	1080	308	1258	225	0	0	582	546	523

Processed Images

subject	neutral	smile	laugh	surprise	fear	shock	disgust	dislike	puzzlement
100	1200	553	513	81	431	34	281	245	636
101	1800	670	853	69	245	0	331	129	549
102	1260	185	1408	0	0	0	105	386	81
103	774	200	856	340	0	0	153	520	70
104	1080	72	2223	826	0	0	1222	0	86
105	617	88	980	87	0	135	440	333	123
106	1260	202	870	203	171	125	281	535	540
107	1260	308	360	0	0	270	126	249	49
500	900	288	1447	0	0	0	317	51	374
501	538	182	1026	115	0	0	95	126	0
502	200	287	790	0	0	0	296	574	100
503	1080	308	1258	225	0	0	582	546	523

Table 2.5: Enumeration of Available Natural Data.

Chapter 3

Methods for Expression Recognition

A multitude of potential methods for expression recognition exist [4, 8, 9], some derived upon previously successful work in face identification and some independently produced for the expression domain. Methods utilizing texture-value features and/or involving optical flow were the focus of research for this thesis. The following are the five methods investigated in this thesis:

1. Optical flow fields classified using Null Space Principal Component Analysis (NSPCA)

2. Optical flow fields classified using Support Vector Machines (SVM)

3. Texture-value features classified using SVMs

4. Combined internal classification outputs of the SVMs used in (2 and (3 above classified again using SVMs - a hierarchy of classifiers

5. Combined feature sets of (2 and (3 above classified using SVMs

3.1 Features

Three different feature types were employed in this thesis. The means for their generation and the highlighting attributes of each feature type are provided below.

3.1.1 Optical Flow

Optical flow [6] generates a matrix of vectors originating from pixel positions in one image and terminating at pixel positions within another image. These vectors represent the mapping between the pixel intensities in one image and the pixel intensities in another. These displacement vectors may be used to describe correspondences between similar images.

A naive approach to generate expression features using optical flow would involve directly taking the flow between each subject's reference image (a neutral expression) and each image of their corpus and use the flow fields in the training and testing of each classifier. These are the flows shown in the top right of Figure 3-1 labeled "Optical Flow 2". Each arrow in the top portion of this figure represents a separate flow field.

It would not be optimal to train on this flow, however, as it contains significant excess information, unrelated to the individual's expression. In particular, this flow still contains information regarding related to the identity of the subject. Directly comparing the flows from two different subjects will not yield viable information regarding their expressions until this information has been removed.

For this purpose a synthetic image[1] created from images outside the database was used. The optical flow can be taken between this synthetic image and the reference image of each subject. Because both the synthetic image and the subject exhibit neutral expressions in these images the optical flow taken between this synthetic image and the subject's reference image will indicate deformations related to the change in identity between the face in the synthetic image and the face in the reference image.

[1] The synthetic image was acquired by morphing images from a website which itself provided the averaged morph of a variety of individuals"http://www.uni-regensburg.de/Fakultaeten/phil_Fak_II/Psychologie/Psy_II/beautycheck/english/"

This flow is in the top left of Figure 3-1 and is referred to as “Optical Flow 1”. This flow was computed for each of the subject’s reference images.

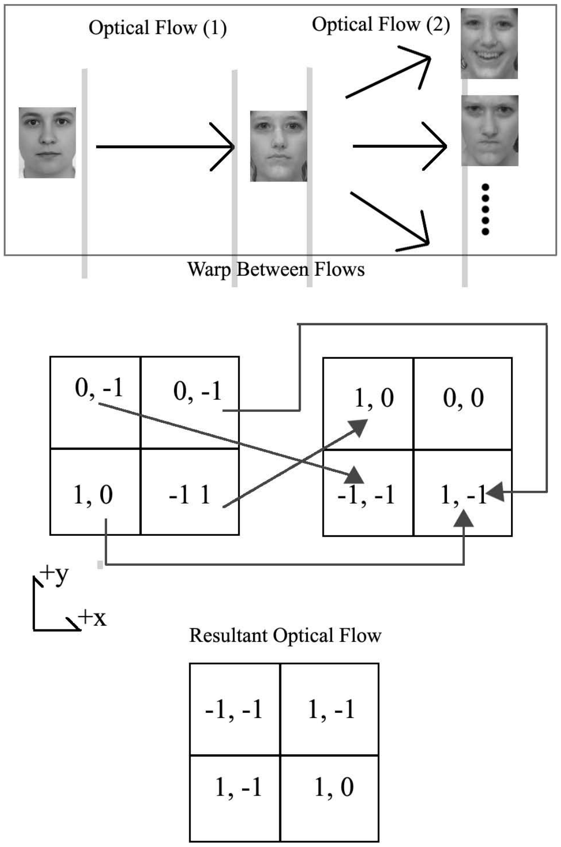


Figure 3-1: Visualization of Optical Flow Warp.

Using these two types of flow fields, it was then necessary to generate new flows for all the subjects which only illustrate changes resulting from the subject’s expression. This was accomplished by mapping each of the subject’s expressions to the synthetic image using these two flows, so that each subject’s expression information was normalized to the identity information of the synthetic image. In other words, the expression taking place on the subject’s face would now appear as if it were taking place on the synthetic face.

This mapping was achieved by using Optical Flow 1 to index into the values of Optical Flow 2. Figure 3-1 illustrates this operation. The operation is equivalent to an image warp ($warp : Flow \times Image \rightarrow Image$) using another optical flow field instead of an image ($warp : Flow \times Flow \rightarrow Flow$). The optical flows resulting from this warp operation were then masked as described in step (7) of Figure 2-2. These masked flows were the optical flows used by the classifier for training and testing.

3.1.2 Texture-value

The same mask applied to the optical flows was again used to extract the equivalent portion of each image from each subject's expression corpus. Histogram equalization was then performed on the gray values within the masked region to remove variations caused by lighting. The outputted image from this equalization was used by the classifier for training and testing.

3.1.3 Combined Feature Files

The optical flow features and the texture-value features were concatenated into one new feature set. Because the ranges of these two feature sets did not agree it was necessary that they be normalized. This was accomplished by computing the mean and variance of a large training set, first for the optical flow features and then for the texture features. The mean of the set was then subtracted from each data point of that method and the result then divided by the standard deviation of the large training set. The particular set used for the calculation of the mean and variance was the training set for the 5 vs. 6 test, which will be described later in this thesis.

3.2 Classifiers

Three classifiers were used to train and test the previously described feature data. They include a modified classification system using Principal Component Analysis (PCA) and the null space of the returned eigenspace, as well as a group of linear and gaussian trained Support Vector Machines(SVMs).

3.2.1 Null Space Principal Component Analysis (NSPCA) Algorithm

For each expression's training set of features, PCA was applied to the covariance matrix of the training set to return the N leading eigenvectors and eigenvalues. These eigenvalues drop off exceedingly quickly - the latter ones being many orders of magnitude less than their predecessors. As the following figure demonstrates, almost all the training sets used in this thesis reach an approximately level value for succeeding eigenvalues after $N = 30$;

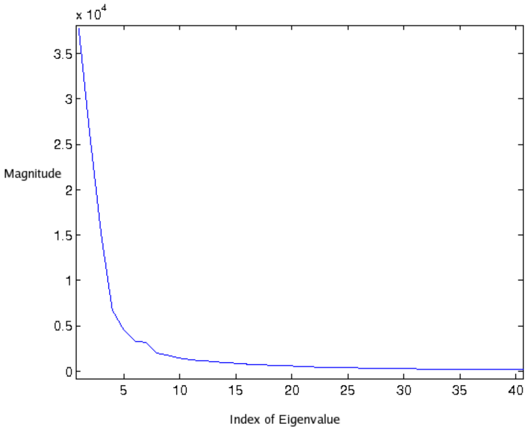


Figure 3-2: Example Covariance Matrix Eigenvalue Decay.

The null space was chosen to be the space orthogonal to the eigenspace spanned by the first $N = 30$ eigenvectors.[2]

Algorithm Walk-Through

The following is a step-by-step explanation of the NSPCA algorithm and can be overlooked by readers familiar with the process.

[2] The "null space" in the null space analysis is not a true null space in the sense that the values are null. Rather they are sufficiently small as compared to the largest eigenvalues that they are being approximated as being so.

Training

The training set is composed of a stack of feature vectors, generating a matrix $\mathbf{B}$ of dimensions $M \times N$ where M is the number of feature points and N the dimension of a feature point. As an example: the masked portion of the face on which the flow is computed is approximately 100×100 pixels. Since each pixel generates an x and y component in the flow, a vector of 20,000 points results. Since M is rarely larger than 8,000 it would be greatly preferred to base the generation of and later computations involving the covariance matrix upon this dimension, rather than the much larger N. Thus, a familiar optimization[3] involving the transpose of the data matrix was used to compute the eigenvalues and eigenvectors of the covariance matrix of each training set. This effectively reduces the order of complexity for the computation from $O(N^2)$ to $O(M^2)$.

Testing

Testing proceeded by iterating over each of the eigenspaces returned for each of the expressions' feature sets and then projecting the test feature into the eigenspace of that feature set. The set whose eigenspace returned the minimal norm was the set whose classification was given to the test feature. This norm is equivalent to the error of the projection illustrated in Figure 3-3. If $\mathbf{A}$ is the matrix of eigenvectors in columns as returned from the PCA, $\overline{\mathbf{x}}$ the mean of the training set, and $\mathbf{x}$ is the flow of the test image, the length l_{ns} of the null space component ns is given by:

$$l_{ns} = ||ns|| = ||(\mathbf{I} - \mathbf{A}\mathbf{A}^T)(\mathbf{x} - \overline{\mathbf{x}})|| \tag{3.1}$$

[3] The details of this optimization are available in the lecture notes of Dr. Bernd Heisele available at: http://stellar.mit.edu/S/course/9/fa03/9.913/index.html

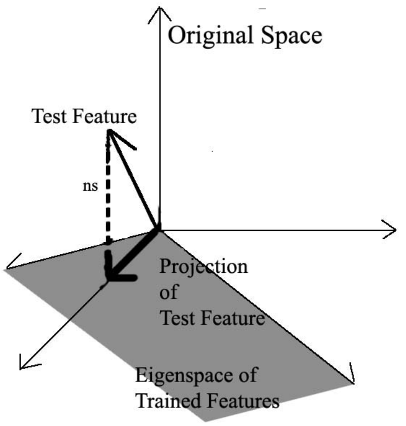


Figure 3-3: Visualization of the Projection and Resultant Distance Metric.

The length, l_{ns}, is inversely proportional to the test feature's similarity to the members of the training set who generated the eigenspace. The smaller the value of l_{ns} the more likely that this test feature should be classified as a member of the training set from which the eigenspace originated.

3.2.2 Support Vector Machines

In addition to the NSPCA classification method described above, SVMs were used for classification in a one-versus-all strategy. All SVMs were run using a linear classifier with the exception of the trials run upon the texture-value method, whose SVM utilized a set of Gaussian classifiers with $\sigma = 20.0$. The C-value was set to one for all experiments.

Hierarchical SVM Classifier

Each SVM produces classifier outputs after a testing session - the maximum value indicating the training class whose expression label will be assigned to the test point. For example, given six possible training categories, the SVM will output six separate values when computing the proper labeling of an inputted test feature. By concatenating these values, as they were output from the training of two different methods, a new dataset of feature points could be created and then fed to another SVM for training and testing. Figure 3-4 illustrates this resulting hierarchical structure.

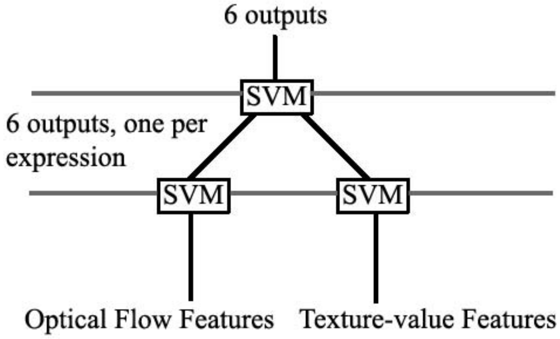


Figure 3-4: Hierarchical SVM Classifier.

This process was performed for the optical flow and texture-value data. Training flows were run through the SVM against **themselves**, to generate the necessary classifier data for training.

Chapter 4

Experimental Results

4.1 Overview

A completed expression recognition system, deployed in field, would ideally be able to

- Recognize expressions from a single individual over the course of several sessions.
- Recognize natural versions of an expression having trained on played versions of that expression.
- Recognize expressions of other individuals, having trained on a different set of individuals.

Thus, a canon of tests were run to verify the applicability of each technique to each of these desired abilities. Individual tests were run to ascertain the first objective, trials trained upon played expressions and then tested upon natural expressions were used to ascertain the second, and finally group tests between groups of subjects were used to ascertain the third.

As remarked in the preprocessing section of this thesis, the image sequences for played expressions were inspected by hand and only those images demonstrating the labeled emotion were retained. Of the resulting images the first third were used for testing and the later two-thirds for training. The final tabulation of training and test played data for each subject is provided in Table 4.1. Recall that subject 103 was removed from experimentation as many of her expressions were ambiguous and difficult to classify consistently.

subject	**Tested**	**Trained**
100	1680	3364
101	1296	2595
102	2244	4494
104	1924	3849
105	1989	3985
106	1602	3209
107	1716	3437
500	1900	3808
501	1232	2468
502	1233	2471
503	1975	3957

Table 4.1: Final Numbers of Played Expression Images used for Testing.

As natural expressions were meant to demonstrate the subtlety of real-world classification, their sequences were not manually edited. Additionally, as they were only used for testing, and not for training in the following experiments, the entirety of available images for each natural expression were used in the natural test sets.

In the following tables, the label "OF/NSPCA" refers to a trial in which the optical flow features were classified using NSPCA. Similarly, "OF/SVM" refers to optical flow features applied to a SVM classifier, "Texture-Value/SVM" to texture-values applied to an SVM classifier, and "Hierarchical-Classifiers/SVM", to the the hierarhcical classifier, and "Combined-Features/SVM" referring to the combined feature types as classified by an SVM.

4.2 Individual Tests

The played training data for each subject was trained upon for each of the five methods and then tested upon each played expression's test data.

Subject	OF/NSPCA	OF/SVM	Texture-Value/SVM	Hierarchical-Classifiers/SVM	Combined-Features/SVM
100	90.2%	89%	91.7%	89.2%	89.5%
101	90.4%	91.7%	89.8%	91.2%	91.3%
102	93.1%	93.3%	94.7%	93.6%	94%
104	92.3%	94.7%	91.3%	94.8%	92.2%
105	79.8%	75.9%	87.8%	76.2%	83.8%
106	93.3%	93.6%	96.4%	93.1%	96%
107	91%	95.5%	93.4%	95.9%	94.3%
500	88.1%	87.4%	86.3%	87.5%	84.7%
501	86.9%	87.7%	95.7%	87.3%	93.9%
502	95.1%	95.3%	98.5%	95.5%	98%
503	97.7%	92.6%	97.1%	92.9%	97.1%
Average	**90.7%**	**90.6%**	**93%**	**90.7%**	**92.3%**
Std. Dev.	**4.5%**	**5.4%**	**3.8%**	**5.4%**	**4.4%**

Table 4.2: Results: Single Experiments.

Generally speaking, all the methods did quite well on this test, which is to be expected considering the similarity between the training and test data. This appears to be true regardless of the classifier or feature set used.

4.3 Played vs. Natural Tests

Consulting Tables 2.4 and 2.5 one will note that only four expressions were common to both played expressions and natural expressions. These four expressions were trained upon using played data, and, as was available per subject, tested upon using the natural expressions. Recall that **all** available images for each natural expression were used for testing, in light of the paucity of images.

Subject	OF/NSPCA	OF/SVM	Texture-Value/SVM	Hierarchical-Classfiers/SVM	Combined-Features/SVM
100	68.5%	25.4%	53%	27%	61.4%
101	59.3%	55.2%	41.3%	56.7%	60.6%
102	37.6%	60%	61.7%	57.9%	56.9%
104	70.1%	67.5%	70%	67.4%	76.4%
105	43.4%	37.9%	33.8%	38.2%	34.6%
106	42.7%	52.7%	56.5%	51.6%	60.4%
107	51.6%	53.5%	72.4%	47.2%	71%
500	6.3%	30.3%	18.2%	27.6%	18.6%
501	33.9%	48.7%	41.1%	50.5%	47.2%
502	80.1%	46%	0%	43.6%	20.1%
503	54.2%	55.5%	43.2%	63.2%	48%
Average	**49.8%**	**48.4%**	**44.7%**	**48.3%**	**50.5%**
Std. Dev.	**19.5%**	**12.1%**	**20.8%**	**12.7%**	**18.3%**

Table 4.3: Results: Played vs. Natural Expressions.

All the methods generally performed worse for this experiment, which is again to be expected as the training data varies significantly from the test data. Future expression databases must be certain to include natural expressions in their training corpus for, as these results demonstrate, the difference between played subject data and natural data is significant. The subtlety of natural expressions challenges the recognition method to achieve a high resolution in its classification technique.

The extremely low recognition rates for subjects 500 and 502 can be explained by the very few natural expressions they possess. The results for these subjects should not be considered statistically significant.

4.4 Group Tests

Two tests were performed to ascertain viability of each method for group recognition. In the first, related to hereafter as 5 vs. 6, the played expressions of five subjects (100, 101, 102, 502, 503) were trained upon, and then tested against the played expressions of six subjects (104, 105, 106, 107, 500, 501). In a second test, hereafter referred to as 5 vs. same 5, the five subjects trained upon previously now had their test data used to test each method.

Subject	OF/NSPCA	OF/SVM	Texture-Value/SVM	Combined-Classifiers/SVM	Combined-Features/SVM
5VS6	35.8%	50.5%	51.5%	47.3%	58.2%
5VSsame5	90.9%	89%	92.6%	88.6%	94.2%

Table 4.4: Results: Group Experiments.

The warping method applied in Figure 3-1 was explicitly chosen for its ability to generalize expression data beyond the identity of a particular subject (as in 5 vs. 6). Surprisingly, however, the four methods involving the optical flow did not perform better than the texture-value method. In fact, OF/NSPCA did much worse than the other methods.

This may not be the case in experiments where the training and test data is less forgiving to the texture-value features. Recall that the training and test data for 5 vs. 6 contains individuals of both Caucasian and African ethnicity, as well both male and female sexes. Had this not been the case, and certain sexes or ethnicities had been present in the training, but not the test set, then the texture-value method would be forced to classify on a range of textures differing from those on which it was trained. Under these conditions the optical flow method should be less affected as it does not rely upon the luminosity of the subject's images, the subject's identity, nor upon the subject's skin color. Possible reasons for why the optical flow features did not dramatically surpass the texture-value features are provided in the following chapter.

Additionally, these results would imply that it is important to choose a proper classifier when working with test data significantly differing from the training data. The poor performance of OF/NSPCA in 5 vs. 6 was improved 10% by instead classifying with an SVM in OF/SVM. Yet, both methods did approximately the same in 5 vs. same 5 where the training and test data were quite similar.

A Receiver Operating Characteristic (ROC) curve permits a more detailed comparison of each method's relative utility. Figures 4-1 and 4-2 illustrate the percent false positive classification on the X-axis and the percent correct classification on the Y-Axis for each of the methods described above on both experiments 5 vs. 6 and 5 vs. same 5. Thus, the faster the graph rises and the larger the area beneath it, the better the classification method.

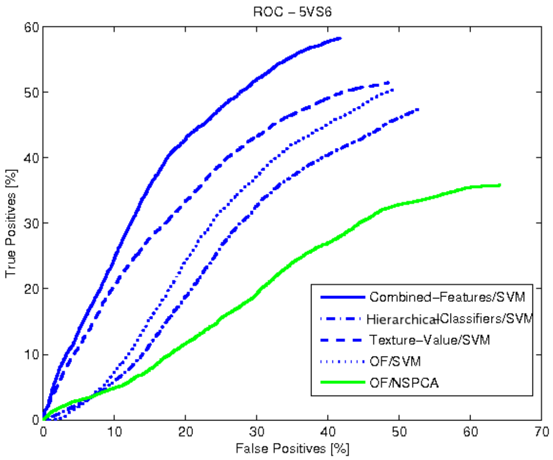


Figure 4-1: ROC curve for 5VS6.

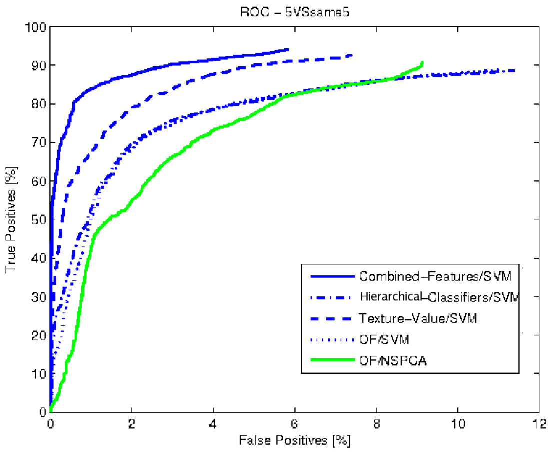


Figure 4-2: ROC curve for 5VSsame5.

Clearly, combining the feature files of the texture-value and optical flow methods generated a significantly improved classifier. It was not anticipated that the texture-value features would surpass the NSPCA as vividly as shown in these curves.

Chapter 5

Discussion

5.1 Postulates for Improved Methods

An explanation for the optical flow's inability to dramatically improve upon the texture-value's results is provided below. Additionally, techniques which will improve the quality of the classification and restore the optical flow's predicted versatility are put forth.

5.1.1 Localization of Relevant Image Regions

Currently, the optical flow takes place over the entirety of the face not masked out by the mask in (7) of Figure 2-2. While this removes significant portions of the image irrelevant to expression recognition, remaining portions of the image may still adversely affect training. As it is, even after masking, the flow still incorporates information about the nose, portions of the cheek region, and chin which were not directly involved with the creation of an expression. This inclusion provides an opportunity for errors in the flow to become prominent during the classification process and should be excluded. These errors are discussed in detail in the following section.

5.1.2 Deformation Models for Improved Optical Flow

During the course of this research, it was made clear that obtaining a satisfactory flow is not only imperative to successful recognition, but extremely difficult to obtain. The failure of the optical flow features to increase performance dramatically over the texture-valued features may be explained by this difficulty. Figure 5-1 exemplifies the nature of a faulty flow.

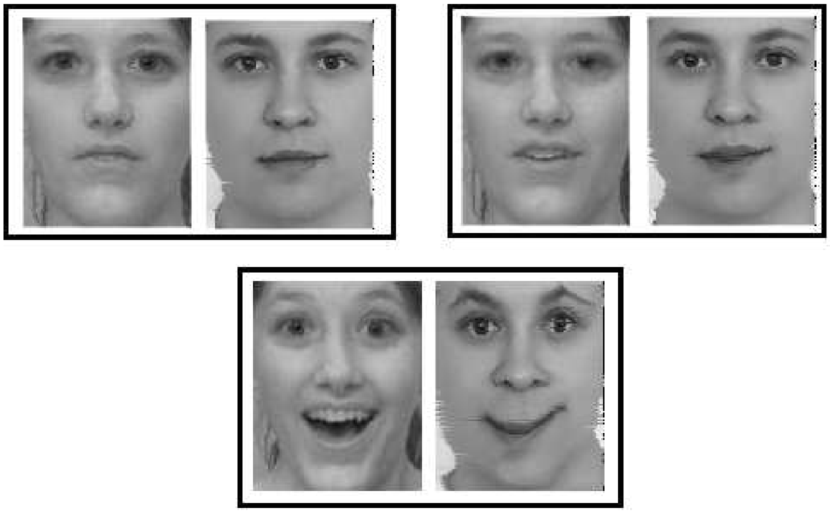


Figure 5-1: Side-by-Side Demonstration of Optical Flow Failure.

Each pair of images in the figure consists of the original image of subject 100, and the resultant image generated when the optical flow is used to warp the synthetic image. As mentioned previously, the goal is to map subject 100's expression onto the synthetic face. In the first pair, in which subject 100 possesses a mostly neutral expression, there is a strong correspondence between the warped image and the original image - signifying a good optical flow. In the next pair, subject 100 begins to open her mouth, and again the flow computation continues to find a good correspondence between her upper and lower lip and the lips of the synthetic image.

By the third pair of images, however, the gap between the lips has proven beyond the capabilities of the flow, and while it is still possible to ascertain the expression present in the synthetic warp, a significant amount of important information has been lost. These anomalies are not particular to the mouth, but include other such features as the eyebrows. Note how the correspondence between the eyebrows has been lost in the third pair and a sharp, arching ridge is present that should not exist.

Future research should direct efforts to constructing a model for optical flows which will prevent this behavior. Specifically, a synthetic version of the facial motion in a 2-dimensional analog should be used to constrain the OF computation (e.g. motion of eyebrows, lips, etc.). This deformable model could be used as a foundation from which new flows are computed before performing the warp operation described in Figure 3-1.

Chapter 6

Conclusion

This thesis has enumerated the properties required in an adequate facial-expression database. The tools used to create and process such a database have been described and the database images resulting from the application of these tools have been tabulated. Three varieties of features and three different classifiers were then used with the images of this database to determine the relative efficacy of five different classification methods. From these trials it was concluded that:

- The best possible classification method was the SVM applied to the combined optical flow and texture-value feature files. Achieving a 94.2% successful classification rate for groups of subjects, and a 92% average for individuals, this classifier was able to classify both individuals and groups well.

- Generalizing data from one set of individuals to another was difficult for all the classifiers. Even the best classifier dropped from 94.2% to 58.2% when the test set was switched from a group of individuals in the training set to a different group of individuals outside that set.

- Natural expressions are much more subtle than played expressions and will challenge a classifier's ability to classify minute changes. In spite of this, the classifiers used in this thesis were still able to classify approximately half of their natural test data correctly.

- For training and test sets that are similar, the classifiers used in this thesis achieve approximately the same recognition rates. Conversely, when the two sets do differ, performance was improved with the use of an SVM rather than the NSPCA classifier.

- The optical flow, as it was computed for this thesis, is occasionally faulty and must be complimented by a model to guide the computation of the flow. The computation and application of such a model should be the focus of future research.

- Finally, the texture-value features performed quite well, despite their simplicity. In all the experiments, they were able to achieve rates commensurate with that of the optical flow. This may not be the case in a database possessing a wider gamut of textures.

6.1 Acknowledgments

I am very grateful to both Dr.Bernd Heisele and Professor Tomaso Poggio for providing me with their support and the opportunity to participate in their research at CBCL. Without their support this thesis would not have been impossible. Finding such a friendly and professional research environment is rare, and I am grateful for the opportunity to work at CBCL.

I would also like to thank my parents and family, who have likewise provided immeasurable support.

Finally, I would like to thank Ms. Hunter of the Course VI department for making the Masters of Engineering Program possible. The Masters of Engineering Program has been both an enlightening and empowering experience and I am very grateful that I was able to be a participant.

This report describes research done at the Center for Biological and Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain and Cognitive Sciences, and which is affiliated with the Computer Sciences and Artificial Intelligence Laboratory (CSAIL). This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. MDA972-04-1-0037, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA- 0218693, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506, and National Institutes of Health (Conte) Contract No. 1 P20 MH66239-01A1. Additional support was provided by: Central Research Institute of Electric Power Industry (CRIEPI), Daimler-Chrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda Research and Development Co., Ltd., Industrial Technology Research Institute (ITRI), Komatsu Ltd., Eugene McDermott Foundation, Merrill-Lynch, NEC Fund, Oxygen, Siemens Corporate Research, Inc., Sony, Sumitomo Metal Industries, and Toyota Motor Corporation.

Bibliography

[1] M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan. Real time face detection and expression recognition: Development and application to human-computer interaction. In *CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, 2003.

[2] The Intel Corporation. *OpenCV Online Reference Manual.* 2005.

[3] M. S. Bartlett J. C. Hager P. Ekman and T. J. Sejnowski. Measuring facial expressions by computer image analysis. In *Psychophysiology*, pages 36:253–263, 1999.

[4] D. Fidaleo and M. Trivedi. Manifold analysis of facial gestures for face recognition. In *Proc. of the 2003 ACM SIGMM workshop on Biometrics methods and appl.*, pages 65–69, 2003.

[5] R. Fischer. *Automatic Facial Expression Analysis and Emotional Classification.* Diploma thesis.

[6] BKP Horn and BG Schunck. Determining optical flow. In *AI Memo 572. Massachusetts Institue of Technology.*, 1980.

[7] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *The 4th IEEE Int. C. on Automatic Face and Gesture Recognition (FG'00)*, 2000.

[8] D. Kim and Z. Bien. Fuzzy neural networks (fnn)-based approach for personalized facial expression recognition with novel feature selection model. In *Proc. of IEEE Int. Conf. on Fuzzy Systems*, pages 908–913, 2003.

[9] A. Lanitis, C. Taylor, and T. Cootes. Automatic interpretation and coding of face images using flexible models. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, pages 743–746, 1997.

[10] J. J. Lien. Automatic recognition of facial expressions using hidden markov models and estimation of expression intensity. In *Carnegie Mellon University, Technical Report CMU-RI-TR-98-31, Ph.D Dissertation*, 1998.

[11] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. *Proceedings, Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara Japan*, pages 200–205, 1998.

[12] J. Movellan and M. Bartlett. *The Next Generation of Automatic Facial Expression Measurement.* Oxford University Press, second edition. in press.

[13] T. Otsuka and J. Ohya. Recognizing abruptly changing facial expressions from time-sequential face images. In *Proc. of Computer Vision and Pattern Recognition Conf. (CVPR98)*, 1998.

[14] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi. The feret evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10), October 2000.

[15] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003.

[16] A. Lanitis C. J. Taylor and T. F. Cootes. Active shape models-their training and application. In *Computer Vision Graphics and Image Understanding*, pages 61:38–59, 1995.

[17] A. J. Toole, J. Harms, and S. L. Snow. A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[18] M. A. O. Vasilescu and D. Terzopoulos. Multilinear image analysis for facial recognition. In *International Conference on Pattern Recognition (ICPR '02)*, 2002.

[19] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 511–518, 2001.

[20] P. Viola and M. Jones. Robust real-time face detection. In *Proc. 8th International Conference on Computer Vision*, volume 20(11), pages 1254–1259, 2001.

[21] L. Williams. Perfromance-driven facial animation. In *ACM SIGGRAPH Conference Proceedings*, pages 24(4):235–242. IEEE, 1990.