M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 488
Accepted to: Workshop on Emotion-Based Agent Architectures, at the Third International Conference on Autonomous Agents, May 1, 1999, Seattle, WA

Offline and Online Recognition of Emotion Expression from Physiological Data

Elias Vyzas and Rosalind W. Picard
MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139-4307
{evyzas, picard}@media.mit.edu

Abstract

We develop a method for offline and online recognition of the emotional state of a person deliberately expressing one of eight emotions. In terms of offline recognition, this paper presents recent improvements to a method previously developed in the MIT Media Lab, which involved recognition using physiological data collected from an actress over many weeks. The improvements involve (1) more robust handling of day-to-day variations in the data, (2) use of longer episodes of data, (3) use of heart-rate information, extracted from a blood volume pressure sensor, and (4) the use of alternative features. The success rates thus increased from 50.62% to 81.25% for all 8 emotions. Additionally, the method has been adapted to run online, so that it can be used for real-time applications. The performance of the real-time version of the algorithm currently lags 8% behind that of the corresponding offline version, but we continue to investigate improvements.

The success rates obtained with the physiological-based recognition are now comparable to those obtained in facial and vocal expression recognition, and offer complementary information or an alternative to such means. The recognition results demonstrated here indicate that there is significant information in physiological signals for classifying the affective state of a person who is deliberately expressing a small set of emotions.

1 Introduction

This paper addresses emotion recognition, specifically the recognition by computer of affective information expressed by one person over many weeks, including lots of day-to-day variations. Recognition is run on a set of features extracted from physiological signals, currently measured from the surface of the skin of a person expressing one of eight emotions. We show improvements over previous results in offline recognition [12]. We also describe a new adaptation of the method that runs online, bringing it closer to a number of real-life real-time applications.

This research is part of a larger effort aimed at giving computers the skills of "emotional intelligence," such as the ability to recognize a person's emotions and to respond appropriately to those emotions. Recognition of emotional information is a key part of human-human communication, and is therefore expected to be necessary in building natural and intelligent human-computer interaction. Software agents and other adaptive interfaces can benefit from recognizing which behaviors cause states such as joy or anger in their users. If a particular behavior pleases a user, it might be reinforced, whereas if a behavior makes a user angry, then the behavior probably needs modification. The idea is that the agent should be adapting to the user with minimal effort on the user's part. Users *naturally* express emotions to the computer, and can do so without having to interrupt the session to click on a special menu or other artificial feedback mechanism. As the computer recognizes the natural expression of the user, it receives information that helps it better adapt to serve that user.

2 Background

A summary of related literature, as well as the experiment, methodology and some of the previous results are mentioned here. They can all be found in greater length in [12].

The research described here focuses on recognition of eight emotional states during deliberate emotional expression by an actress. These states were: Neutral (no emotion) (N), Anger (A), Hate (H), Grief (G), Platonic Love (P), Romantic Love (L), Joy (J), and Reverence (R). The specific states one would want a computer to recognize will depend on the particular application. The eight emotions used in this research are intended to be representative of a broad range, which can be described in terms of the "arousal-valence" space commonly used by psychologists [7]. The arousal axis ranges from calm and peaceful to active and excited, while the valence axis ranges from negative to positive. For example, anger was considered high in arousal, while reverence was considered low. Love was considered positive, while hate was considered negative.

There has been prior work on emotional expression recognition from speech and from image and video; this work, like ours, has focused on deliberately expressed emotions. The problem is a hard one when you look at the few benchmarks which exist. In general, people can recognize affect in neutral-content speech with about 60% accuracy, choosing from among about six different affective states [10]. Computer algorithms can match this accuracy but only under more restrictive assumptions, such as when the sentence content is known. Facial expression recognition is easier, and the rates computers obtain are higher: from 80-98% accuracy when recognizing 5-7 classes of emotional expression on groups of 8-32 people [13, 3]. Facial expressions are easily controlled by people, and easily exaggerated, facilitating their discrimination.

Emotion recognition can also involve analyzing posture, gait, gesture, and a variety of physiological features in addition to the ones described in this paper. Additionally, emotion recognition can involve prediction based on cognitive reasoning about a situation, such as "That goal is important to her, and he just prevented her from obtaining it;

therefore, she might be angry at him." Such a framework for analysis of affective dynamics has been developed under Affect Control Theory [5, 11]. The best emotion recognition is likely to come from pattern recognition and reasoning applied to a combination of all of these modalities, including both low-level signal recognition, and higher-level reasoning about the situation [8].

For the first part of the research described here, four physiological signals of an actress were recorded during deliberate emotional expression. The signals measured were electromyogram (EMG) from the jaw, representing muscular tension or jaw clenching, blood volume pressure (BVP) and skin conductivity (GSR) from the fingers, and respiration from chest expansion. Data was gathered for each of the eight emotional states for approximately 3 minutes each. This process was repeated for several weeks. The four physiological waveforms were each sampled at 20 samples a second. The experiments use 2000 samples per signal, for each of the eight emotions, gathered over 20 days. Hence there are a total of 32 signals a day, and 80 signals per emotion.

Let one of the four raw signals, the digitized EMG, BVP, GSR, and Respiration waveforms, be designated by S. The signal is gathered for 8 different emotions each session, for 20 sessions. Let S_n represent the value of the n^{th} sample of the raw signal, where $n = 1...N$ and $N = 2000$ samples. Let $\tilde{S}_n$ refer to the normalized signal (zero mean, unit variance), formed as:

$$\tilde{S}_n = \frac{S_n - \mu}{\sigma} \quad i = 1, ..., 4$$

where μ and σ are the means and standard deviations explained below. We extract 6 types of features for each emotion, each session:

1. the means of the raw signals (4 values)

$$\mu = \frac{1}{N}\sum_{n=1}^{N} S_n \quad i = 1, ..., 4 \qquad (1)$$

2. the standard deviations of the raw signals (4 values)

$$\sigma = \left(\frac{1}{N-1}\sum_{n=1}^{N}(S_n - \mu)^2\right)^{1/2} \quad i = 1, ..., 4 \qquad (2)$$

3. the means of the absolute values of the first differences of the raw signals (4 values)

$$\delta_1 = \frac{1}{N-1}\sum_{n=1}^{N-1}|S_{n+1} - S_n| \quad i = 1, ..., 4 \qquad (3)$$

4. the means of the absolute values of the first differences of the normalized signals (4 values)

$$\tilde{\delta}_1 = \frac{1}{N-1}\sum_{n=1}^{N-1}\left|\tilde{S}_{n+1} - \tilde{S}_n\right| = \frac{\delta_1}{\sigma} \quad i = 1, ..., 4. \qquad (4)$$

5. the means of the absolute values of the second differences of the raw signals (4 values)

$$\delta_2 = \frac{1}{N-2}\sum_{n=1}^{N-2}|S_{n+2} - S_n| \quad i = 1, ..., 4 \qquad (5)$$

6. the means of the absolute values of the second differences of the normalized signals (4 values)

$$\tilde{\delta}_2 = \frac{1}{N-2}\sum_{n=1}^{N-2}\left|\tilde{S}_{n+2} - \tilde{S}_n\right| = \frac{\delta_2}{\sigma} \quad i = 1, ..., 4 \qquad (6)$$

Therefore, each emotion is characterized by 24 features, corresponding to a point in a 24-dimensional space. The classification can take place in this space, in an arbitrary subspace of it, or in a space otherwise constructed from these features. The total number of data in all cases is 20 points per class for each of the 8 classes, 160 data points in total. Note that not all the features are independent; in particular, two of the features are nonlinear combinations of the other features.

2.1 Methodology

There is no guarantee that the features chosen above are all appropriate for emotion recognition. Nor is it guaranteed that emotion recognition from physiological signals is possible. Furthermore, a very limited number of data points—20 per class—is available. Hence, we expect that the classification error may be high, and may further increase when too many features are used. Therefore, reductions in the dimensionality of the feature space need to be explored, among with other options. We focus on three methods for reducing the dimensionality, and evaluate the performance of these methods.

The **Sequential Floating Forward Search (SFFS)** method [9] is chosen due to its consistent success in previous evaluations of feature selection algorithms, where it has recently been shown to outperform methods such as Sequential Forward and Sequential Backward Search (SFS, SBS), Generalized SFS and SBS, and Max-Min, [6] in several benchmarks. Of course the performance of SFFS is data dependent and the data here is new and difficult; hence, the SFFS may not be the best method to use. Nonetheless, because of its well documented success in other pattern recognition problems, it will help establish a benchmark for the new field of emotion recognition and assess the quality of other methods.

The SFFS method takes as input the values of n features. It then does a non-exhaustive search on the feature space by iteratively adding and subtracting features. It outputs one subset of m features for each m, $2 \leq m \leq n$, together with its classification rate. The algorithm is described in detail in [9].

Fisher projection (FP) [2] is a well-known method of reducing the dimensionality of the problem in hand, which involves less computation than SFFS. The goal is to find a projection W of the data to a space of fewer dimensions than the original where the classes are well separated. Due to the nature of the Fisher projection method, the data can only be projected down to $c-1$ (or fewer if one wants) dimensions, assuming that originally there are more than $c-1$ dimensions and c is the number of classes. It is important to keep in mind that if the amount of training data is inadequate, or the quality of some of the features is questionable, then some of the dimensions of the Fisher projection may be a result of noise rather than a result of differences among the classes. In this case, Fisher might find a meaningless projection which reduces the error in

the training data but performs poorly in the testing data. For this reason, projections down to fewer than $c-1$ dimensions are also evaluated in the paper. Note that if the number of features n is smaller than the number of classes c, the Fisher projection is meaningful only up to at most $n-1$ dimensions. Therefore in general the number of Fisher projection dimensions d is $1 \leq d \leq \min(n, c) - 1$. For example, when 24 features are used on all 8 classes, all $d = [1, 7]$ are tried. When 4 features are used on 8 classes, all $d = [1, 3]$ are tried.

As mentioned above, the SFFS algorithm proposes one subset of m features for each m, $2 \leq m \leq n$. Therefore, instead of feeding the Fisher algorithm with all the 24 features, we can use the subsets that the SFFS algorithm proposes as our input to the Fisher Algorithm. Note that the SFFS method is used here as a simple preprocessor for reducing the number of features fed into the Fisher algorithm, and not as a classification method. We call this hybrid method **SFFS-FP**.

The Maximum a Posteriori (MAP) classification is used for all Fisher Projection methods, while the SFFS came with a built-in k-nearest-neighbor classifier. The leave-one-out method is chosen for cross validation because of the small amount of data available.

2.2 Results

Some relevant results from the classification algorithms are shown in Table 1. All methods performed significantly better than random guessing, indicating that there is emotional discriminatory information in the physiological signals.

The classification rates obtained by SFFS and SFFS-FP are reported in Table 1.

Number of Features	Without Day Matrix			With Day Matrix	
	SFFS (%)	Fisher (%)	SFFS-FP (%)	Fisher (%)	SFFS-FP (%)
24	40.62	40.00	46.25	49.38	50.62

Table 1: Classification rates for 8 emotions from all 20 days (160 data points in total) of Data Set A and different methods used. The Day Matrix adds 19 features to the data fed to the Fisher Algorithm.

3 Day dependence

As mentioned previously, the data were gathered in 20 different sessions, one session each day. During the classification procedure, we noticed high correlation between the values of the features of different emotions in the same session. In previous work [12] we first quantified this phenomenon by building a day (session) classifier and then used it to improve the emotion classification results by including the day information in the features. Here we first summarize the previous results and then present a more robust handling of the day information.

3.1 Day classifier

We use the same set of 24 features, the Fisher algorithm, and the leave-one-out method as before, only now there are $c = 20$ classes instead of 8. Therefore the Fisher projection is meaningful from 1 to 19 dimensions. The resulting "day classifier" using only the Fisher projection and the leave-one-out method with MAP classification, yields a classification accuracy of 133/160 (83%), an extremely high success rate.

3.2 The Day Matrix

According to the results of the previous section, the features extracted from the signals are highly dependent on the day the experiment was held. Therefore, we would like to augment the set of features to include both the **Original** set of 24 features and a second set incorporating information on the day the signals were extracted. A **Day Matrix** was constructed, which included a 20-number long vector for each emotion, each day. It was the same for all emotions recorded the same day, and differed among days. We chose the 20-number vector as follows: For all emotions of day i all entries are equal to 0 except the i'th entry which is equal to a given constant C. This gave a 20x20 diagonal matrix for each emotion. The problem was that when the feature space included the Day Matrix, the Fisher projection algorithm encountered manipulations of a matrix which was close to singular. We could still proceed with the calculations but they were less accurate. Here we present a more robust version where the vector is 19-number long and does not encounter singularity problems.

Let us think of a case where the data come from only 2 different days and only 1 feature is extracted from the data (This is the only way the following manipulations can be visualized, but it can trivially extend to more features). Although the feature values of one class are always related to the values of the other classes in the same way (for example the mean EMG for anger may always be higher than the mean EMG for Joy), the actual values may be highly day-dependent (Fig. 1a). To alleviate this problem an extra dimension can be added before the features are fed to the Fisher Algorithm (Fig. 1b). If the data came from 3 different days, 2 extra dimensions would have to be added rather than one (Fig. 1c), etc. Therefore, in the general case $D-1$ extra dimensions are needed for data coming from D different days, and 19 extra dimensions are needed in our case. The above can be also seen as using the minimum number of dimensions so that each of D points can be at equal distance from all others. Therefore the $D-1$ dimensional vector will contain the coordinates of one such point for each day. This vector is the same for all emotions recorded the same day. The classification improvement for 8 emotions can be seen in Table 1.

4 Improved use of data and new features

4.1 Data

In previous work [12], we used data consisting of 2000 samples-per-signal, for each of the eight emotions, gathered over 20 days.

The data were originally gathered in 34 sessions where the 8 different emotions were expressed one after the other. Each full session lasted around 25 minutes, resulting in around 28 to 33 thousand samples per signal, with each emotion being around 2 to 5 thousand samples long, due to the randomness of the Clynes method of eliciting the

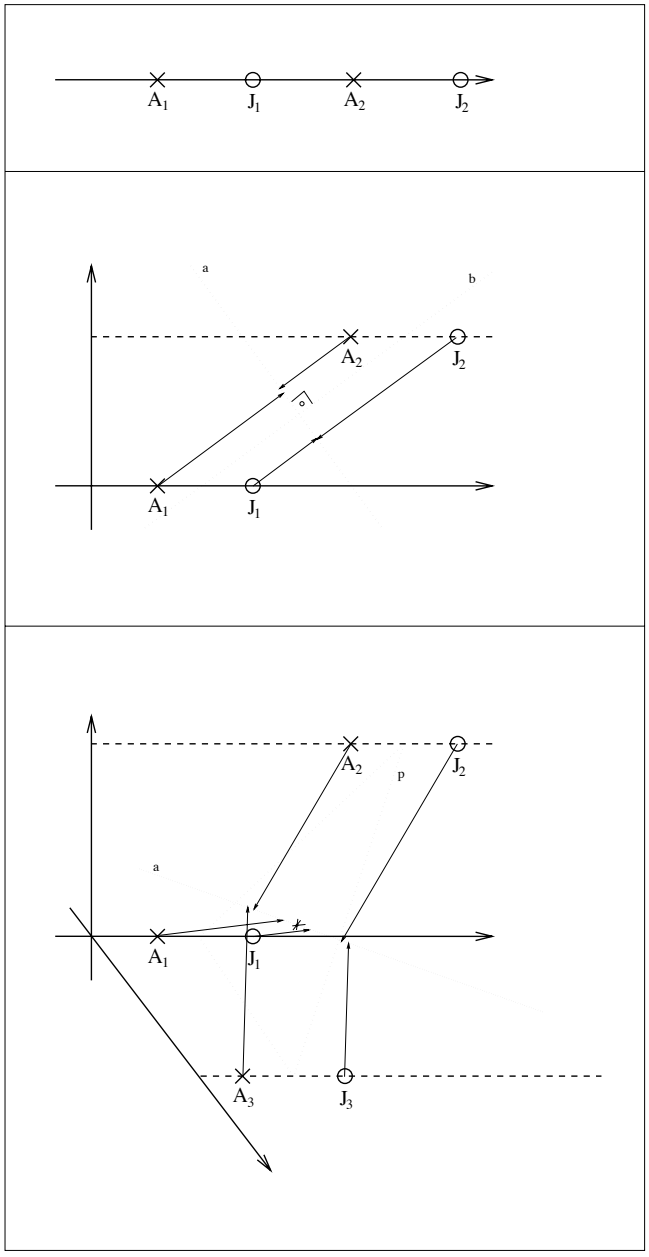

Figure 1: Fictitious example of a highly day-dependent feature for 2 emotions from 2 different days. (a) The feature values for (A)nger and (J)oy from 2 different days. (b) Addition of an extra dimension allows for a line b to separate Anger from Joy. The data can be projected down to line a, so the addition of the new dimension did not increase the final number of features. (c) In the case of data from 3 different days, addition of 2 extra dimensions allows for a plane p to separate Anger from Joy. The data can again be projected down to line a, not increasing the final number of features.

Data	Without Day Matrix (%)	With Day Matrix (%)
Set A	42.97	46.09
Set B	54.69	54.69

Table 2: Comparative classification rates for the 16 common days (128 data points in total) between Data Sets A and B, using 24 features fed to the Fisher Algorithm. The results suggest that using the longer data (Set B) improves classification performance.

emotional states [1]. In several occasions one or more sensors failed during parts of the experiment. The first 20 sessions were the ones used in the previous sections, choosing 2000 samples from each emotional state while trying to avoid parts where the sensors had failed. The question which remained was if any information could be extracted from the uninterrupted data, like transition characteristics, or if an online classifier could be built. Therefore, we revisited the data from the full sessions and chose 20 days in which the sensors did not fail during any part of the experiment. 16 of the original days and another 4 which had not been used before were included. We call this new set of data "Set B", with "Set A" being the original data mentioned in the previous sections. Some comparative results between the common days of the two slightly different sets of data can be seen in Table 2.

4.2 Features

Using peak detection on the Blood Volume Pressure signal, the Heart Rate can be calculated. The same 6 features proposed in Section 2 can be extracted from the Heart Rate as well. Additionally, a set of 11 other features have been proposed [4] for use with these physiological data. We would like to see if the inclusion of any of the above features can improve classification. The results can be seen in Tables 3, 4 and 5. Note that the total number of different features is 40 (rather than 41) because the mean EMG that was proposed in [4] was already included in the original 24 features.

We can see that in most cases, a small number m_{SFFS} of the original features gave the best results in SFFS. For SFFS-FP a slightly larger number $m_{SFFS-FP}$ of features tended to give the best results. These extra features found useful in SFFS-FP but not in pure SFFS, could be interpreted as containing some useful information, but together with a lot of noise. That is because feature selection methods like SFFS can only accept/reject features, while the Fisher algorithm can also scale them appropriately, performing a kind of "soft" feature selection and thus making use of such noisy features.

5 Online Recognition

Each day of Data Set B contains a continuous stream of data running through 8 different emotions. This data set is then appropriate for training and testing an online algorithm.

Number of Features	Without Day Matrix			With Day Matrix	
	SFFS (%)	Fisher (%)	SFFS-FP (%)	Fisher (%)	SFFS-FP (%)
24	49.38	51.25	56.87	54.37	63.75
30 (incl. HR)	52.50	56.87	60.00	58.75	63.75
11 (other)	60.62	70.00	70.63	61.25	63.12
40 (incl. HR, other)	65.00	77.50	81.25	77.50	78.75

Table 3: Comparative classification rates for 8 emotions from all 20 days (160 data points in total) of Data Set B and different features and methods used. The Day Matrix adds 19 features to the data fed to the Fisher Algorithm.

Number of Features	Without Day Matrix		With Day Matrix
	m_{SFFS}	$m_{SFFS-FP}$	$m_{SFFS-FP}$
24	14	16	19
30 (incl. HR)	5	7	22
11 (other)	11	7	7
40 (incl. HR, other)	8	25	32

Table 4: Number of features m proposed by the SFFS algorithms that gave the best results in Data Set B. When a range of SFFS algorithms performed equally well, only the one proposing the fewest features is listed.

Number of Features	Without Day Matrix		With Day Matrix	
	Fisher	SFFS-FP	Fisher	SFFS-FP
24	7	7	4	4
30 (incl. HR)	4	5	3	4
11 (other)	5	6	5	3
40 (incl. HR, other)	7	5	7	6

Table 5: Number of dimensions used in the Fisher Projections which gave the best results, out of a maximum of 7 dimensions. When a range of Fisher Projections performed equally well, only the one using the fewest dimensions is listed.

5.1 The iterative algorithm

Most of the data manipulation in this thesis has been done using MATLAB which is relatively slow compared to C/C++ and other compiled programming languages but has very good vector/matrix manipulation abilities. Any real-life real-time application will probably not be using MATLAB, so manipulating large vectors at every time step will probably make the whole process too slow. Therefore, in the online version of the algorithm we will only use features whose values can be updated at every time step with minimal computational cost. The 6 features per signal proposed previously can be iteratively updated using the following algorithm (where S_{N+1} is the signal of the time step just incorporated in the data and W is the width of the moving window in number of time steps):

For $3 \leq N < W$

$$\mu_{N+1} = \frac{N}{N+1}\mu_N + \frac{1}{N+1}S_{N+1} \tag{7}$$

$$\sigma_{N+1} = \begin{array}{l}\left(\frac{N-1}{N}\sigma_N^2 + \frac{1}{N}S_{N+1}^2\right. \\ \left. -\frac{2}{N}S_{N+1}\mu_{N+1} + \frac{1}{N}\mu_{N+1}^2\right)^{1/2}\end{array} \tag{8}$$

$$\delta_{1_{N+1}} = \frac{N-1}{N}\delta_{1_N} + \frac{1}{N}|S_{N+1} - S_N| \tag{9}$$

$$\tilde{\delta_{1}}_{N+1} = \frac{\delta_{1_{N+1}}}{\sigma_{N+1}} \tag{10}$$

$$\delta_{2_{N+1}} = \frac{N-2}{N-1}\delta_{2_N} + \frac{1}{N-1}|S_{N+1} - S_{N-1}| \tag{11}$$

$$\tilde{\delta_{2}}_{N+1} = \frac{\delta_{2_{N+1}}}{\sigma_{N+1}} \tag{12}$$

And for $N \geq W$

$$\mu_{N+1} = \mu_N + \frac{1}{W}\left(S_{N+1} - S_{N+1-W}\right) \tag{13}$$

$$\sigma_{N+1} = \begin{array}{l}\left(\sigma_N^2 + \frac{1}{W-1}(S_{N+1}^2 - S_{N+1-W}^2)\right. \\ \left. -\frac{W}{W-1}\left(\mu_{N+1}^2 - \mu_N^2\right)\right)^{1/2}\end{array} \tag{14}$$

$$\delta_{1_{N+1}} = \begin{array}{l}\delta_{1_N} + \frac{1}{W-1}\left(\left|S_{N+1} - S_N\right|\right. \\ \left. -\left|S_{N+2-W} - S_{N+1-W}\right|\right)\end{array} \tag{15}$$

$$\tilde{\delta_{1}}_{N+1} = \frac{\delta_{1_{N+1}}}{\sigma_{N+1}} \tag{16}$$

$$\delta_{2_{N+1}} = \begin{array}{l}\delta_{2_N} + \frac{1}{W-2}\left(\left|S_{N+1} - S_{N-1}\right|\right. \\ \left. -\left|S_{N+3-W} - S_{N+1-W}\right|\right)\end{array} \tag{17}$$

$$\tilde{\delta_{2}}_{N+1} = \frac{\delta_{2_{N+1}}}{\sigma_{N+1}} \tag{18}$$

The estimates for the first few steps can be calculated using the offline formulae (Eqns. 1-6)

The above iterations assume a continuous feed of data, therefore we will be using the long continuous data of Set B, as mentioned earlier. Using all 5 signals (EMG, BVP,

GSR, Respiration, and HR), gives a total of 30 features that can be calculated for every position of the moving window, for each one of the days.

5.2 Training data

Given that this is an online algorithm, it is not clear if we should use data from emotions of one day in the training of the classifier for other emotions of the same day. Therefore, *assuming that a person does not re-train the algorithm during the day*, we only use features from other days to train the classifier. Because of the small amount of days available, we use the leave-one-out method. This means that a new classifier is trained using 19 days and tested on the one left out, with the process repeated for all 20 days. Each day has around 30 thousand time steps, so a moving window can produce around that many sets of 30 features. But using all these sets for training would make the problem computationally very hard, requiring extreme amounts of disk space, memory and time, and would be almost useless, as consecutive time steps have very highly correlated features. Therefore, we arbitrarily choose to use a subset of 200 sets of features per emotion, updating around every 15 time steps. This produces 30400 training sets of features (200 sets of features per emotion times 8 emotions per day times 19 days). These are then fed into the Fisher Algorithm to produce a reduced dimensionality Fisher Projection.

5.3 Testing data

Using the Fisher Projection matrix, we calculate the posterior probabilities for all the sets of features (around 30 thousand data points) of the day we are testing and classify each one as coming from the emotion with the highest posterior probability.

5.4 Data labeling and moving window size

In the offline version, features were calculated from segments of data known to fully belong to only one emotion. In the online version, features are calculated based on data from a moving window. When the window includes the transition from one emotion to the next, features are calculated from data coming from 2 different emotions. It is not clear if these features should be included in the training of the classifier, and to which emotion. Similarly it is not clear if the classifier should be expected to classify these features to the previous or the next emotion during the testing phase. We expect our decisions on the training phase to influence the performance of the classifier in the testing phase.

The objective of an online emotion classifier is to first recognize as *correctly* as possible the emotional state of the user (high classification rate), and second to recognize it as *soon* as possible (high sensitivity). The former suggests a large window size, to minimize variance in the features within a class. It would also require that the features be considered as belonging to the previous emotion if most of the window is still in the previous emotion. On the contrary, the latter suggests a small window size, and the features of a window including the smallest part of a new emotion to be considered as belonging to the new emotion. Taking into account the above tradeoffs, we built and compared several classifiers, varying the following parameters:

W: We compare 5 different window sizes W (100, 200, 500, 1000, and 2000 time steps long). We also try combinations of 2, 3, 4 and all 5 window sizes. This is done by feeding to the Fisher Projection Algorithm a multiple of the 30 features calculated from each different window size for each data point (60 features when using 2 windows, 90 features when using 3 windows, etc.) Besides the 5 single-window cases, there are 10 pairs, 10 triplets, 5 quadruplets and 1 case of all 5 window sizes used, therefore a total of 31 different window size combinations.

W_{train_1}: A data point's features are used in the training of the new emotion when it is at least W_{train_1} time steps into the new emotion. We compare classifiers with $0 < W_{train_1} \leq W$. Normalizing provides $w_{train_1} = \frac{W_{train_1}}{W}$, $0 < w_{train_1} \leq 1$.

W_{train_2}: A data point's features are used in the training of the previous emotion when it is at most W_{train_2} time steps into the new emotion. We compare classifiers with $-\frac{W}{2} \leq W_{train_2} \leq \frac{W}{2}$. Normalizing provides $w_{train_2} = \frac{W_{train_2}}{W}$, $-0.5 \leq w_{train_2} \leq 0.5$.

W_{test_1}: A data point is expected to be classified as belonging to the new emotion when it is at least W_{test_1} time steps into the new emotion. We compare classifiers with $0 < W_{test_1} \leq W$.

W_{test_2}: A data point is expected to be classified as belonging to the previous emotion when it is at most W_{test_2} time steps into the new emotion. We compare classifiers with $0 \leq W_{test_2} \leq W$.

5.5 Definition of performance

In the case of an online algorithm, there are options for how to define performance. We could try to combine the posterior probabilities of all data points in one emotion and end up with an overall posterior probability from which we could classify the whole segment. Alternatively we could use simple voting among the classification results of all data points within one emotion to come up with an overall classification of the whole segment. None of these methods are natural, because in real life we will not know the emotion boundaries of the data we are trying to classify. (Although such pre-segmented classification is what was used in the facial and vocal expression recognition results alluded to earlier.)

Another measure of performance is the *data point* classification success rate. This is the ratio of the total number of data points correctly classified over the total number of data points in the day for which a classification was attempted. The results analyzed later use this definition of performance, but overall segment-classification performance will also be mentioned.

5.6 Results

In all 31 window-size combinations, the best results were obtained when the data were projected down to 7 $(c-1)$ Fisher dimensions. This is probably because the increase in training data helps in reducing the effect of noise in the features, making all 7 dimensions contain useful information, unlike in the offline version.

In all single-window cases, the larger the window size, the better the results. In all other cases, the larger the maximum window size used, the better the results.

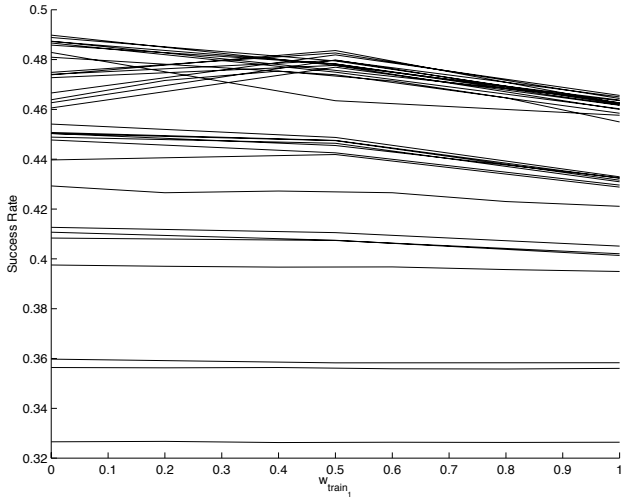

Figure 2: Success rate vs. w_{train_1} for different combinations of window sizes. Using data points from the start of a new emotion, even though the window still includes data from the previous emotion in the training, seems to slightly improve the results.

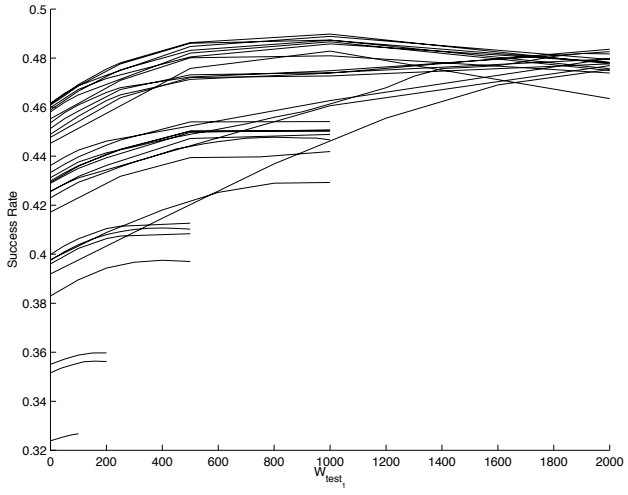

Figure 3: Success rate vs. W_{test_1} for different combinations of window sizes. Using data points from the start of a new emotion, even though the window still includes data from the previous emotion in the testing, seems to slightly worsen the results.

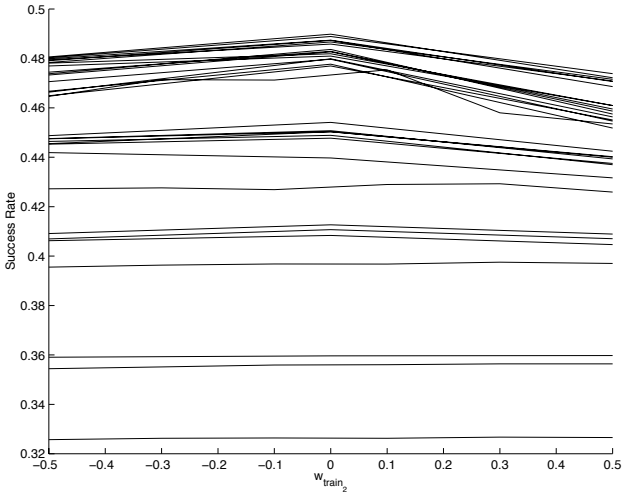

Figure 4: Success rate vs. w_{train_2} for different combinations of window sizes. Excluding data points from the end of an emotion segment in the training slightly improves the results.

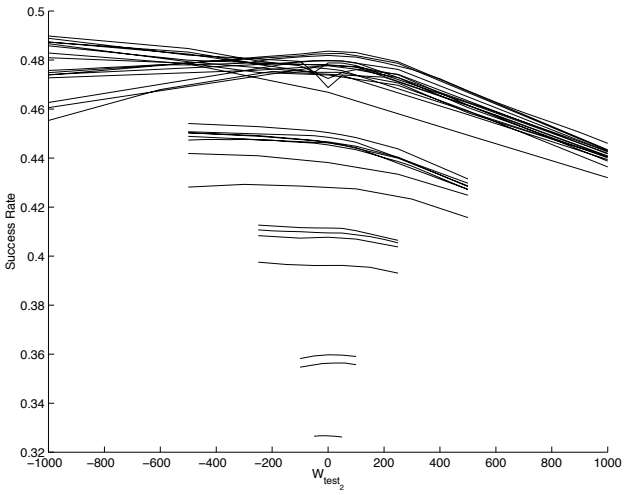

Figure 5: Success rate vs. W_{test_2} for different combinations of window sizes. Excluding data points from the end of an emotion segment in the training significantly improves the overall results.

In all cases, the results when using a combination of window sizes were at least as good, and in most cases significantly improved, over using any subsets of these window sizes.

Using data points from the start of a new emotion, even though the window still includes data from the previous emotion ($w_{train_1} << 1$) in the training, seems to slightly improve the results (Fig. 2). On the contrary, using these data points in the testing, slightly worsens the overall results (Fig. 3). Therefore, they help improve the training of the classifier, but they themselves are not classified as well as the middle section of the emotions.

Excluding data points from the end of an emotion segment ($w_{train_2} < 0$) in the training, slightly improves the results (Fig. 4). Similarly, excluding these data points from the testing significantly improves the overall results (Fig. 5). It seems that the data towards the end of each emotion segment does not help in the training of the classifier, and is not classified as well as the middle section of each emotion segment. We inquired with the actress who provided the data, and she indicated that trying to express a specific emotion steadily for 3 minutes often got boring; hence the data towards the end of each segment might not be as representative of the emotion as the earlier and middle portions of the segment.

The highest data point classification success rate was obtained when combining all 5 window sizes, and it was 48.98%. It should be noted that the segment classification success rate reached 60%, while the offline version using the same methods (Fisher Projection method, 30 features, without Day Matrix) gave a (segment classification) success rate of 56.87% (Table 3). Unfortunately, in most real-life applications, presegmented data will not be available.

6 Conclusions

The results here confirm and expand upon our earlier results, which suggested that there is significant information in physiological signals for classifying the affective state of a person who is deliberately expressing a small set of emotions.

Success rates above 80% when recognizing 8 emotions are extremely high, even compared to the other existing methods of emotion recognition. Nevertheless it is very important to keep in mind that these were intentionally expressed emotions, of only one subject, expressed in the same sequence every time (with unknown interactions between emotions) and all had similar duration (something not necessarily true with 'real' emotions). Therefore, plenty of work has to be done until a robust and easy-to-use emotion recognizer is built. A first step was made, by looking into online emotion recognition. Results from the online classifier were very encouraging, comparable to the offline version's results using the same features and methods.

In the future, an emotion recognizer could incorporate a model of an underlying mood, changing over longer periods of time. The question is how frequently should the estimates of the baseline be updated to accommodate for the changes in the underlying mood. Also, it appears that although the underlying mood changes the features' values for all emotions, it affects much less the relative positions with respect to each other. We are investigating ways of exploring this, and expect it to yield much higher recognition results.

References

[1] Dr. M. Clynes. *Sentics: The Touch of the Emotions.* Anchor Press/Doubleday, 1977.

[2] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley-Interscience, 1973.

[3] Irfan Essa and Alex Pentland. Coding, analysis, interpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, July 1997.

[4] J. Healey and R. W. Picard. Digital processing of affective signals. In *IEEE Int. Conf. on Acoust., Sp., and Sig. Proc.*, Seattle, 1998.

[5] D. R. Heise. Affect control theory: Concepts and model. *Journal of Mathematical Sociology*, 13(1-2):1–33, January-February 1987.

[6] A. K. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, February 1997.

[7] P. J. Lang. The emotion probe: Studies of motivation and attention. *American Psychologist*, 50(5):372–385, 1995.

[8] R. W. Picard. *Affective Computing.* The MIT Press, Cambridge, MA, 1997.

[9] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, November 1994.

[10] K. R. Scherer. Ch. 10: Speech and emotional states. In J. K. Darby, editor, *Speech Evaluation in Psychiatry*, pages 189–220. Grune and Stratton, Inc., 1981.

[11] L. Smith-Lovin. Affect control theory: An assessment. *Journal of Mathematical Sociology*, 13(1-2):171–192, January-February 1987.

[12] Elias Vyzas and Rosalind W. Picard. AAAI 1998 fall symposium, emotional and intelligent: The tangled knot of cognition. In *AAAI*, Orlando, FL, Oct. 1998.

[13] Y. Yacoob and L. S. Davis. Recognizing human facial expressions from log image sequences using optical flow. *IEEE T. Patt. Analy. and Mach. Intell.*, 18(6):636–642, June 1996.