Effect of Experimental Factors on the Recognition of Affective mental states through Physiological Measures

Rafael A. Calvo[1], Iain Brown[2], Steve Scheding[2]

[1]School of Electrical and Information Engineering, The University of Sydney
[2]Australian Centre for Field Robotics, The University of Sydney

Abstract

Reliable classification of affective mental states through processing of physiological response requires the use of appropriate machine learning techniques, and the analysis of how experimental factors influence the data recorded. While many studies have been conducted in this field, the effect of many of these factors is yet to be properly investigated and understood. This study investigates the relative effects of number of subjects, number of recording sessions, sampling rate and a variety of different classification approaches. Results of this study demonstrate accurate classification is possible in isolated sessions and that variation between sessions and subjects has a significant effect on classifier success. The effect of sampling rate is also shown to impact on classifier success. The results also indicate that affective space is likely to be continuous and that developing an understanding of the dimensions of this space may offer a reliable way of comparing results between subjects and studies.

Keywords: Emotion recognition, physiological signal processing, data mining, affective computing, human-machine interaction.

1. Introduction

It has been proposed [1] that the next big step in improving the way computers communicate with humans is to adopt an interaction paradigm that imitates aspects of human-human communication; namely, an awareness of a user's affective states (a combination of emotion and other mental states such as boredom or tiredness), so that the system can react to these states. Research into affective computing investigates how

computers can interpret and simulate emotions to achieve more sophisticated human-computer interaction.

There have been several approaches proposed for determining the affective states of subjects. Some of the more prevalent research techniques are based on facial patterns, gestures, speech and posture analysis as well as studies linking physiological response to emotional state. Each technique has its own challenges. Often, somatic motor expressions of emotion are heavily dependent upon the individual, making any global recognition system impossible. It is hoped that the affective–physiological connection is so rudimentary that strong similarities will be observable independent of the subject.

The great challenge of physiological signals is the abundance of available data. Hundreds of features can be extracted by considering all the physiological responses. Heart and muscle activity, brain activity, blood pressure, skin temperature, respiration, and sweat production are all rich sources of information concerning the physiological responses of the human body. Machine learning techniques for processing this data likely hold the key to understanding which responses are indicative of changes in mental and affective state.

This paper contributes a comparison of eight classification techniques and an analysis of the relative effect of a number of experimental factors on the success rate of affect classification. These factors include: number of sessions, number of subjects, sampling rates and classification algorithms used. Affective content is a rich source of information within human communication and learning as it helps clarify both the content and context of the exchange. Indeed, research has shown that along with cognitive processes, affective processes are essential for healthy human functioning [2]. Affect recognition, therefore, is one of the fundamental goals to be achieved in order to develop more effective computer systems. While the primary research focus is to investigate affective systems, research in this area has the potential to strongly benefit associated fields such as psychology and teaching.

Section 2 reviews the literature, focusing on psychophysiological techniques which use the subject's physiological signals as input to a classification algorithm. Section 3 presents an experimental session, and describes the protocol followed for recording physiological signals from three subjects while they elicited a sequence of emotions. The tools used to record and then process the signals for this session are also described. Section 4 provides some details about the eight classification techniques evaluated and the research questions that arise on how different humans elicit emotions (e.g Do we elicit emotions consistently? Do all humans do it in similar ways?). The basic tenet of these open research questions is that the accuracy of the classifiers provides an indication of how complex the emotion identification in a given data set is, and that this complexity is at least partially due to way humans elicit emotions. Section 5 looks at the results obtained for the different classification techniques discussing their accuracy and training time in different situations. Section 6.

2. Background

In recent years several studies started investigating the potential for using biometric data for the classification of affective state [3-7]. Despite a longstanding debate amongst psychologists on the so called 'autonomic specificity', or the possibility of using autonomic nervous system (ANS) recordings to recognize affective state. This recent work [3-7] provides some evidence that the discrimination among some affective states is possible,

Emotion recognition is inherently multi-disciplinary, and draws on the fields from psychology, physiology, engineering and computer science. It is not at all surprising, then, that the approaches taken to study in this field also have a tendency to vary greatly. While the research goals of each study overlap there is wide variety in equipment used, signals measured, features extracted, evaluations used and in the format of presented results. These studies have had different levels of success (e.g. Picard, 81%, Kim, 78.4%), and with different limitations

Picard [4] built automatic classifiers for the recognition of emotion and showed the relationship physiology and the elicitation of emotions, and that it is consistent within an individual, but it provides no insight as to whether there is any consistency between individuals. The study by Kim [6] uses a large number of subjects (young children, 5-8yrs). Their recognition accuracy was much lower, however this maybe due to the lack of consistency in their sample population. This study also addressed the issue of the inherent subjectivity of the subject-elicited technique (individual understanding of what emotive nouns refer to), by using an immersive, multi-modal environment to trigger the emotion. However it is difficult to create a multi-modal environment where each of the modes is coherently and seamlessly presented with the others. In cases where this is not achieved, the lack of coherency between triggering stimuli has been shown to heavily reduce the effectiveness and believability of the environment [8], which in turn will influence the quality of emotions elicited.

This paper investigates some of the effects in classification results by variations in factors such as: number of sessions, number of subjects, sampling rates, and algorithms used for classification. The study also considers the subjective evaluation of each affective elicitation in the three dimensions of arousal, valence and dominance. Until the effects of individual decisions made in the formulation, processing and analysing of the different papers mentioned is properly understood it is hard to see how the results of each study can be effectively viewed together.

3. Subjects and Methods

The signals chosen for this study were the electrocardiograph (ECG), electromyograph (EMG) and galvanic skin response (GSR). The ECG measures the voltage change across

the chest due to the electrical activity of the heart. In this case the signal was measured between the wrists and used an electrode connected to the inside of one ankle as a reference node. The EMG measures the electrical impulses across muscle groups that are generated by activation of that muscle group. Electrodes were placed on either end of the masseter muscle group and a reference electrode was placed on the inside of one of the ankles. The masseter muscle group has been used in previous studies [9], [5] and was chosen due to its reliability and ease of measurement. GSR can refer to many different readings; in this study variation in skin conductance was measured. Skin conductance is directly related to sweat production and is therefore has been used directly to measure anxiety levels, however in this study the features extracted from GSR are treated numerically. GSR was measured by subjects placing their index and middle fingers on each of two electrodes in a plastic bar. Subjects were asked to maintain a constant pressure on the electrodes as a variation in pressure affects the results. The equipment used for recording the signals was a Biopac M150 base unit with the appropriate modules for ECG, EMG and GSR. Signals were recorded to a HP Tablet PC using the AcqKnowledge 3.8.2 Software supplied with the equipment.

In this study the combination of factors used were; subject-elicited, lab setting, feeling, open-recording and emotion-purpose [4]. It was believed that though there was a small risk that the lab setting, open-recording and subject awareness of the study's purpose may affect the quality or effectiveness of the emotions elicited, it was necessary to do this.

A modified version of the Clynes protocol for eliciting emotion [10] was chosen for generating the subject emotion. The Clynes protocol was used in an earlier study by Picard [4] and asks subjects to elicit eight distinct emotions, (*no emotion, anger, hate, grief, platonic love, romantic love, joy, and reverence*). The Clynes protocol typically uses physical expression to give somatosensory feedback, given that the correct equipment was not available, subjects were offered a stress ball to hold in their free hand to use as an object of physical expression. Each emotion was elicited for a three minute period, separated by a period of rest.

After subjects were prepared for the study the emotions were elicited in order. In this study subjects were not told exactly what was meant by each emotion (other than its name) allowing individual, subjective, interpretations of each affective label. After each emotion was elicited, subjects were asked to rate the emotion in each terms of Arousal, Valence and Dominance on the Self Assessment Manikin pictorial scale [8]. Three subjects (Male 60, Male 40, Female 30 yrs old) Three sessions were recorded for each subject on different days. The sessions with Subject 1 were recorded at 40Hz, while the sessions of Subjects 2 and 3 were recorded at 1000Hz, after deciding to see the effect of a higher sampling rate on the ability to classify the data. Although the number of subjects is small, the aggregate data is very large. Each of the three sessions for each three subjects contains 24 minute of recordings, for 3 physiological signals at 1000 samples per second.

The raw data was preprocessed using Matlab. The signal data was organised into thirty overlapping 30 second windows for each emotion recording in each session. 120 features were extracted for each 30 second window using the Augsburg Biosignal Toolbox [12]. The features extracted were primarily the mean, median, standard deviation, maxima and

minima of several characteristics in each signal. The data was then processed by WEKA, a machine learning toolbox [9].

4. Classification

Eight classification algorithms were evaluated using 10-fold cross validation:

1. ZeroR: predicts the majority class in the training data; used as a baseline.
2. OneR: uses the minimum-error attribute for prediction [10].
3. Function Trees (FT): classification trees that could have logistic regression functions at the inner nodes and/or leaves.
4. Naïve Bayes: A standard probabilistic classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data [9].
5. Bayesian Network: using a hill climbing algorithm restricted by sequential order on the variables, and using Bayes as optimisation criteria.
6. Multilayer Perceptron (MLP): using one hidden layer with 64 hidden units.
7. Linear Logistic Regression (LLR) using boosting.
8. Support Vector Machines: Finds the maximum margin hyperplane between 2 classes. Weka's SMO with polynomial kernel was used [11] with c=1.0, epsilon=1e-12.

An underlying hypothesis of this study is that different emotions manifest themselves in distinct physiological states. Another hypothesis is that the classifiers' performance gives an indication of an internal 'consistency' of the data. If the performance is bad for all algorithms, the data is harder to model. A number of specific problems arise when the classifier performance is used to make other inferences, including:

1. Intra-Subject, Single Session

Subjects might not elicit emotions in the same way on different days. To build a classifier and to test it on data from a single session means excluding the factors of inter-session variation Even for classifiers 'custom' built for a single subject, most applications would require high multisession accuracy.

2. Intra-Subject, All Sessions

A subject specific classifier can be trained and tested by combining data from a number of sessions. By combining the windows from the three sessions for each subject into a single data set, the classifiers' accuracy indicates how variation in affective elicitation deteriorates the accuracy. This is probably caused by differences in the appraisal of emotion, intensity and quality of the elicitation (how close to the emotion the subject was able to elicit).

3. Inter-Subject

The 'universality' of emotions –the assumption that different people elicit emotions in a similar way- has been disputed. Depending on the application, it might be necessary to build a classifier based on data recorded from another subject. For this evaluation, data included both the day-to-day baseline variation in emotion and also the variation in

subject interpretation of affective labels. Consequently seeing how the inter-subject data set classification compares to the combined and individual sessions will give insight into how much variation exists between subjects.

5 Results

Table 1 shows the classifiers' accuracy and training time on a PC with an Intel Core 2 Duo Processor (1.83GHz) and 2GB DDR2 RAM. MLP had the highest percentage of correctly classified samples, however data sets take a long time to process (36 minutes to process 9 minutes of recorded signals), making it unsuitable for real time applications. SVM, LLR and Functional Tree (FT) algorithms are faster, and give high accuracy. Of these methods the SVM algorithm gives the most consistent results for the shortest processing time. The FT algorithm also demonstrated unsuitability by failing to be compatible with all data sets. Though often quicker, the remaining algorithms give significantly lower or less consistent results than the SVM algorithm. Hence the SVM algorithm was used as the primary algorithm for comparing the confusion matrices and misclassified sample analysis. Table 1 also shows the processing time of each algorithm for a 3min recording data set.

There is a noticeable decay in classifier accuracy as the data sets become more complex, however even the most complicated data still gives 42% success using the chosen SVM algorithm. This remains three times higher than chance classification.

	ZeroR	OneR	FT	Naïve Bayes	Bayes Net	MLP	LLR	SVM
S2D1– 40	12.5%	50.4%	89.2%	66.3%	81.3%	92.9%	90%	94.6%
S2D1–1K	12.5%	48.3%	96.7%	61.7%	N/A	97.1%	97.5%	95.8%
S2DA–40	12.5%	55.3%	76.7%	43.6%	64.3%	90.8%	72.6%	74.7%
S2DA-1K	12.5%	59.2%	88.9%	38.6%	N/A	97.8%	86.9%	85.7%
Time to Process	0 s	1 s	1.5min	2 s	N/A	36min	8min	41 s
SADA- 40Hz	12.5%	55.4%	N/A	22.8%	59.3%	70.6%	41.8%	42.2%

Table 1: Results of the different classification algorithms used for each data set. S#D# refers to the subject number and session (day) number. 40/1K refers to the sampling rate, (Hz).

5.2 Variation of results across different sample rates

Table 2, gives a summary of the classifier success for each of the different data sets. Individual sessions displayed strong classifier success across all data sets. For individual sessions the difference in sample rate is fairly small, with classifier success varying by only a few percent in any case. In all but one case the accuracy for 1000Hz is better than the 40Hz equivalent. This is not shown to be true of other algorithms, and is most

profoundly noticed where BayesNet failed to process the higher sample rate data set. The results show a progressively increasing difference between the success rates of classification for high and low sample rates as the data sets become more complicated. Although this evidence is far from conclusive it does suggest that the sample rate is a factor to be considered when making physiological recordings.

The different accuracy for the 40Hz and the 1000Hz data sets is not restricted to the SVM classifier. It should also be noted that the effect of sample rate variation was more pronounced in some, but not all techniques. The BayesNet technique for example showed a tendency to fail at higher sample rates, as did the Functional Tree approach. The more consistently correct classifiers, MLP, LLR and SVM, however, all showed classification improvement at higher sample rates. More detailed studies will provide a more complete picture of the effect sample rate has on emotion identification.

Subject	1	2	3	Combined Sessions
1 – 40Hz	96.3%	92.1%	95.4%	80.4%
2 – 40Hz	94.2%	97.5%	95.8%	74.7%
2 – 1000Hz	95.8%	97.1%	98.8%	85.7%
3 – 40Hz	90.5%	95%	92.1%	68.1%
3 – 1000Hz	99.2%	99.6%	96.3%	79%
All Subjects (40Hz)	N/A	N/A	N/A	42.2%

Table 2: Percentage of samples correctly classified for data sets using the SVM Algorithm.

5.3 Variation of results across different sessions and subjects

Comparing the results of the individual sessions, some emotions were consistently poorly classified, others consistently well classified, and others varied from session to session. Platonic love and romantic love stand out as emotions that are often misclassified, while anger and the no emotion baseline were consistently well classified. Table 3 shows percentages of emotions misclassified as other types. For example, Subject 1's romantic love samples are 4% misclassified as 'No emotion' and 3% misclassified as 'Hate'.

The consistency of emotion elicited is better identified from the combined data set of all of a subject's sessions. Subject 3, for example, shows very high classification success in each session individually, but displays the lowest classification success in the combined data set. This suggests that for each individual session, the consistency of emotion elicited for each 3-minute block was very good, but that the character of emotion elicited from session to session was not as consistent. Subject 1, in contrast, shows greater variation in the individual sessions, but better consistency across the three sessions.

In the intra-subject data sets, all three subjects displayed relatively high misclassification in romantic love. The confusion matrices for the three subjects showed one subject with high misclassification towards hate, one with high misclassification towards platonic love and the other with a misclassification split between anger, joy and

platonic love. These variations are subject dependent and are likely caused by developed associations as well as variations in mood, concentration and interpretation of the meaning of affective labels.

In the inter-subject data set, romantic love, hate and platonic love showed the worst results for classification, while anger, reverence and joy showed the best classification results. Anger is correctly identified but other emotions tend to be misclassified as anger.

Data Set	**Worst Classified**			**Best Classified Emotion**			**Most Commonly Misclassified**
	1	2	3	1	2	3	
S1– 40	J, Re	Re, Ro	P, Ro	A, G, N, P	G	A, G, J, N, Re	Romantic(4%N, 3%H), Reverence(5%J, 3%N), Hate(2%A), Platonic(3%Ro, 2%H)
S2 – 40	J, H, Ro	P	G	A, N, Re	A, H, J, No, Re	H, J	Platonic (3%G, 3%J), Grief (4% P)
S2 – 1K	G, H, J	P, J	-	P	A, H, N, Re, Ro	All	Platonic(4%J), Joy(4%P), Hate(2%G), Grief(2%H, 2%A)
S3 – 40	A, H, Ro	P, G	Re, Ro	Re	A, H, N	G, N, P	Romantic(8%J, 2%Re), Reverence(7%Ro), Platonic(3%G, 2%J), Hate(4%A, 3%G)
S3 – 1K	-	-	Re, Ro	All	All	A, G, H, N, P	Reverence(4%Ro, 1%A), Romantic(4%Re)
S1DA – 40	Ro, H, J			A, No, P			Romantic(10%H, 6%N, 5%J,5%G), Hate(12%A, 5%Re, 4%J), Joy(16%Re, 8%Ro)
S2DA – 40	N, G, P			Re, A			No Emotion(14%J, 10%P, 7%Ro, 5%G), Grief(8%J, 8%N, 4%H, 4%P, 4%Ro), Platonic(13%Ro, 7%J, 7%N)
S2DA – 1k	P, G			A, Re			Platonic(14%Ro, 9%J, 6%N), Grief(10%N, 6%H)
S3DA – 40	Ro, P, Re, G			A, H, N			Romantic(17%J,14%A, 13%P, 7%Re), Platonic(9%Ro, 9%J, 8%G, 7%Re), Reverence(8%G, 6%J, 6%A), Grief(8%P, 6%Re, 6%A)
S3DA – 1k	P, H, Re			N			Plat(19%G, 8%A, 8%J), Hate(12%A, 8%Re), Reverence(4%H,4%N, 4%P, 4%Ro)
SADA	Ro, H, P			A, Re, J			Rom(14%P, 12%A), Hate(18%A, 13%N, 12%Re), Plat(15%Ro, 12%A, 10%N), Grief(11%A, 11%P), No Em(11%J, 11%P), Joy(13%Re, 9%Ro), Rev(12%N, 11%A), Anger(10%H, 8%P)

Table 3: Misclassification results for all data sets.

6. Conclusions

The method used in this study utilised a subject-elicited, lab setting, feeling, open-recording and emotion-purpose framework. This particular choice of factors highlighted the individual variation in subject's interpretation of emotive labels. As a consequence,

future studies will utilise a detailed description of the emotion to be elicited, or use the induced-emotion approach. Subjects also had preferred techniques for eliciting the emotions, some preferred to visually focus on something, while another preferred to elicit with closed eyes. For this study, the process had the luxury of being flexible and each subject was able to find a way to elicit emotions that they found comfortable.

The strong consistency of classifier success (> 90%) across the nine primary data sets (Table 2) supports the hypothesis of correlation between emotion state and physiological state. Although there is no guarantee that the emotion elicited is an accurate portrayal of the affective label requested, the high success in classification does show that the physiological manifestation caused by each of the eight categories was sufficiently distinct to allow discrimination and classification against the 7 other categories. If further data sets continue to show good discrimination, they will add to the mounting case in support the hypothesis of correlation.

A noteworthy result was the consistency of misclassification within a subject's data sets. Subject 3's romantic love samples were often misclassified as joy, and all subjects showed some misclassification between the negative emotions; anger, hatred and grief. Subjects also showed variation between sessions of which emotions were well classified, and which were relatively poorly classified, this may point to influence from the variation in day-to-day baseline as noted by Picard [2]. It is likely, for example, that on a day where a subject is feeling sad, that many samples might be misclassified as grief, while emotions which are sufficiently distinct, such as joy, might show strong classification success in contrast.

Further studies will continue to use the SAM diagrammatic survey for subject self assessment, but this will be supplemented with a quality assessment rating, ("How well did you feel you elicited the required emotion?"). This rating will help give an understanding of why misclassifications occur within sessions, and whether these misclassifications are predictable.

This study was successful in demonstrating that key factors such as number of sessions, number of subjects, sampling rates, and algorithms used for classification, all play a role in the success of classification. This study also supports the hypothesis that emotions lie in a continuous space. A future challenge will be to identify the axes of this space and determine an appropriate transform from physiological signals into these metrics.

While this study gives an important foundation for recognising the importance of these factors a complete understanding of the ways in which these factors do affect the results can only be properly obtained through more detailed studies.

Acknowledgments

The authors would like to thank the subjects who volunteered their time for this project.

7. References

[1] R. Picard, *Affective Computing*: The MIT Press, 1997.

[2] A. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*: Harper Perennial, 1995.

[3] E. Vyzas and R. Picard, "Offline and online recognition of emotion expression from physiological data," 1999.

[4] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 23, pp. 1175-1191, 2001.

[5] J. Wagner, N. Kim, and E. Andre, "From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification," *Multimedia and Expo, IEEE International Conference on,* vol. 0, pp. 940-943, 2005.

[6] K. Kim, S. Bang, and S. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and Biological Engineering and Computing,* vol. 42, pp. 419-427, 2004.

[7] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion Recognition Using Bio-sensors: First Steps towards an Automatic System," in *Affective Dialogue Systems*, 2004, pp. 36-48.

[8] M. M. Bradley and P. J. Lang, "Measuring emotion: the Self-Assessment Manikin and the Semantic Differential," *J Behav Ther Exp Psychiatry,* vol. 25, pp. 49-59, 1994.

[9] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*: Morgan Kaufmann, 2005.

[10] R. C. Holte, "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets," *Machine Learning,* vol. 11, pp. 63-90, 1993.

[11] C. P. John, "Fast training of support vector machines using sequential minimal optimization," in *Advances in kernel methods: support vector learning*: MIT Press, 1999, pp. 185-208.