Classification Accuracy from the Perspective of the User: Real-Time Interaction with Physiological Computing

Stephen H. Fairclough
Liverpool John Moores University
Byrom Street, Liverpool, UK
s.fairclough@ljmu.ac.uk

Alexander J. Karran
Liverpool John Moores University
Byrom Street, Liverpool, UK
alexander.j.karran@gmail.com

Kiel Gilleade
Liverpool John Moores University
Byrom Street, Liverpool, UK
gilleade@gmail.com

ABSTRACT

The accurate classification of psychophysiological data is an important determinant of the quality when interacting with a physiological computing system. Previous research has focused on classification accuracy of psychophysiological data in purely mathematical terms but little is known about how accuracy metrics relate to users' perceptions of accuracy during real-time interaction. A group of 14 participants watched a series of movie trailers and were asked to subjectively indicate their level of interest in a binary high/low fashion. Psychophysiological data (EEG, ECG and SCL) were used to create a binary classification of interest via a Support Vector Machine (SVM) algorithm. After a period of training, participants received real-time feedback from the classification algorithm and perceptions of accuracy were assessed. The purpose of the study was to compare mathematical classification accuracy with the perceived accuracy of the system as experienced by the users. Results indicated that perceived accuracy was subject to a number of psychological biases resulting from expectations, entrainment and development of trust. The F1 score was generally a significant predictor of perceived accuracy.

Author Keywords

Physiological Computing; EEG; Psychophysiology; tagging media

ACM Classification Keywords

H.1.2. User/Machine Systems; H.5.1 Multimedia Information Systems; H.5.2. User Interfaces; I.2.6 Learning; I.5.1 Models

CHI 2015, April 18–23, 2015, Seoul, Republic of Korea.
Copyright 2015 © ACM 978-1-4503-3145-6/15/04...$15.00.
http://dx.doi.org/10.1145/2702123.2702454

INTRODUCTION

Psychophysiological data can be collected implicitly during human-computer interaction and used to represent the affective or cognitive state of the person. This dynamic and quantified representation of the user represents a basis for adaptive software mechanics. The same logic can be applied to the derivation of media tags during the passive consumption of movies, music or still images. In this case, psychophysiological data is collected and classified as the person experiences the media, which are subsequently classified to yield tags related to emotional experience [1]. This form of implicit human-centred tagging [2] provides a method for understanding human behavior and the effects of media on the user – as well as enabling a number of interactive media applications, such as interactive narratives tailored to induce a specific psychological state in the viewer [3]. This type of implicit interaction represents a form of physiological computing [4] constructed upon a generic control loop [5].

Physiological computing as implicit interaction

Research on human-centred tagging of media has focused on the measurement of emotional states [6], which may be categorized as discrete states (happy, angry, calm etc.) or within the two-dimensional space of the circumplex model [7]. The generation of affective tags provides a useful means to: (1) assess whether media produced the intended emotional state, and (2) to assemble a repertoire of media over a period of time known to induce specific emotional states in that particular individual. The former represents a 'test audience' usage case where automated psychophysiological quantification effectively replaces subjective self-report. The second instance emphasizes the construction of personalised database (of media clips) via human-centred tagging that may be utilized to elicit desirable emotional states and mitigate undesirable ones, e.g. affective music player [8]. Recent work in the domain of cultural heritage [9] departed from this tradition by measuring a psychological state of interest as a cognitive-affective state. In this example, measures from EEG and autonomic psychophysiology were combined to operationalize the degree of interest elicited by exposure to media. Interest is defined as a combination of: (1) attention, (2) stimulation and (3) high levels of either positive or negative emotion. Tagging media with respect to items that elicit high interest can be used to create personalised and engaging trajectory through any information space [10]. This approach is concerned with the classification of magnitude along a single psychological dimension (high vs. low levels of interest) as opposed to the categorization of discrete emotional states.

Classification accuracy

Current research on the classification of psychophysiological data places enormous emphasis on the application of machine learning algorithms in this context [11]. The general methodology for the construction of a classifier is to generate a training set, which is subsequently used to train a classifier and represents a template for all subsequent acts of categorisation. The first obstacle for classification is the derivation of an optimal set of training data. It is hoped that training data provides a good mapping in terms of a quantitative discrimination (between the states to be classified) and lead to a reduction of classification errors. Those factors most likely to actively contribute to classification errors are:

1. The influence of noise from non-psychological sources, noise is a 'fact of life' for ambulatory psychophysiology
2. The degree of divergence between the estimated mapping provided by the training set and the best mapping possible. This factor is determined by the representativeness of the training set. The degree of divergence between what is measured by the system now and what was measured during training is called Bias.
3. The sensitivity of the classifier to the training set. It is important to note that different approaches to signal classification differ with respect to their susceptibility to specific and idiosyncratic qualities of the training dataset. The degree of sensitivity to the training set exhibited by the classifier is called Variance.

Therefore, a good dataset for training may be defined as one that encompasses the full range of physiological responses for classification and has been acquired under realistic conditions. The acquisition of such a dataset often requires a sustained period of monitoring and a cumulative approach to data collection, i.e. training dataset becomes more inclusive as data is acquired over time/episodes of usage. The accumulation of training data over a period of time should reduce bias and variance as the resulting data that informs the process of classification is both representative and generalizable [12].

Perceived accuracy & user experience

The quality of the classification emerging from the training dataset is generally assessed using 'hard' markers of mathematical accuracy, e.g. cross-validation. It is assumed that accurate classification as represented by a mathematical index will translate into good performance with respect to the categorization of psychological states within the context of human-computer interaction. This assumption disregards the obvious fact that users of physiological computing systems will possess varying degrees of subjective self-awareness. The co-existence of 'hard' markers of classification accuracy (from the system) and 'soft' markers of subjective self-awareness is characteristic of interactions with physiological computing systems. The combination of 'hard' and 'soft' markers of classification accuracy yields a third category of accuracy – which equates to the perceived accuracy of the system from the perspective of the user (see [4] for a more detailed discussion).

The perceived accuracy of the classification engine is an important determinant of the user experience. A system that is perceived to be accurate in the short-term will create a positive impression that encourages further use. In the long-term, an acceptable level of perceived accuracy will engender trust in the technology [13]. The question of what is an acceptable level of accuracy for a physiological computing system has been addressed by previous research [14,15]. These authors simulated various levels of accuracy with respect to control of an input device and task difficulty respectively in order to explore levels of user acceptance and tolerance for system error.

Current study

The current paper will focus on the relationship between perceived classification accuracy and mathematical accuracy of a physiological computing prototype working in real-time. A system was designed to make a binary (high/low) classification of the interest level experienced by the user during the viewing of 40 movie trailers. Data from EEG, ECG and skin conductance level were collected, quantified and classified in real-time. Classification of psychophysiological data was achieved using a Support Vector Machine (SVM) working on a subject-dependent basis, i.e. a SVM was generated that was specific to each individual participant.

The training dataset used to generate the SVM classifier was obtained at four different points in time as each participant viewed the series of movie trailers. The initial system build for classification was achieved on the basis of a small dataset whereas the final system build utilized a significantly larger training dataset for classification. Four system builds are included in the experimental design in order to explore how the acquisition of training data influences both 'hard' and 'soft' markers of classification accuracy.

At the end of each movie trailer, the participant was required to provide a subjective binary estimate of interest and subsequently received feedback on the classification of interest produced by the system.

The purpose of this paper is to assess users' perceptions of psychophysiological classification based upon real-time feedback within the context of an interaction with a working system. We also wished to explore the feasibility of subject-dependent classification where the classification algorithm for psychophysiological data is trained to each user following a brief period of initial exposure. The focus of current research in this field is on 'hard' mathematical markers of classification accuracy that are often derived on a retrospective basis [11]. Our contribution is to study the association between these 'hard' markers and the perception of accuracy from users who have received explicit feedback from the system during a real-time interaction.

The study was designed to explore four research questions:

1. How does the accumulation of training data influence both mathematical classification accuracy and perceived classification accuracy? It is assumed that both estimates of classification will improve over each successive build of the system (due to the cumulative acquisition of training data and an associated reduction of variance and bias in the training set)
2. How do user perceptions of accuracy change with sustained exposure to the system? Is there evidence for any systematic bias?
3. For subject-dependent classification of psychophysiology in real-time, how long is required for a system to generate a sufficient corpus of data to train a classifier?
4. What is the relationship between mathematical accuracy and perceived accuracy? Do users tend to over- or under-estimate classification accuracy?

SYSTEM DESCRIPTION

Conceptual Model

The conceptual model for the system used in the study is illustrated in Figure 1. Four types of measure are collected from the participant, each one maps onto three process sub-components, which are forwarded to the classification engine in order to categorise the level of interest as high or low. The concept of interest is divided into three types of process: (1) activation, i.e. does the content stimulate the autonomic nervous system? (2) cognition, i.e. does the content engage novel problem-solving and consolidate memory formation in the rostral prefrontal cortex of the brain, and (3) valence, i.e. does the content provoke a strong positive or negative emotional response (see [10] for a detailed description of how the concept of interest and associated measures were derived). If psychophysiological data indicates that content is stimulating, cognitively engaging and emotionally provocative, it is deemed to be of interest to the user. The activation processor is indexed by autonomic indicators, such as heart rate (HR) and skin conductance level (SCL). The level of electrocortical activation from the EEG signal over the rostral prefrontal cortex [16] was used as a measure of cognition. Valence was represented by frontal EEG asymmetry [17].

Measures from the physiological sensors were forwarded to the component processors. These features were used as inputs to the classification engine in the composite model that expressed interest as a binary (high/low) state. The output from the classification represents a control input for a hypothetical process of software adaptation.

Real-time interest classification

A data processing pipeline was designed to: (1) extract the relevant psychophysiological features, (2) quantify each category of psychophysiological response, (3) classify the response as a binary state of interest and label this response. Because the system works in real-time on a subject-dependent basis, an initial exposure of the participant to video clips was used to train the first build of the SVM classifier. The classifier was re-trained on a new data set three times after this initial build, e.g. every 7-9 movie trailers, in order to deliver a steady accumulation of the training dataset.

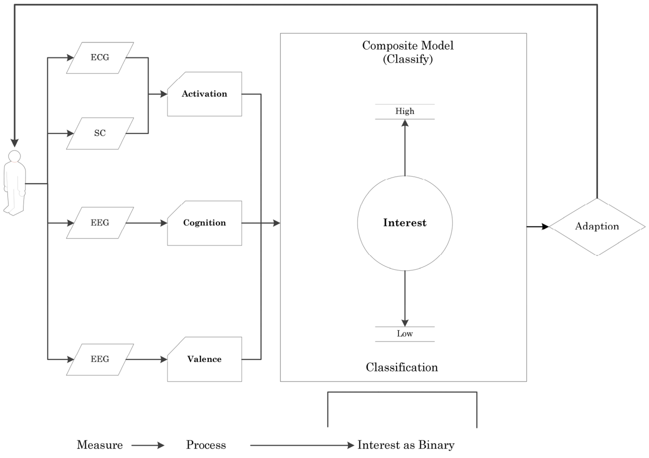

Figure 1. System Model

Due to the complexity of the application framework two elements have been extracting from the overall structure to highlight how they functioned within the application. The first element took the form of a video player sub-window, which also acted as the means for gathering and processing the subjective responses to each 60sec video after viewing. Execution of the video player function drew a clip from the pool of video material. After the video had been displayed a new window appeared to prompt the user to deliver a subjective response (high/low interest). These responses were processed and forwarded to the export module for association with psychophysiological responses for that particular video.

The second element received feature vector output from the data export process and checked whether it was necessary to construct a new version of the classifier. If true, a check was performed to determine if the current request corresponded with the first build of the classifier. If this condition was satisfied, a further check was performed to determine whether two instances of both classes (i.e. two examples of high and low class data) existed within the training dataset. If true, a classifier was constructed. However, if a classifier already existed and a new classifier build was required, then data collected for the current stimulus period (i.e. corresponding to approx. 10 movie trailers) was added to the existing training set and a new classifier was constructed based upon this aggregated dataset. If no new classifier build was required, the train classifier process was bypassed and new vectors were classified and output.

Data acquisition

Psychophysiological data was imported in real-time from two ambulatory physiological sensor technologies: the Nexus X MkII © (used to capture autonomic ECG and SCL responses) and the Enobio © (used to capture EEG responses). These data were buffered internally and the process pipeline split into two top-level protocols to process autonomic data and EEG data respectively.

Physiological responses from the autonomic system were measured using the Electrocardiogram (ECG, sampled from the torso) and SCL (distal phalanges, second and forth finger, non-dominant hand). Both channels were sampled at 512Hz. Three channels of electroencephalographic (EEG) data were recorded, measuring alpha (8-12Hz) and beta (13-30Hz) activity, using the Enobio wireless 3-channel sensor (sampled at 250Hz) with ground contacts on left ear lobe and inner ear (Starlabs Inc). A mobile sensor forehead band was fitted and nasion aligned to ensure sensor placement at Fp1, Fp2, Fpz and electrodes attached.

The autonomic data processor included filtering for both electrocardiogram (ECG) and skin conductance level (SCL) of a 0.5 to 35Hz bandpass and 35Hz lowpass respectively. The ECG data was subsequently forwarded to a beat detection process in order to determine the mean and standard deviation of the inter-beat-interval (IBI). The filtered SCL data was forwarded to the epoch analysis module to produce the mean and standard deviation of SCL. The resulting derivatives from ECG and SCL were forwarded to a feature store for eventual output.

Data analysis

The EEG data processor performed filtering (Bandpass 0.05-35Hz) and epoch analysis on three channels of EEG activity derived from Fp1, Fp2 and Fpz. These data were subjected to a Fast Fourier Transform (FFT) analysis to determine the total amplitude spectra of the signal in the alpha and beta bandwidths. Data from the FFT were forwarded to calculate cognition and valence. The former was expressed as a ratio of electrocortical activation (β /α) at sites Fp1, Fp2, and FPz. Valence was represented by frontal EEG asymmetry, which is expressed as the difference between the natural logs (ln) of the total power in the alpha band of the right and left hemispheres. The resulting eight derivatives (see Table 1) are forwarded to the feature store and exported to the train classifier process

Measure	Signal Derivatives for Classification
Heart Rate	IBI (mean) IBI (s.d.)
Skin Conductance	SCL (mean)
EEG	Ratio β /α (Fp1) Ratio β /α (Fp2) Ratio β /α (Fpz) ln(αFP2)-ln(αFP1)

Table 1. A list of measures and those signal derivatives used as input to the classification algorithm.

All derivatives were extracted from a 60 second epoch (i.e. duration of each movie trailer) to deliver a total of 40 stimulus events. For autonomic measures features were captured using a 12sec data window with a moving window of 6sec. For EEG, features are captured using a 12sec of data with moving 6sec window. This approach was used to construct a feature vector every 6sec resulting in 10 -1 (due to the overlapping data 12 second data windows) per sixty seconds stimulus epoch. This approach delivered a potential of 360 – (n * 9) classification vectors, where n equals the total number of vectors used to train the classifier initially.

Classification

The training of the classifier occurred within MatLab using the deployment command line processor for real-time data interaction

To train and ascertain estimated performance of the classifier in real-time, the sequential minimal optimisation [18] and hold-out cross-validation methods are used on the aggregated training data. The hold-out cross-validation method partitions the data into two parts, by randomly assigning data to either training or testing sets, ensuring that the classifier is trained and tested with novel data and is analogous to a real world task. This method of cross-validation has been shown to provide a more accurate assessment of potential classifier performance in comparison to k-fold cross-validation when applied to small datasets, such as those gained from real-time applications [19]. When coupled with a loose grid search algorithm, these methods form the basis for the training and parameterisation of the SVM in real-time, providing the optimal settings for the box constraint and sigma values of the SVM radial basis function (RBF) kernel for each new instance of training data as the system is used. That is, for each new build of the system, training data is aggregated and cross-validated to create a new classifier in real-time, in this instance to prevent over fitting of the classifier to the training data and reduce computation time the box constraint and sigma values are set to a maximum of 2.

Using this approach a feature vector was constructed every 6 seconds resulting in 10 -1 (due to the overlapping data 12 second data windows) classification vectors per sixty second stimulus epoch. This gave a potential of 360 – (*n* * 9) classification vectors, where *n* equals the total number of vectors used to train the classifier initially. Thus, nine classifications were performed in real-time per stimulus epoch and a majority vote was performed between class outputs at the end of each epoch to determine the resulting class for that epoch.

METHODOLOGY

Participants

16 participants (9 female) aged 19-25yrs. took part in the experiment. However, only 14 participants data were used for analysis. Data from two participants were excluded because their responses did not meet a criterion that was required to train a classifier on four occasions. Specifically, these two participants did not produce the required instances of high and low classes over the maximum 10 videos needed to train the system (i.e. the number of videos remaining in the database was insufficient to permit three further iterations of the classifier). All procedures and measures were approved by the University Research Ethics Committee prior to data collection.

Experimental design

The experiment was designed as a repeated measures study (i.e. the same participants took part in all build sessions). A "Wizard of Oz" [20] prototyping approach was derived in order to convey feedback to the user in real-time. The application included four build phases over the course of the experiment designed to investigate how the accumulation of data into the training set influenced classification accuracy:

- build 1 was the initial classifier training phase and required responses from at least two of each of the target classes (high and low)
- build 2 aggregated the database of responses from build 1 into a new training data set and the SVM was re-built
- build 3 aggregated the responses from builds 1 and 2 into a new training data set and the SVM was re-built
- build 4 aggregated the responses from the previous 3 builds and into a new training data set and the SVM was re-built

Materials

The stimulus material took the form of movie video trailers from four genres of film: science fiction, comedy, action and horror. The presentation of each movie trailer lasted 60 seconds; each genre contained 10 trailers. Videos were displayed on a 42" LCD TV screen at 720p resolution and audio was reproduced through television stereo speakers at an easy listening volume of 70 dB. Participants sat at an approximate distance of 1m directly in front of the television and within easy reach of a computer connected mouse. Video display and user interactions were captured using a computer with two display outputs; one screen output the video and subjective response collection application interface and the other displayed the classifier interface. The presentation order of the movie trailers was randomised for each participant, with each video presentation drawing from the pool of 40 until all material was exhausted.

Procedure

After receiving instruction about the experimental procedure, participants were required to provide written consent. Electrodes were placed on the torso for ECG and on the distal phalanges of second and forth finger of the non-dominant hand for SC. Participants were asked to sit comfortably but remain as still as possible.

The experimental procedure was completed in two parts, initial training (build 1) and subsequent instances of classification/feedback. During the build 1 mode a video trailer of 60 seconds duration randomly chosen from a pool of 40 was displayed to the participant. After each video, participants were shown a simple interface on screen that asked "was this content interesting? Yes/No." Once feedback was received, another screen appeared to allow the next video in the sequence to be played ("Play next video? Yes/No"). This procedure was repeated until the experimenter received a message indicating that a classifier was being constructed.

Once a classifier had been constructed, the sequence of events was as follows:

(1) participant views movie trailer

(2) participant provides a subjective rating of whether the movie trailer was rated as high or low interest. This rating was added to the training dataset for subsequent builds of the classifier

(3) The system provides feedback (high/low interest) to the experimenter for that item based on the current build of the classification algorithm

(4) Feedback from the system is conveyed verbally to the participant by the experimenter.

This mode of interaction continued during builds 2-4.

The experimenter was the same person throughout the trial and it was made clear to the participants that he was simply a conduit to system feedback.

RESULTS

The classification accuracy of the system was measured in mathematical terms using holdout cross-validation (as described above) and the F_1 score. The holdout method of cross-validation uses the entire dataset as both training and testing data by splitting the data arbitrarily according to criteria; that is, data is randomly assigned to either training or testing according to the "set size" determined before classification (in this case 60% training, 40% testing). The dataset contains both the classification vectors (observations) and its associated label (subjective judgments), testing the SVM model involves classifying the remaining (40%) novel instances of test data, to determine accuracy. The labels (subjective judgments) associated with the test vectors (observations) are known to the experimenter but unknown to the SVM model.

The perceived accuracy of the system was expressed as the degree of agreement between the binary classification produced by the system in real-time and subjective self-assessment from the user. Perceived accuracy was expressed as a percentage figure (i.e. % of agreement) and in terms of four categories of outcome:

- True Positive (both system and participant rated the movie as high interest)
- False Positive (system rated the movie as high interest but participant produced a rating of low interest)
- True Negative (both system and participant rated the movie as low interest)
- False Negative (system rated the movie as low interest but participant produced a rating of high interest).

Research Question 1: How does the accumulation of training data influence classification accuracy over successive builds of the system?

The first question posed by the study addressed the relationship between the accumulation of training data and mathematical accuracy (and the perception by the user of mathematical accuracy). The classification engine was build through four phases. The first build was based upon a relatively small number of stimuli (movie clips) and is assumed to produce the lowest level of classification accuracy. By the same logic, it is assumed that the fourth and final build would deliver the highest level of accuracy because it is based upon the largest corpus of training data.

A univariate ANOVA was conducted to assess the influence of build on mathematical accuracy. This analysis revealed no significant influence of build on either index of mathematical accuracy [F(3,11)=1.40, p=.30]. A second ANOVA was performed to explore the influence of build on perceived accuracy. This analysis revealed a marginal main effect [F(3,11)=3.15, p=.06]; post-hoc analyses indicated that perceived accuracy was significantly higher for build 1 and 4 compared to build 2 (p=.04), in addition, perceived accuracy increased during build 4 compared to build 3 (p=.05). These finding suggest that perceived accuracy followed a quadratic pattern over the number of builds, being high at build 1, falling by 10% at build 2 and rising to original level by build 5. All descriptive statistics are illustrated in Figure 2.

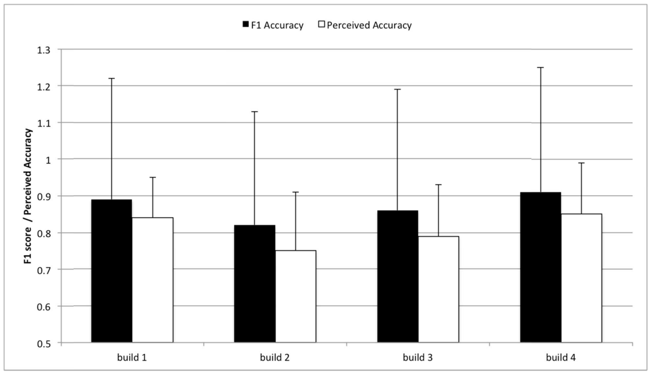

Figure 2. Means and standard deviation for mathematical accuracy (F1) and perceived accuracy over four successive system builds (N=14).

Research Question 2: How does perceived classification accuracy change over successive builds and sustained exposure to the system?

The second question concerned the effect of each build on user perceptions of accuracy with respect to pattern of responses, i.e. True Positive, False Positive, True Negative, False Negative. A 4 (build) x 2 (positive/negative) x 2 (true/false) ANOVA was conducted on these data. This analysis indicated a greater frequency of positive responses (M=3.04) compared to negative responses (M=0.95) [F(3,11) = 105.7, p<.01], i.e. participants tended to indicate high rather than low interest. In addition, there were a greater number of true (i.e. correct) responses (M=3.23) compared to false responses (M=0.75) [F(3,11) = 106.3, p<.01].

A significant interaction between positive/negative response category and true/false [F(3,11) = 11.06, p<.01] revealed a higher number of correct (M=5.12) compared to false responses (M=0.96) in positive/high interest category. A significant interaction between build and true/false [F(3,11) = 8.56, p<.01] indicated that the number of correct responses was higher during build 4 (M=4.04) compared to build 2 (M=2.75) or build 3 (M=2.93). These findings suggest a general bias towards positive (i.e. high interest) cases and true (i.e. correct) responses from the participants. It was also apparent that participants perceived the frequency of correct responses (in both positive and negative categories) to peak following the fourth and final build of the classification algorithm. The effect of build on positive responses is illustrated in Figure 3.

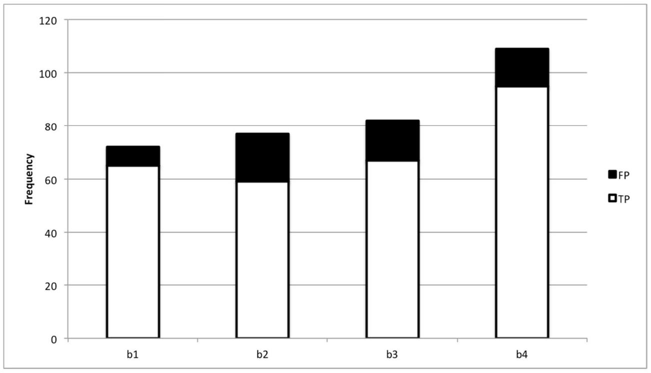

Figure 3. Frequency of True Positives (TP) and False Positives (FP) over four successive builds of the classification algorithm (N=14).

Research Question 3: How long is required to train a subject-dependent classifier?

The study was designed to train a classification algorithm to the individual on a subject-dependent basis using data from initial exposure to the system. There is a question regarding the feasibility of this approach with respect to user acceptance, i.e. can a classifier be created for all participants within a short period of time. We found that the system took an average of 7 videos (i.e. 7 minutes) to construct the SVM. The maximum number of videos was 12 but this was an upper limit imposed by the experimental protocol (see *Participants* section for further explanation). The minimum number of videos used to create the SVM was 4.

Research Question 4: What is the relationship between mathematical classification accuracy and perceived accuracy?

The final question to be explored concerns the relationship between mathematical measures of classification accuracy and users' perceptions of accuracy during interaction. There are two aspects to be investigated, the first concerns the differential between perceived accuracy and mathematical accuracy, i.e. do users tend to over- or under-estimate with respect to mathematical classification accuracy. A difference score was calculated to express the differential between perceived accuracy and mathematical accuracy [Perceived Accuracy – Mathematical Accuracy], i.e. positive score indicates that perceived accuracy was greater than mathematical accuracy. The mean difference score illustrates that estimates of perceived accuracy were generally between 4-9% lower than the F1 score but variability was substantial. A 4 (build) x 2 (index of mathematical accuracy) ANOVA was performed to explore the impact of build number on this differential. No significant differences were found with respect to influence of build number (see Table 2 for descriptive statistics).

The second aspect of this analysis relates to the degree of association between perceived classification accuracy and its mathematical analogue. There is also a secondary issue related to whether the degree of association between perceived and mathematical accuracy changes over time with each successive build, i.e. does the association between 'hard' and 'soft' markers of accuracy change with each successive build. A series of Pearson's r correlations were calculated to express the degree of association between mathematical and perceived accuracy over all four builds of the classification algorithm. The resulting r scores (Table 4) demonstrate a positive correlation for F1 that is statistically significant in the case of all but the second build phase.

Build	F1 *r*	Mean Diff.
1	0.58*	.09 [.29]
2	0.37	.04 [.30]
3	0.82**	.05 [.24]
4	0.73**	.07 [.27]

Table 2. Pearson's r correlation coefficients between mathematical classification accuracy and perceived accuracy. Mean Diff. refers to the differential score produced when mathematical accuracy was subtracted from perceived accuracy, standard deviation in parentheses. * p<.05, **p<.01 (N=14).

DISCUSSION

The purpose of this study was to explore the relationship between mathematical classification accuracy and the perception of accuracy from users during a real-time interaction with explicit feedback. In addition, the experiment was designed to explore how the size of the training dataset influenced both 'hard' and 'soft' markers of classification accuracy.

The system was designed to classify psychophysiological data in real-time on a subject-dependent basis. The classification algorithm (SVM) was initially created for each individual based upon exposure to an average of seven movie trailers of 1min duration. After the initial build of the classifier (build 1), the SVM was rebuild every 7-9 movie trailers to yield three subsequent versions of classifier, each one based upon on cumulative increase of training data.

The effect of system builds on classification accuracy

Due to high levels of bias and variance in the initial training set, it was assumed that both mathematical accuracy and perceived accuracy would be poor during build one and exhibit a linear increase as the training dataset used to train the classifier increased over successive builds. The analysis of both markers of mathematical accuracy (Figure 2) revealed no significant support for this prediction. It was notable that a high (e.g. 89%) level of mathematical accuracy was achieved during the first build based upon approximately seven minutes of data collection. This finding is encouraging with respect to designing a system upon a process of subject-dependent classification where each algorithm is created dynamically for each user.

The absence of any subsequent increase in mathematical accuracy could be a 'ceiling effect' i.e. the high level of accuracy accomplished during build one left little room for further improvement. Alternatively the 40min duration of the experiment may have been insufficient to observe the effect of exposure/data collection on classification accuracy. Extending the period of exposure and size of the training dataset beyond the current experiment represents one avenue for further research.

The perceived accuracy of the classification engine tended to fluctuate over the four builds (Figure 2). The statistical analysis support a view that perceived accuracy increased during the final build compared to builds 2 and 3. High level of perceived accuracy observed during build one could reflect a 'halo effect' due to the novelty of the system and the incorporation of technical apparatus, such as sensors. When participants progressed to the second build phase, perceived accuracy dropped by 9%, a trend that was mirrored by the F1 score (Figure 2). The decline of perceived accuracy during the second build may reflect how increased exposure to the system classification led to a less optimistic (and more accurate) appraisal of the capabilities of the technology. This type of analytic understanding of a technical system informs the development of trust between user and technology [13].

Factors influencing perception of system accuracy

It is important to note that positive responses (high interest) were more frequent than those in the negative (low interest) category. This is unsurprising given that our participants watched movie trailers, which are designed to pique the interest of the viewer. However, this imbalance did bias the participant to perceive classification accuracy in terms of the more frequent (high interest) category. If we consider the ratio of true (correct) to false (incorrect) responses in the positive category (Figure 3), participants experienced one incorrect response for every 9.3 classifications during the first build. This was the highest accuracy recorded by participants and represented an initial positive bias. This index fell to one error for every 3.3 classifications during the second build phase (Figure 3). The proportion of incorrect responses in the positive/high interest category subsequently decreased during builds 3 (one error per 4.5 classification) and 4 (one error per 6.8 classifications). Given that negative responses (low interest) were relatively infrequent, it is argued that the proportion of correct responses in the high interest category were largely responsible for driving the perceived classification accuracy.

The number of classifications judged to be correct (in both high and low interest categories) significantly increased during the fourth and final build compared to builds 2 and 3. In addition, the correlation between mathematical scores of classification accuracy and perceived accuracy were positive and significant during final two builds of the system (Table 2). It can be argued that build four represented instances of classification based upon the largest training dataset, where bias and variance are both reduced, and convergence between mathematical and perceived measures of accuracy was a natural consequence of this factor.

Alternatively, the presence of class imbalance within our classification system may have functioned as a form of implicit bias. Participants learned that output from classification tended to favour the 'high interest' category, which represented a subtle mechanism of entrainment whereby participants tended to choose the positive category without conscious realization. It is also possible that participants treated the experiment as a game where the goal was to correctly match their response with one produced by the system. Hence, participants were predisposed towards the high interest category, which in turn leads to the creation of classification engine with an implicit bias towards this category - and repeated feedback from the system both reinforces and amplifies this bias towards the high interest category. A final possibility is that participants tended towards agreement with the system classification during the fourth build due to fatigue accumulated during the test session.

Methodological issues

This issue of class bias with respect to binary classification presented a dilemma for assessment of the current system. The obvious solution is to create a balanced training set where both outcomes are equally likely but that can be problematic where bias is an inherent property within a database. Even if the classification system were initially trained using perfectly balanced data, bias would eventually creep into the classification engine when it was re-trained according to the preferences of the individual. A systematic exploration of class bias using this type of subject-dependent classification based upon an incremental training dataset is one topic for further work. The issue of bias due to feedback could be explored systematically by varying the protocol used in the current study. For example, the results of each classification could be withheld from the participant and shared at the end of the experiment.

The current study used a protocol where system feedback was conveyed to participants via the experimenter. This form of feedback was selected due to the technical limitations of the system and was far from ideal. The presentation of feedback at the interface is likely to exert a strong influence on the perception of accuracy. The use of a human agent introduces an unwelcome level of 'experimenter bias' into the experience of the participants.

It was surprising that perceived accuracy and mathematical scores of accuracy were generally within 10% of one another (Table 2). There was a general tendency for perceived accuracy to be higher than mathematical accuracy. The study does indicate that F1 score was generally a good predictor of perceived accuracy; it was significantly positively correlated with perceived accuracy in all but one of the four builds (Table 4). The results of the correlation suggest that F1 score from a classification engine may provide a reasonable estimate of perceived accuracy from users but more research is required to support this claim.

CONCLUSION

The purpose of this study was to explore the relationship between mathematical and perceived classification accuracy using psychophysiological data in a real-time application. It was found that mathematical accuracy remained stable throughout the experiment whilst perceived accuracy showed some fluctuation related to bias and developing expectations from the user. The study indicated that perceived accuracy tended to be an over-estimation of mathematical accuracy (F1 score) but there was a high degree of positive correlation between F1 and perceived accuracy.

ACKNOWLEDGEMENTS

This work was partly funded by the European Commission as part of ICT-2009.4.1 (Digital Libraries and Digital Preservation) under the ARtSENSE project (270318).

REFERENCES

1. Soleymani, M, Pantic, M & Pun, T. 2012. Multimodal emotion recognition in response to videos. IEEE Transactions on Affective Computing, 3(2), 211-223.
2. Soleymani, M. & Pantic, M. 2012. Human-centred implicit tagging: overview and perspectives. In Proceedings of the 2012 IEEE International Conference on Systems, Man and Cybernetics, 3304-3309.
3. Gilroy, S. W., Porteous, J., Charles, F., Cavazza, M., Soreq, E., Raz, G., Ikar, L., Or-Borichov, A., Ben-Arie, U., Klovatch, I. & Hendler, T. (2013). A Brain-Computer Interface to a Plan-Based Narrative. In F. Rossi (ed.), *IJCAI*, 633-2
4. Fairclough, S.H. 2009. Fundamentals of physiological computing. Interacting With Computers, 21, 133-145.
5. Fairclough, S.H. & Gilleade, K.E. 2012. Construction of the biocybernetic loop: a case study. In *Proceedings of the 14th ACM international conference on Multimodal interaction* (ICMI '12). ACM, New York, NY, USA, 571-578.
6. Koelstra, S., Mühl, C., Soleymani, M., Lee, J., Yazdani, A., Ebrahimi, T, Pun, T., Nijholt, A., and Patras, I. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. IEEE Transactions on Affective Computing, 3(1), 18-31.
7. Russell, J.A. 1980. A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161-1178.
8. van der Zwaag, M.D., Janssen, J.H & Westerink, J.H.D.M. 2013. Directing physiology and mood through music: validation of an affective music Player, IEEE Transactions on Affective Computing*,* 4(1), 57-68.
9. Karran, A.J., Fairclough, S.H. and Gilleade, K.E. 2013. Towards an adaptive cultural heritage experience using physiological computing. In CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13). ACM, New York, NY, USA, 1683-1688.
10. Karran, A.J. & Kreplin, U. 2014. The drive to explore: physiological computing in a cultural heritage context. In Fairclough, S.H. & Gilleade, K. (Eds.) Advances in Physiological Computing. Springer. 169-195.
11. Novak, D., Mihelj, M. & Munih, M. 2012. A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing. Interacting With Computers, 24, 154-172.
12. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F. & Arnaldi, B. 2007. A review of classification algorithms for EEG-based Brain-Computer Interfaces. Journal of Neural Engineering, 4, R1-R13.
13. Miller, C. A. 2005. Trust in adaptive automation: The role of etiquette in tuning trust via analogic and affective methods. Paper presented at the First International Conference on Augmented Cognition, Las Vegas, NV.
14. Van de Laar, B, Bos Plass-Oude, D., Reuderink, B., Poels, M. & Nijholt, A. 2013. How much control is enough? Influence of unreliable input on user experience. IEEE Transactions on Cybernetics, 43(6), 1584-1592.
15. Novak, D., Nagle, A. & Riener, R. 2014. Linking recognition accuracy and user experience in an affective feedback loop. IEEE Transactions on Affective Computing, 5(2), 168-172.
16. Ramnani, N. & Owen, A.M. 2004. Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. Nature Reviews Neuroscience, 5, 184-194.
17. Coan J.A., Allen J.J. 2004. Frontal EEG asymmetry as a moderator and mediator of emotion. Biological Psychology. 67(1) 7-49.
18. Platt J.C. 1999. Fast training of support vector machines using sequential minimal optimization. In Advances in kernel methods, Bernhard Schlkopf, Christopher J. C. Burges, and Alexander J. Smola (Eds.). MIT Press, Cambridge, MA, USA 185-208
19. Isaksson, A., Wallman, M., Göransson, H. & Gustafsson, M.G. 2008. "Cross-validation and bootstrapping are unreliable in small sample classification". Pattern Recognition Letters, 29(14), 1960-1965
20. Kelley, J. F., "An iterative design methodology for user-friendly natural language office information applications". ACM Transactions on Office Information Systems, March 1984, 2:1, pp. 26–41.