International Journal of Human-Computer Studies 122 (2019) 232–241

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Observers' physiological measures in response to videos can be used to detect genuine smiles

Md Zakir Hossain*, Tom Gedeon

Computer Science and Information Technology, 108 North Road, Acton 2601, Australia

ARTICLE INFO

Keywords:
Classification accuracy
Smiles
Observers
Physiological measures
Video Stimuli

ABSTRACT

We investigated a method to detect genuine smiles from observers' physiological states. We recorded two physiological measures from people observing videos of smiles: pupillary response (PR) and galvanic skin response (GSR). Smile videos were from two benchmark databases (MAHNOB and AFEW). MAHNOB videos were classified as showing genuine or real smiles and AFEW videos were classified as not showing real smiles, based on their process of elicitation. A leave-one-observer-out procedure was employed to investigate classification performance using k-nearest neighbor (KNN), support vector machine (SVM), simple neural network (NN), and ensemble classifiers. Different noise removal techniques and a feature selection method — canonical correlation analysis with neural network (NCCA) — were applied to find minimally correlated features for the classes. Using these methods, the highest classification accuracy of 97.8% for PR and 96.6% for GSR signals were found via the ensemble classifier. In comparison, the observers (n = 20) correctly judged smiles as real only 58.9% of the time (on average) to 68.4% (by voting), which is similar to the literature, showing our data is similar in quality. Overall, our results demonstrate that user-independent analyses of physiological measures can substantially outperform individual self-reports for detecting real smiles.

1. Introduction

It could be highly beneficial to discriminate genuine facial expressions (spontaneous/felt/real) from other types (posed/acted) robustly and reproducibly in many situations like social interaction, public security, and so on. As an example, a police officer or computerized tool may make assumptions about a suspect's veracity or not according to whether their facial expressions are genuine or acted. One of the most frequently displayed facial expressions is a smile (Dibeklioglu et al., 2015). Smiles are interpersonal 'tools' and nonverbal behaviours that are sometimes more significant than spoken words (Birdwhistell, 1970) and carry extra information to strengthen, supplement or contradict what is being said. Shlenker (1980) indicated that smiling can help people to increase likability. Gifford et al. (1985) show evidence that people who smile more, and use more gestures, are identified as having better social skills. Smiling with eye contact is also perceived to have a positive influence on how people respond to a question (Parsons and Liden, 1984). Thus, the smile is an extremely useful facial expression, and accurately recognizing whether a smile is genuine or acted would seem beneficial for successful social interaction. The smile is, however, a complex, multi-purpose, dynamic expression that conveys not only the meaning of happiness, but can also be identified as rapport, sarcasm, frustration, empathy, surprise, polite disagreement, pain and even more (Hoque et al., 2011). In this paper, we refer to smiles which are elicited by stimuli to generate positive smiles involuntarily as real or genuine smiles interchangeably. Voluntary smiles can include posed and acted smiles as well as other kinds of smiles. It is important to accurately detect genuine smiles to understand the affective state underlying the meaning of this most real kind of smile.

Previous work examining whether genuine smiles can be discriminated from fake smiles has focused on analysing the smile images/videos directly. For example, Valstar et al. (2007) tested discrimination of genuine from fake smiles by analysing 202 videos of smiling people. The dynamic and morphological characteristics of the smiles of virtual agents were studied in (Ochs et al., 2010). Ambadar et al. (2009) included co-activation of Orbicularis oculi, smile controls, mouth opening, amplitude, and asymmetry of amplitude as morphological features and duration of smiles, onset and offset velocity, asymmetry of velocity, and head movements as dynamic characteristics. Further, facial feature analysis has been evaluated using image data to measure the timing of face motion during smiles in (Cohn and Schmidt, 2004), this study showed that dynamic characteristics were more informative than morphological characteristics. Ambadar et al. (2009) showed specific physical characteristics of smiles (e.g., smile controls, mouth

* Corresponding author.
E-mail address: zakir.hossain@anu.edu.au (M.Z. Hossain).

https://doi.org/10.1016/j.ijhcs.2018.10.003
Received 22 November 2017; Received in revised form 22 October 2018; Accepted 23 October 2018
Available online 24 October 2018
1071-5819/ © 2018 Elsevier Ltd. All rights reserved.

opening, amplitude, asymmetry of amplitude, asymmetry of velocity, duration, head movements etc.) influenced what those smiles were perceived to mean. The morphological and dynamic features of smiles in the case of face-to-face interaction were also studied in (Hoque et al., 2011). The facial and prosodic features of displayers' video clips were analysed in (Hoque and Picard, 2011), with the purpose of recognizing smiles from both acted and naturally elicited data. A two-layer deep Boltzmann machine was applied to smiling image data in (Gan et al., 2015). An informative feature set was extracted from smiling faces and an automatic technique was implemented for analysing smiling videos to discriminate genuine from posed smiles in (Dibeklioglu et al., 2015), with 92.90% accuracy. Although all of the above studies were found to reliably discriminate real from fake smiles, they focused on analysing the video/image data directly and did not measure physiological signals from observers watching the smiling video stimuli, and also they did not examine human physiological reactions and self-judgements in a single experiment.

As mentioned above, smiles generally reflect positive affect (Ekman et al., 1990), but can arise from a variety of emotions (Ekman and Friesen, 1982). Affect detection from a displayer's physiological measures is an ongoing research topic (Zhou et al., 2011; Liu et al., 2008). In general, recognition from video is easier for users than recognition from static images (Picard, 2000). Observers may experience certain feelings (Kim and Andre, 2008; Soleymani et al., 2009) from watching video clips or listening to music that are related to their physiological state. Observers' physiological changes are also associated with emotional states (Kim and Andre, 2008) and less susceptible to social masking (Kim, 2007). The physiological reactions in the body and brain change in response to different stimuli and are recognised as self-judgements (Damasio, 1994). On the other hand, affective self-reports might be held in doubt because errors in self-judgements are not negligible – observers might misrepresent or cannot always remember different smiles during an experiment or might want to please the experimenter (Soleymani et al., 2012b). The approach of using observers' physiological signals to decode affective responses (Soleymani et al., 2009) to smiles is an alternative way of accessing the displayers' internal state. In this regard, two physiological signals — pupillary response (PR) and galvanic skin response (GSR) — were analysed in an attempt to detect genuine smiles.

The pupillary response is the measure of pupil diameter over time. Among other things, pupil diameter is influenced by light, cognition, attention, and emotion (Bradley et al., 2008; Partala and Surakka, 2003). The pupillary reflex has been found to vary significantly during the identification of smiles or emotions after removing luminance effects (Soleymani et al., 2012b). Principal Component Analysis (PCA) can be used to reliably separate the effect of changes in luminance from other effects (Oliveira et al., 2009). However, pupillary responses also change with different emotional states (Bradley et al., 2008; Partala and Surakka, 2003). GSR is another important physiological signal that has been found to be sensitive to emotional changes (Kim and Andre, 2008; Healey and Picard, 1998). Recent research has indicated that reactions to happiness and sadness can be distinguished from GSR (Levenson et al., 1990). We recorded and analysed both of these signal from observers while they watched brief video stimuli showing the key emotional expression.

We collected video clips from benchmark datasets to use as stimuli (Fig. 1) for the observers. We use the expression "displayer" to indicate the person in the video performing an emotion, such as a genuine smile, whereas the "observer" is the person watching the video. Firstly, nineteen video clips were collected from two benchmark datasets ((Dhall et al., 2014; Soleymani et al., 2012), and (Petridisa et al., 2013)) and processed using MATLAB R2015a to convert them to grey scale, with each clip lasting 10 s. The use of greyscale and 10 s was to eliminate differences between data sources (see Stimuli Collection section). Secondly, physiological signals were recorded from twenty observers, while watching the video stimuli, along with their judgments as to whether the smiles were genuine, collected via a Likert scale. Thirdly, the data from the observers of the nineteen stimuli we collected were analysed. The analysis consisted of several stages: signal normalisation, de-noising, smoothing, feature selection and classification. For classification, K-nearest neighbor (KNN), support vector machine (SVM), neural network (NN), and ensemble classifiers were utilised to discriminate between genuine and acted smiles by analysing the physiological signals separately. Canonical correlation analysis with neural network (NCCA) was applied to select minimally correlated features. Finally, the results showed that classification accuracies were much higher than chance and significantly higher than the same observers' own judgements.

2. Methodology

2.1. Stimuli collection

The original video stimuli were randomly selected from two databases: AFEW (Acted Facial Expressions in the Wild) (Dhall et al., 2014) and MAHNOB (Multimodal Analysis of Human Nonverbal Behaviour in Real-World Settings) (Soleymani et al., 2012a; Petridisa et al., 2013). The AFEW database contains data from professional actors displaying various emotions. In the case of smiles, they were asked to perform or instructed to display a smile and thus we could classify these as acted smiles, and here we include these as not being genuine smiles. We chose 10 stimuli from the larger AFEW database at random. For comparison, 9 real smiles' stimuli were collected from MAHNOB database (5 from the HCI-tagging database (Soleymani et al., 2012a) and 4 from the Laughter database (Petridisa et al., 2013)) where participants' smiles were elicited by watching a sequence of funny or otherwise pleasant video clips and thus classified as real/genuine smiles. Again, this was a random subset of the smiles available. The characteristics and file names of the smiling video samples are listed in Table 1.

The collected video samples were not in the same format, colour or duration, so we made them as similar as possible. The MATLAB R2015a platform was employed to convert them to mp4 format, grey scale and duration of 10 s each. Firstly, the emotion expression portion of each sample was cropped and the frames were extracted. Secondly, every frame of each sample was converted into grey scale and only the face portions were retained. Finally, the processed frames of each sample were converted back into video samples and used as stimuli to record physiological signals and self-report from the observers. In this case, either frame rates were decreased or processed frames were used in a repetitive manner to make each stimulus up to 10 s long. This means that the observers saw the shorter facial expressions for longer, so the time for viewing overall was the same for all videos. We also checked the luminance variation over all frames of each stimulus using MATLAB SHINE toolbox (Willenbockel et al., 2010), the results are shown in Fig. 2. It can be seen from Fig. 2 that luminance does not vary much among the frames of each particular stimulus and the average luminance is in a range of 60 to 85 ALU (arbitrary linear unit).

2.2. Experimental setup

In order to detect genuine smiles, a structured experiment was implemented to record and evaluate the observers' judgements and physiological signals. The selected 19 stimuli were presented to observers after a short introduction page. At the end of each stimulus, each observer was required to select an option based on a 5 point Likert scale (-2 to $+2$) with an additional option of 'No Smile' as shown in Fig. 3, to indicate whether they judged the facial expression to be a smile, and the degree to which it was showing a real smile expression or not. The total duration of the experiment was around 10 minutes.

In this study, the observers' GSR signals were recorded using Neulog (https://neulog.com/) sensors at a sampling rate of 10 Hz from the index and middle finger of the left hand. Eye activities (pupil dilation)

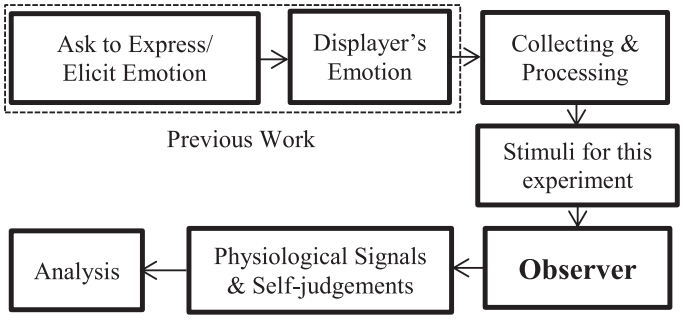

Fig. 1. Basic block diagram of the experiment.

Table 1
Collected video samples shown as stimuli.

Sl.	Source	File Name	Category	Notation
1	MAHNOB_HCI (Soleymani et al., 2012a)	P2-Rec1-2009.avi	Genuine / Real / Felt / Spontaneous smiles	R1
2		P4-Rec1-2009.avi		R2
3		P8-Rec1-2009.avi		R3
4		P14-Rec1-2009.avi		R4
5		P24-Rec1-2009.avi		R5
6	MAHNOB_Laughter (Petridisa et al., 2013)	S001-001.mp4		R6
7		S008-002.mp4		R7
8		S009-001.mp4		R8
9		S011-001.mp4		R9
10	AFEW (Dhall et al., 2014)	000,329,320.avi	Do not belong to the above category	A1
11		000,404,000.avi		A2
12		002,809,954.avi		A3
13		011,309,840.avi		A4
14		013,818,854.avi		A5
15		000,758,680.avi		A6
16		004,025,454.avi		A7
17		005,513,240.avi		A8
18		001,912,000.avi		A9
19		003,652,360.avi		A10

were recorded using a Facelab (Seeing Machines) remote eye-tracker system with a sampling rate of 60 Hz. The recording system is shown in Fig. 4.

2.3. Data acquisition

Twenty (9 female, 11 male) students, with mean age of 26.9 ± 6.3, voluntarily participated as observers of the videos in this experiment, from the Australian National University. Each observer had normal or corrected to normal vision. We recruited voluntary participants, as they provide highly reliable outcomes compared to paid participants, when they complete experiment tasks (Redi and Povoa, 2014). Our participants voluntarily took part in our experiment, and they all finished the task. All the methods related to the experiment were approved by our University's Human Research Ethics Committee prior to data acquisition.

Upon arrival at the laboratory, each observer signed the consent form and was seated on a static chair, facing a 17 inch LCD monitor in a sound-attenuated, dimly lit, closed room. Sensors were attached to measure their GSR signals. Observers were given a brief introduction to the experimental procedure. Their chairs were moved forward or backwards to adjust the distance between the chair and eye tracker. Nine point calibration was performed, where a spot was displayed on the monitor and observers asked to track it, for calibrating the eye tracker and starting the experiment. Observers were instructed to limit their body movements in order to reduce undesired artefacts in the signals. During the experiment, all observers used their right hand for moving the mouse or typing. The stimuli were presented to the observers in an order-balanced way. After completing the experiment, the sensors were removed and the observers were thanked for their participation.

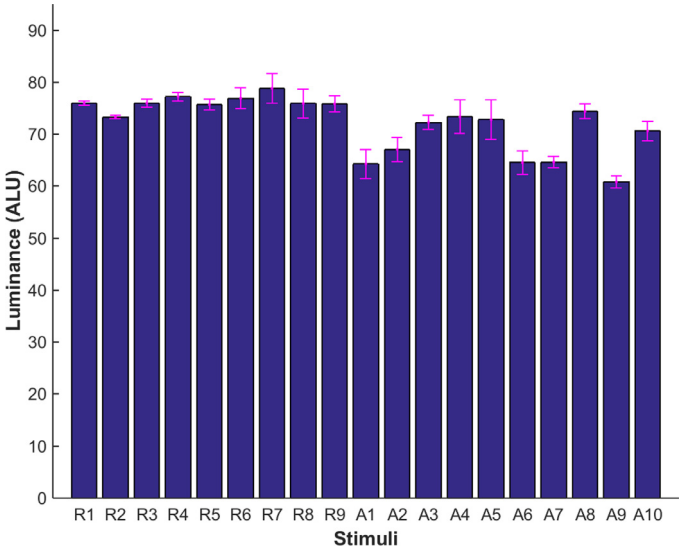

Fig. 2. Average (± std.) luminance (in arbitrary linear units (ALU)) over all frames of each stimulus.

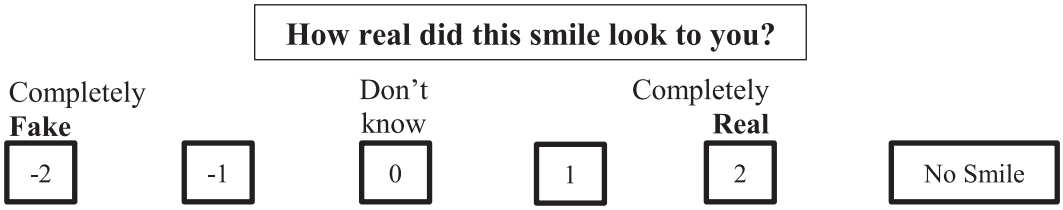

Fig. 3. Five point Likert scale to accumulate participated observer's self-report.

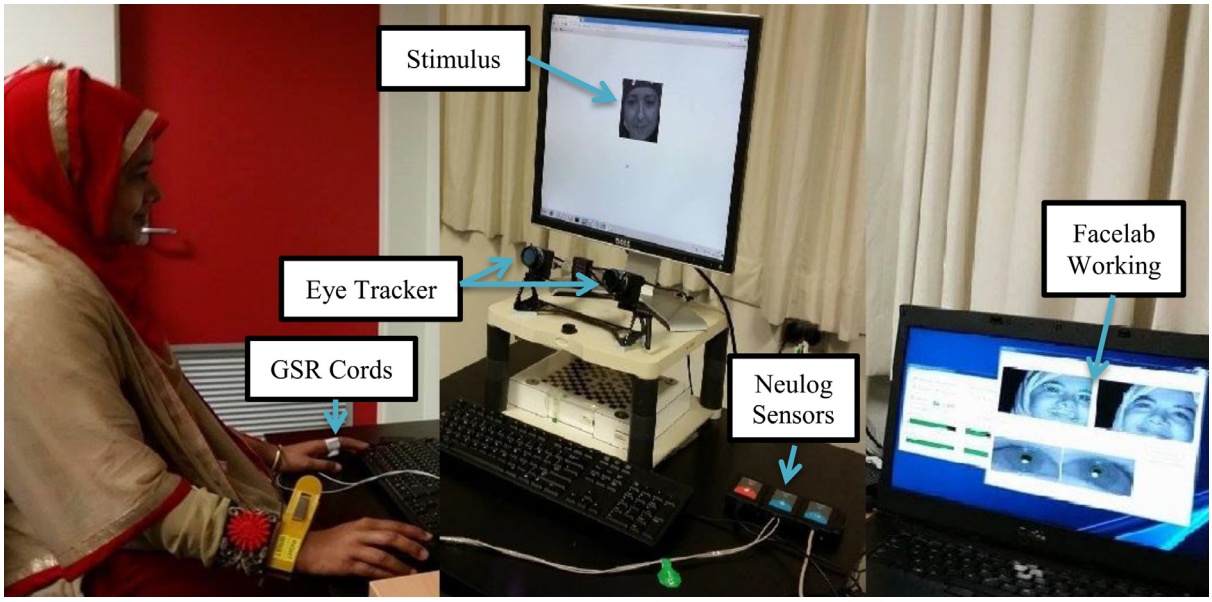

Fig. 4. Experimental setup to record data.

Fig. 5. Pupil signal processing procedure.

2.4. Data processing

The recorded physiological signals were extracted and three data sets were created: left eye pupil diameter (LEPD), right eye pupil diameter (REPD), and GSR. All the extracted features were numerical. It was necessary to standardize the features to reduce the between-observer differences (Hossain et al., 2016a). Maximum value normalization was applied to each data set separately. In this normalization, the maximum value from a given signal of each observer was computed over all videos watched and all features of that particular signal were divided by their computed maximum value. Thus, all data for each observer varied between 0 and 1 for each video. Data for each stimulus were then separated and the final data sets were constructed. Thus, every data set had 19 patterns (consisting of physiological signal sensor measurements) corresponding to the 19 stimuli (that is, for the emotion videos) with a number of features (here we treated each time point of a signal as a feature) over 10 s duration for each pattern.

In the case of pupil data, the Seeing Machines eye tracker provides the position of the projected eye gaze on the monitor, the pupil diameter and the moments when the eyes are closed or blinking. The missing data segments due to eye blinks were measured as zero by the eye tracking machine, and cubic spline interpolation was applied to reconstruct the pupil size during the blinking time (Mathôt, 2013). Then, the interpolated signal was smoothed using 10-point Hann moving window average, to filter out noise and unrelated features (Zheng et al., 2014). According to Pamplona et al. (2009), pupil diameter varies due to effects caused by lighting, and the pupillary light reflex magnitude changes between different people. The magnitudes of pupil diameter time series were normalized according to the maximum value normalization technique. Principal component analysis (PCA) had been shown to be effective in separating the effect of changes in luminance from stimulus relevance (Oliveira et al., 2009). This was performed here by subtracting the first principal component from the normalized and smoothed pupil diameter data (Soleymani et al., 2012b). The pupil signal processing steps are shown in Fig. 5.

Trends of LEPD and REPD were observed after removing noise. Similar types of trends were found when comparing the left eye (Fig. 6(a)) and right eye pupil diameters (Fig. 6(b)). In the case of GSR signals, twenty point median filter was applied on normalized GSR signals to smooth and remove the effect of noise from the raw signals as suggested by Guo et al. (2013). The trends of GSR signals are depicted in Fig. 6(c).

Fig. 6 illustrates the time point average of physiological signals over observers when viewing all video stimuli. In the case of pupil dilation, it can be seen that the pupil constricted from stimulus onset and reached a minimum, and then a sharp dilation started and continued until reaching a maximum point. Then, either a smooth dilation or constriction started and continued, which is sustained in a consistent range, until the very end of our analysis window. It is stated in the literature that brain signals (such as EEG signals) spend 0.2–0.5 s to detect emotional stimuli (Lithari et al., 2010; Bilalpur et al., 2017) where peripheral physiological signals take 2–3 s (Partala and Surakka, 2003;

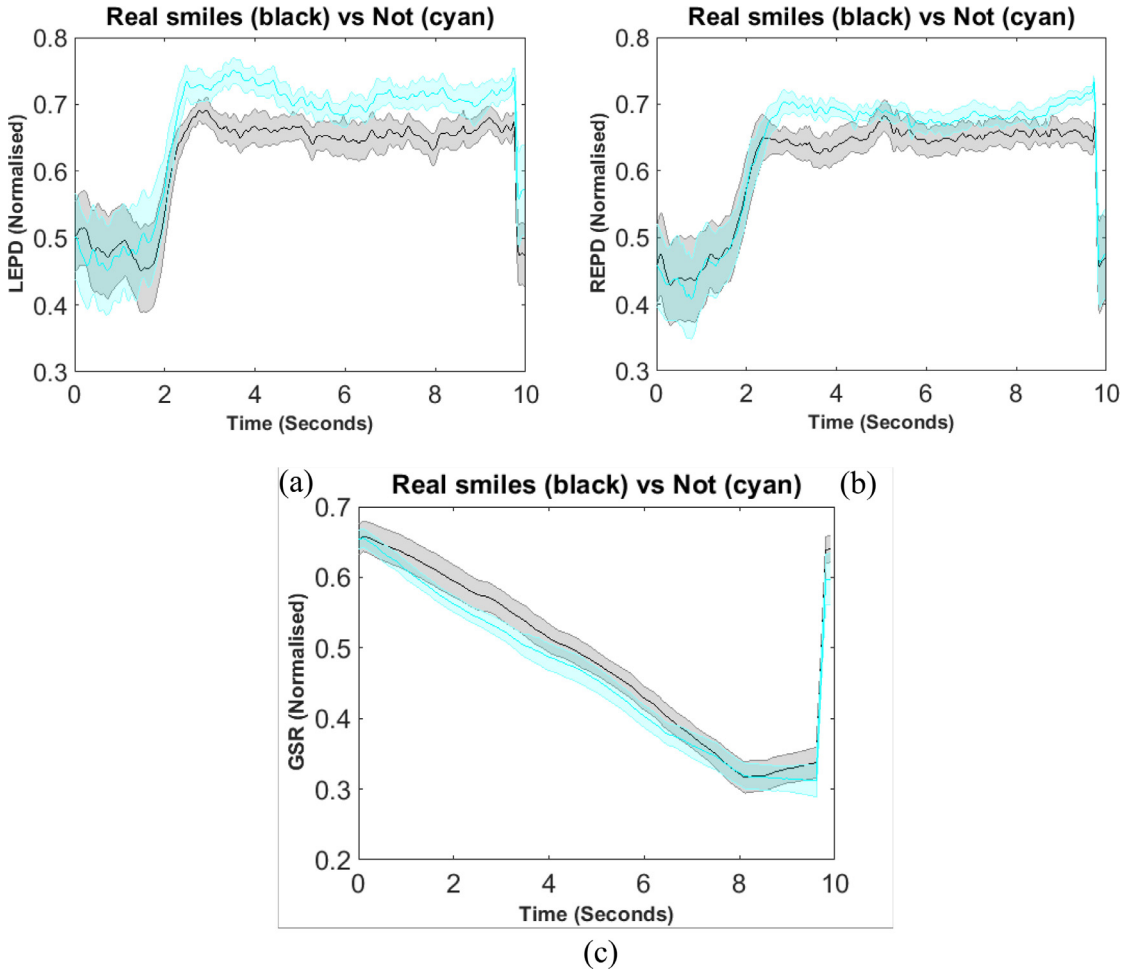

Fig. 6. Average trends of (a) LEPD, (b) REPD, and (c) GSR over observers.

Hossain et al., 2016b). It was mentioned in (Partala and Surakka, 2003) that the neutral and emotional stimuli were separated at about 1 s where peaks for all stimuli were reached at about 2–3 s. Finally, negative and positive stimuli were separated from the peak amplitudes. In our case, we only used (positive) emotional stimuli such as real and other smiles where the curves up to about 2.5 s are quite similar for real smiles versus the rest, but differ strongly from 2.5 to 4.5 s.

The trends for GSR signals are different to pupillary responses, but quite similar in timing when considering the real smiles' stimuli, and show the most divergence between real and not real smiles between 2 and 4.5 s. Two-Sample Kolmogorov-Smirnov (K-S) test (Marsaglia et al., 2003) shows that the average GSR signal ($p = 0.3435$) is not significantly different (this is not significant, but the pattern of variations between real and fake smiles have differences – between 2 and 6 s – which is selected by our feature selection method) while LEPD ($p < 0.001$) and REPD ($p < 0.001$) signals differed significantly for real smile physiological signals as compared to the other smile signals.

2.5. Feature selection

Feature selection is an important technique that reduces large numbers of features by discarding unreliable, redundant and noisy features, with the aim of achieving comparable or even better performance Huang et al., 2007). Thus, we employed a feature selection technique to find informative features relevant to our aim from these signals. In this case, a correlation technique is applied to find informative features that are relevant to the classification task. Canonical Correlation Analysis with Neural Network (NCCA) (Hossain et al., 2016c) is a training and learning process that searches for informative features according to the classification classes. This feature selection technique is applied here to search for minimally correlated features considering the real versus acted classes. There are many highly correlated features in physiological signals while watching stimuli by the same observers. We believe that the best source of differentiating information is in the minimally correlated features, and that there will be meaningful information in those features because observers were watching different types of videos, only some being real smiles. The following joint learning rules (Eqs. (1)–((3)) were considered, where $i, j, w, s, f, \lambda, \eta\&\eta_0$ represent the pattern index, feature index, weight, input features, output features, Lagrange multipliers and constant learning rates respectively.

$$f_i = w_i s_i = \sum_j w_{ij} s_{ij} \tag{1}$$

$$\Delta w_{ij} = \eta s_{ij} (f_{i+1} - \lambda_i f_i) \tag{2}$$

$$\Delta \lambda_i = \eta_0 (1 - f_i^2) \tag{3}$$

The feature selection process is explored in Fig. 7, the input features s_{ij} are initially all the features across all participants and videos at the sampling frequency of each sensor. We created two groups: Group 1 (signals while watching real smiles' stimuli) and Group 2 (signals while watching the other stimuli). The NCCA updates the values of weights (w_{ij}) by minimising the correlation between two sets of variables, such as groups 1 and 2 (s_{ij}) Lai and Fyfe, 1999). The activation is fed forward from input features (s_{ij}) to the corresponding output (f_i) through the respective weights (w_{ij}). Here, $i = 9*20$ for Group 1 and

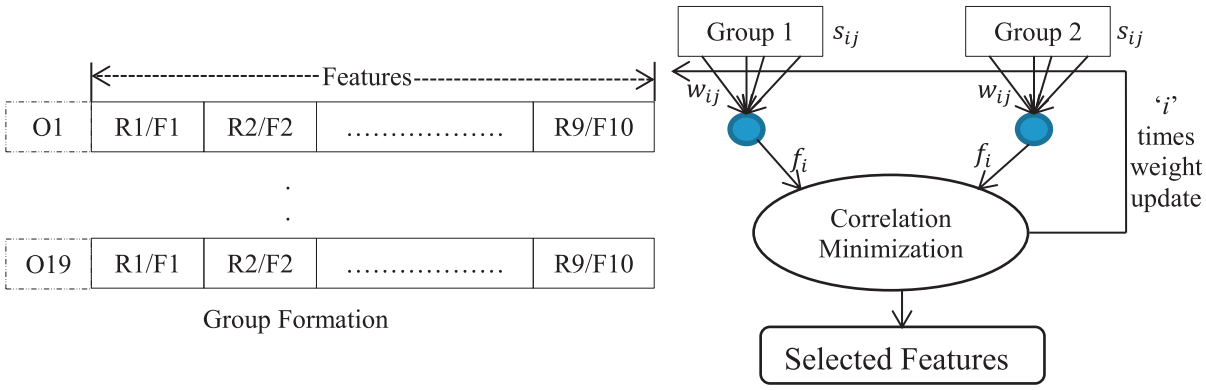

Fig. 7. Feature selection using NCCA system.

10*20 for Group 2 where there are 20 observers who watched 9 real smiles and 10 other smiles. The values of *j* are 100 and 600 for GSR and PR signals respectively where there are 100 and 600 features for a single GSR and PR signal respectively, as we considered each time point of physiological signals as a feature and the GSR sensor records at 10 Hz and the eye gaze detector (PR) at 60 Hz. The values of the weights are strengthened by updating the values of Lagrange multipliers (Lai and Fyfe, 1999). In this case, initially $\lambda = 0.015$, $\eta = 0.01$, *and* $\eta_0 = 0.5$ were chosen based on common values from the literature, and then weights and Lagrange multipliers were updated according to Eq. (2) and Eq. (3). Then minimally correlated features are selected from groups 1 and 2 according to Eq. (1)–((3) (Hossain et al., 2016c). Suppose we want to select 50 features from a total of 100 GSR features, and NCCA ranked the features according to the correlation between Group 1 and Group 2. Then the 50 features which are ranked with minimum correlation compared to other features are selected and considered for classifiers. In this case, the features of one observer are taken as the test set and the rest of the observers' features are used to train the classifier, including only using the NCCA feature selection step on the training set to avoid biasing the classifier by the effect of the test set. This process is repeated for each observer, and the average classification accuracy is reported.

3. Results and discussion

We employed four classifiers, each with two classes (real smiles and not) to compute classification accuracies from the average classification results of 20 observers. The classifiers were k-nearest neighbour (KNN), support vector machine (SVM), neural network (NN), and a voting ensemble. We used default performance parameter settings in this MATLAB version as the Euclidean distance metric and 7 nearest neighbours for KNN, sequential minimal optimization method and Gaussian radial basis kernel function with a scaling factor of 5 for SVM, scaled conjugate gradient training function with 10 hidden nodes for NN, and employed an ensemble aggregating the decision of these three classifiers respectively. The mean square error performance function is used to compute classification accuracies from each classifier. The analysis was performed with an Intel® Core™ i5-5200 U with 2.20 GHz, 8.00 GB RAM, Operating System 64-bit laptop using MATLAB R2015a. The features (at each time point) of one observer were only used for testing, while features from some or all of the other observers only were used to train the classifier. The NCCA system was applied only on the training set to select features, in order to avoid any effect of biasing on the test set. This process was repeated for each observer and thus, a leave-one-observer-out procedure was performed to compute the classification accuracies. The average classification accuracies over observers for GSR signals are explored in Table 2.

As we considered each time point as a feature for each physiological signal, thus there were a total of 100 features for each GSR signal;

Table 2
Average classification accuracies over observers for GSR features.

No. of Features	KNN		SVM		NN		Ensemble	
	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.
10	78.1	10.7	74.0	13.8	75.6	18.9	78.8	12.1
20	81.9	07.1	75.8	14.0	79.3	21.9	83.8	05.2
30	79.6	11.2	79.9	14.4	75.1	21.3	83.8	09.5
40	80.5	09.8	83.0	15.8	85.0	18.4	86.9	08.6
50	80.1	10.3	87.5	13.2	90.8	12.0	91.6	05.9
60	85.1	06.1	93.0	06.6	94.1	07.3	94.3	04.5
70	83.3	09.1	94.1	07.6	94.3	09.8	94.3	05.3
80	86.5	05.3	94.0	08.3	93.0	11.9	94.0	05.3
90	85.6	05.3	95.6	05.6	95.7	02.4	95.9	03.9
100	85.1	06.9	96.1	05.2	96.5	03.3	**96.6**	**03.3**

10 sec (video length) x 10 samples per sec (sampling frequency). It is clear from Table 2 that the ensemble classifier shows higher accuracies compared to other classifiers, with the highest accuracy of 96.6% (± 3.3) for 100 (all) features. It is worth mentioning that the pattern of differences on GSR signals could allow this discrimination, even though the *overall average* values of the GSR signals do not differ significantly as we showed in Section 2.4. In the case of each PR signal, there were 600 features (10 sec (video length) x 60 samples per sec (sampling frequency). The average classification accuracies over observers for PR features are shown in Table 3.

It can be seen from Table 3 that the ensemble classifier shows higher accuracies for PR features also. The LEPD shows the highest accuracy of 97.8% (± 0.6) for 450 selected features where the REPD shows the highest accuracy of 97.3% (± 0.8) for 550 selected features respectively. It can also be seen from Tables 2 and 3 that the classification accuracies do not change much for some ranges of selected features, such as accuracies varied only from 95.9% to 96.6% for selected features of 90 to 100 in case of GSR, and from 97.5% to 97.8% for selected features of 350–600 in case of LEPD, and from 96.4% to 97.3% for selected features of 400 to 600 in case of REPD, respectively. This indicates that we can represent whole signals by a smaller number of features, without much decrease of accuracy, whenever required.

We also checked the effect of varying the number of observers used in training, on classification accuracies. As the number of training observers increase, classification accuracy increases from lower number of observers as shown in Fig. 8. It is also noticeable from Fig. 8 that the accuracy does not increase much after a certain point, here 9 observers is that point. This outcome is reported from 450 selected LEPD features and the ensemble classifier.

We also verified the number of training videos as illustrated in Fig. 9. It can be seen from Fig. 9 that accuracies increase when the number of training videos are increases from 1 to 9. After that point, the classifiers' accuracies do not change much. These results are reported from the 450 LEPD features and the ensemble classifier.

Table 3
Average classification accuracies over observers for PR features.

No. of Features	LEPD								REPD							
	KNN		SVM		NN		Ensemble		KNN		SVM		NN		Ensemble	
	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.
50	74.3	08.0	79.8	08.2	77.4	19.1	80.5	07.1	80.3	07.9	81.9	08.5	82.9	11.2	83.0	08.0
100	79.5	07.4	81.3	06.3	82.1	18.4	84.9	06.5	82.5	04.4	88.9	05.1	87.4	02.5	89.1	05.1
150	84.8	05.3	94.3	02.6	94.3	04.1	94.9	02.4	88.0	03.5	94.3	01.4	93.8	01.4	94.4	01.4
200	88.3	04.4	96.6	02.0	96.4	01.5	96.9	02.0	89.0	03.8	95.8	01.2	95.4	00.9	95.9	01.3
250	88.6	04.1	95.1	01.6	95.9	06.2	96.5	02.1	89.3	02.7	95.4	00.9	95.6	01.4	95.9	01.2
300	88.8	04.8	95.8	01.4	96.6	01.0	97.0	01.1	91.3	02.6	95.3	00.8	95.5	01.0	95.8	01.2
350	88.3	03.6	97.1	01.1	97.1	06.0	97.5	01.1	92.3	01.8	95.9	01.2	95.1	00.6	96.0	01.3
400	90.0	02.9	96.8	00.8	95.6	01.5	97.4	01.1	92.6	01.5	96.8	01.2	95.5	01.0	96.9	01.1
450	91.6	02.6	96.3	01.2	96.4	00.6	**97.8**	**00.6**	92.6	01.5	96.6	01.2	95.3	00.8	96.8	01.2
500	91.5	03.6	97.4	01.2	97.4	00.6	97.6	01.0	92.9	01.5	96.3	01.3	95.4	01.2	96.4	01.3
550	91.9	01.1	97.4	00.6	97.3	00.8	97.6	00.6	92.9	01.2	97.1	00.8	95.6	01.6	**97.3**	**00.8**
600	92.1	00.9	97.3	01.2	97.4	00.6	97.5	01.2	93.1	01.4	97.1	00.8	95.3	00.8	**97.3**	**00.8**

Observers' self-reported judgements were also calculated from five point Likert scale results. To compute the self-report scores, the percentage of stimuli correctly selected by individual observers was calculated. The average score over 20 observers was 52.7% (on average). On the other hand, Frank et al. (1993) found that observers were 56.0% correct at discriminating genuine from fake smiles in his experiment. We also tried to find whether accuracies improve if we consider the "No Smile" choice as not being smiles in that they were not real smiles (some comments by participants made it clear that they only chose real when they were sure it was smile), leading to an accuracy of 58.9%. We also considered voting (more than 50% of the observers characterise a stimulus as real or not) from all observers. This does improve the results, the observers' judgments as a group is 68.4% accurate. Below, and in our conclusion, we suggest why this low outcome makes sense as compared to the high results for physiological signals from the same subjects. We next discuss evidence from the literature which supports both the low and high results we have achieved.

We also tried to compare our outcomes with others. But most of the work in the literature on smile genuineness has been performed either by surveys and hence just verbal responses, or just as computer vision approaches for analysing image/video based smile characteristics and hence have no verbal responses. The survey results are similar to ours, ranging from 56% to 69% (Hoque and Picard, 2011; Frank et al., 1993).

On the other hand, this paper is based on experimental observers' physiological responses while watching emotion expressions in videos. This difference makes it difficult to compare the computer vision based analyses of image/video results in the literature directly with our outcomes. Valstar et al. (2007) discriminated fake from genuine smiles by analysing displayer's video data from face, head and body actions, and found classification accuracy of 94%. With the use of morphological characteristics with the ratio of duration to amplitude, then a linear discriminant function distinguished between displayer's genuine and fake smiles with the classification accuracy of 93% (Cohn and Schmidt, 2004). Gan et al. (2015) reported their highest classification accuracy was 91.7% at discriminating displayer's genuine from fake smiles using a deep Boltzmann machine. Dibeklioglu et al. (2015) proposed an automatic system and described an informative set of facial features of smile videos to distinguish displayer's genuine from fake smiles with their highest classification rate of 92.9%. On the other hand, we have found 96.6%, 97.3%, and 97.8% accuracies by analysing observers' GSR, right eye, and left eye pupillary responses, respectively. Numerically, our results are similar and slightly better on the randomly chosen subsets from the databases. It is important to note that computer vision techniques can use a huge number of facial expressions to recognise smiles where they are using smiling faces as a primary source (Valstar et al., 2007; Dibeklioglu et al., 2015). In our case, we use observers' physiological signals while watching smile videos, thus this is impossibly long in a single experiment to consider all of the videos with humans due to boredom. We have shown by statistical significance calculations that our results are unlikely to be due to chance, and have also shown that we have sufficient observers (as Fig. 8 shows) and sufficient videos (as Fig. 9 shows). This result makes intuitive sense: the

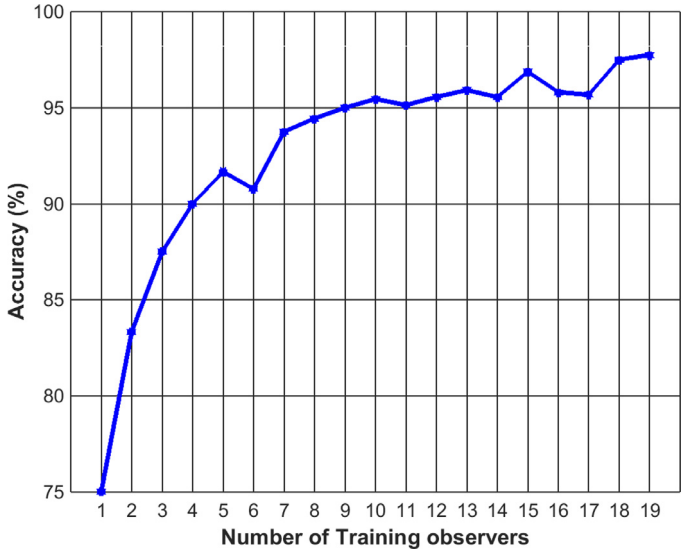

Fig. 8. Variation of accuracies with the increasing number of training observer.

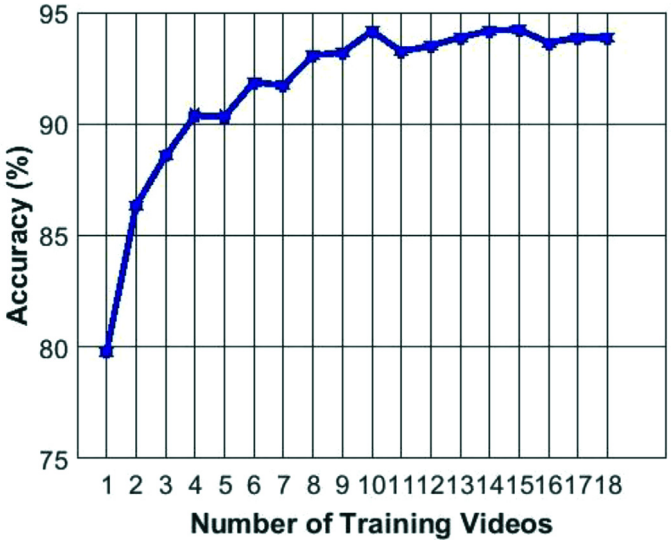

Fig. 9. Variation of accuracies with the increasing number of training videos.

computational approaches try to use all the information available in the image / video, which is the same information available to the human observer. A slightly better performance by the human could be due to the greater amount of training by the human observer from their life prior to the experiment. It is also difficult to compare our results with the emotion recognition work such as Hoque and Picard (2011), because they use different stimuli and emotion classes. We note that those emotion classes are not directly related to detecting genuine smiles, as their fake smiles can include more obvious fake smiles, as we can deduce from their definitions.

We believe that a computer vision algorithm *should* be able to perform the same task with the same level of accuracy as our human observers. The nearest result for a similar task in the literature is 4% less accurate (Valstar et al., 2007). Such computer vision algorithms require significant amounts of detailed knowledge and hard work to encode the characteristics of real smiles (and this is not as yet fully understood in the Psychology literature), or require large numbers of real smiles for training a deep learning neural network or similar non-parametric classifier to improve their results. There are many smile images on the web, but which of them are genuine? There are only small numbers of smiles in databases where we can know with confidence that they are real as they were elicited as such. Arguably, even expecting a smile to be elicited could lead to a subject trying to help the experimenter and smiling partly consciously or at least self-consciously.

Eventually we would expect evidence such as our work to be used to improve computer vision algorithms for this task. The key benefit from our technique is that the learning of the really hard tasks (detecting genuine smiles in this case) has already been done by the human being in their normal life. The physiological signal collection and analysis is much more straightforward than designing such a computer vision algorithm, and our signal collection and analysis can be applied to other emotions (Chen et al., 2017), while a completely new computer vision algorithm would need to be developed for each emotion.

Human verbal responses are easy to quantify and collect during experiments, and reflect conscious human behaviour in a fashion which superficially seems to be objective. It is a one-dimensional approach to understanding human behaviour that fails to address cognitive and biological processes and does not account for the unconscious mind's thoughts, feelings, and desires (Mills, 2000). A further limitation of this process is that facial expressions can be intentionally controlled, and observers may misrepresent their reaction to the viewed emotional faces (Soleymani et al., 2012b; Hess and Kleck, 1994). So in areas which are emotionally significant for human beings, we should not expect verbal responses to be objective and reliable (Horikawa et al., 2013; Plested et al., 2017). In some settings, explicit verbal responses are not possible to collect (Horikawa et al., 2013).

On the other hand, physiological responses are automatic reactions that trigger physical responses to stimulus, and have the advantage of immediately being affected by observing facial changes that cannot be faked voluntarily or assessed visually (Kim and Andre, 2008; Soleymani et al., 2012b; Soleymani et al., 2012a). Most of us are familiar with these automatic and instinctive physiological responses we experience every day, but we typically remain unaware about their details, such as how galvanic skin responses and pupillary responses change due to stimuli and under pressure (Teatero and Penney, 2015). There is also evidence that observers' physiological responses can form or evaluate another's mental state (Shah et al., 2017), perhaps via subjective feelings which allows us to judge others' facial expression. An important benefit of physiological measurement is that it is not easy to control voluntarily and provides spontaneous and non-conscious outcomes. Thus, it is therefore plausible that observers' sub-conscious 'choices' (physiological responses) can provide higher accuracy compared to their conscious choices (verbal responses). Overall, the results of the experiment shows that we can classify real and other smiles via observers' innate and non-conscious physiological responses that are controlled by the autonomic nervous system.

We note that many physiology-based works assume a coherent relationship between explicit (conscious) user responses and sub-conscious physiological signals. E.g., (Soleymani et al., 2009; Soleymani et al., 2012a) employ user ratings as the ground truth, and study the effectiveness of physiological responses towards predicting the user ratings. Our works show that there appears to be substantial similarity between conscious and sub-conscious responses in the area of emotion recognition which leads to the observed difference in survey and physiology-based classifications. The only similar work we can find is in musical emotion recognition where the ground truth does not come from an individual's own reactions, where physiological classification gives results in the range of 87–95% [(Kim and Andre, 2008), Lin et al. (2009)]. In comparing observers' recognition of depression level, a similar result to ours was obtained, with physiological signal classification being 79% as compared to user responses being 47% (Plested et al., 2017).

4. Conclusion

In this paper, two types of physiological signals were investigated,

while watching emotion containing video stimuli, along with recording the observer's judgements via a Likert scale, to discriminate genuine from acted smiles. It was a challenging task, because the recorded physiological signals were highly noisy. Different noise removal techniques with an advanced feature selection method were applied and the highest classification accuracy was found to be 97.8% by analysing 450 features of LEPD. The observers were only 52.7% (on average) to 68.4% accurate (by voting) according to their self-report. These results are in the normal range reported in the literature over multiple studies for determining real smiles from surveys. The result of 52.7% accuracy in conscious (self-report) discrimination of smile genuineness seems too low to be plausible given how important smiles are for humans, though perhaps humans are only accurate in groups, with our 20 subjects being about 15% better as a group. A result of 97.8% from pupillary response suggests that at non-conscious levels we are very good at detecting genuine smiles, perhaps reflecting the fact that this identification can feed into our emotional responses to others, and perhaps even that there is a benefit to a relatively low level of conscious identification of genuine smiles – it may be important in social interactions to be able to accept smiles or other expressions at 'face value', consciously. Our future work will consider extending this work to single images, more complex videos, group expressions in still and video images, to other facial expressions, and the use of virtual or synthesised faces (Asthana et al., 2009).

Acknowledgments

The authors wish to thank all the observers for their participation, and the anonymous referees for their helpful comments to improve our paper.

References

Ambadar, Z., Cohn, J.F., Reed, L.I., 2009. All smiles are not created equal: morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. J. Nonverbal Behav. 33 (1), 17–34.
Asthana, A., Gedeon, T., Goecke, R., Sanderson, C., 2009. Learning-based face synthesis for pose-robust recognition from single image. In: British Machine Vision Conference 2009, pp. 1–10 British Machine Vision Association and Society for Pattern Recognition, 2009.
Bilalpur, M., Kia, S.M., Chawla, M., Chua, T., Subramanian, R., 2017. Gender and emotion recognition with implicit user signals. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. Glasgow, UK. pp. 379–387.
Birdwhistell, R., 1970. Kinetics in Context. University of Pennsylvania Press, Philadelphia.
Bradley, M.M., Miccoli, L., Escrig, M.A., Lang, P.J., 2008. The pupil as a measure of emotional arousal and autonomic activation. Psychophysiology 45 (4), 602–607.
Chen, L., Gedeon, T., Hossain, M.Z., Caldwell, S., 2017. Are you really angry? Detecting emotion veracity as a proposed tool for interaction. In: Proceedings of the 29th Australian Conference on Human-Computer Interaction (OzCHI '17). Brisbane, QLD, Australia. pp. 5 pages.
Cohn, J.F., Schmidt, K.L., 2004. The timing of facial motion in posed and spontaneous smiles. Int. J. Wavelets, Multi-resolution Inf. Process. 2, 1–12.
Damasio, A.R., 1994. Descartes' Error: Emotion, Reason, and the Human Brain. Grosset/Putnam.
Dhall, A., Goecke, R., Joshi, J., Sikka, K., Gedeon, T., 2014. The second emotion recognition in the wild challenge. In: 16th ACM Int'l Conf. on Multimodal Interaction.
Dibeklioglu, H., Salah, A.A., Gevers, T., 2015. Recognition of genuine smiles. IEEE Trans. Multimed. 17 (March (3)), 279–294.
Ekman, P., Davidson, R.J., Friesen, W.V., 1990. The Duchenne smile: emotional expression and brain physiology: II. J. Pers. Soc. Psychol. 58, 342–353.
Ekman, P., Friesen, W.V., 1982. Felt, false, and miserable smiles. J. Nonverbal Behav. 6, 238–252.
Frank, M.G., Ekman, P., Friesen, W.V., 1993. Behavioral markers and recognizability of the smile of enjoyment. J. Pers. Soc. Psychol. 64 (January (1)), 83–93.
Gan, Quan, Wu, Chongliang, Wang, Shangfei, Ji, Qiang, 2015. Posed and spontaneous facial expression differentiation using deep Boltzmann machines. In: Affective Computing and Intelligent Interaction tiation using deep Boltzmann machines, pp. 643–648.
Gifford, R., Ng, C.F., Wilkinson, M., 1985. Nonverbal cues in the employment interview: links between applicant qualities and interviewer judgments. J. Appl. Psychol. 70 (4), 729–736.
Guo, R., Li, S., He, L., Gao, W., 2013. Pervasive and unobtrusive emotion sensing for human mental health. In: 7th International Conference on IEEE on Pervasive Computing Technologies for Healthcare (PervasiveHealth), pp. 436–439 5-8 May.
Healey, J., Picard, R., 1998. Digital processing of affective signals', in acoustics, speech and signal processing. In: Proc. of IEEE int'l conference, pp. 3749–3752.
Hess, U., Kleck, R.E., 1994. The cues decoders use in attempting to differentiate emotion-elicited and posed facial expressions. Eur. J. Soc. Psychol. 24, 367–381.
Hoque, M., Morency, L.P., Picard, R.W., 2011. Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science. Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science 6974. pp. 135–144.
Hoque, M.E., Picard, R.W., 2011. Acted vs. natural frustration and delight: many people smile in natural frustration. In: 9th IEEE International Conference on Automatic Face and Gesture Recognition (FG'11). Santa Barbara, CA, USA. March.
Horikawa, T., Tamaki1, M., Miyawaki, Y., Kamitani, Y., 2013. Neural decoding of visual imagery during sleep. Sci. Immun. 340 (6132), 639–642.
Hossain, M.Z., Gedeon, T., Sankaranarayana, R., 2016a. Observer's galvanic skin response for discriminating real from fake smiles. In: The 27th Australasian Conference on Information System, pp. 1–4.
Hossain, M.Z., Gedeon, T., Sankaranarayana, R., Apthorp, D., Dawel, A., 2016b. Pupillary responses of asian observers in discriminating real from fake smiles: a preliminary study. In: 10th International Conference on Methods and Techniques in Behavioral Research, pp. 170–176.
Hossain, M.Z., Kabir, M.M., Shahjahan, M., 2016c. A robust feature selection system with Colin's CCA network. Neurocomputing 173 (January (3)), 855–863 Elsevier.
Huang, J., Cai, Y., Xu, X., 2007. A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern Recogn. Lett. 28, 1825–1844.
Kim, J., 2007. Bimodal emotion recognition using speech and physiological changes. In: Grimm, M., Kroschel, K. (Eds.), Robust Speech Recognition and Understanding. I-Tech Education and Publishing, Vienna, Austria, pp. 265–280.
Kim, J., Andre, E., 2008. Emotion recognition based on physiological changes in music listening. IEEE Trans. Pattern Anal. Mach. Intell. 30 (December (12)), 2067–2083.
Lai, P.L, Fyfe, C., 1999. A neural implementation of canonical correlation analysis. Neural Netw. 12 (10), 1391–1397.
Levenson, R.W., Ekman, P., Friesen, W.V., 1990. Voluntary facial action generates emotions-specific automatic nervous system activity. Psychophysiology 27, 363–384.
Lin, Y.P., Wang, C.H., Wu, T.L., Jeng, S.K., Chen, J.H., 2009, April, April. EEG-based emotion recognition in music listening: a comparison of schemes for multiclass support vector machine. In: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, pp. 489–492.
Lithari, C., Frantzidis, C.A., Papadelis, C., Vivas, AnaB., Klados, M.A., Kourtidou-Papadeli, C., Pappas, C., Ioannides, A.A., Bamidis, P.D., 2010. Are females more responsive to emotional stimuli? a neurophysiological study across arousal and valence dimensions. Brain Topogr. 23 (1), 27–40.
Liu, C., Conn, K., Sarkar, N., Stone, W., 2008. Physiology-based affect recognition for computer-assisted intervention of children with **autism spectrum d**isorder. Int. J. Hum. Comput. Stud. 66 (9), 662–677.
Marsaglia, G., Tsang, W., Wang, J., 2003. Evaluating Kolmogorov's distribution. J. Stat. Softw. 8 (18).
Mathôt, S., 2013. A simple way to reconstruct pupil size during eye blinks. available at. http://dx.doi.org/10.6084/m9.figshare.688001.
Mills, J.A., 2000. Control: A History of Behavioral Psychology. NYU Press, New York.
Ochs, M., Niewiadmoski, R., Pelachaud, C., 2010. How a virtual agent should smile? Morphological and Dynamic Characteristics of virtual agent's smiles. In: Intelligent Virtual Agent Conference (IVA). Philadelphia, USA.
Oliveira, F.T.P., Aula, A., Russell, D.M., 2009. Discriminating the relevance of web search results with measures of pupil size. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA. ACM, pp. 2209–2212.
Pamplona, V.F., Oliveira, M.M., Baranoski, G.V.G., 2009. Photorealistic models for pupil light reflex and iridal pattern deformation. ACM Trans. Graphics 28 (4), 1–12.
Parsons, C.K., Liden, R.C., 1984. Interviewer perceptions of applicant qualifications: a multivariate field study of demographic characteristics and nonverbal cues. J. Appl. Psychol. 69 (4), 557–568.
Partala, T., Surakka, V., 2003. Pupil size variation as an indication of affective processing. Int'l J. Human-Comput. Stud. 59, 185–198.
Petridisa, S., Martineza, B., Pantica, M., 2013. The MAHNOB laughter database. Image Vis. Comput. 31 (2), 186–202.
Picard, R.W., 2000. Affective Computing. the MIT Press, Cambridge, Massachusetts, London, England, pp. 175.
Plested, J.F., Gedeon, T., Zhu, X.Y., Dhall, A., Goecke, R, 2017. Detection of universal cross-cultural depression indicators from the physiological signals of observers. In: 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, pp. 185–192.
Redi, J., Povoa, I., 2014. Crowdsourcing for rating image aesthetic appeal: better a paid or a volunteer crowd? In: International ACM Workshop on Crowdsourcing for Multimedia, pp. 25–30 07 November.
Shah, P., Catmur, C., Bird, G., 2017. From heart to mind: linking interoception, emotion, and theory of mind. Cortex 1–4.
Shlenker, B.R., 1980. Impression Management: The Self-Concept, Social Identity, and Interpersonal Relations. Brooks/Cole Pub. Co., Monterey, CA, pp. 344.
Soleymani, M., Chanel, G., Kierkels, J.J.M., Pun, T., 2009. Affective characterization of movie scenes based on content analysis and physiological changes. Int'l J. Semant. Comput. 3 (June (2)), 235–254.
Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M., 2012a. A multimodal database for affect recognition and implicit tagging. IEEE Trans. Affective Comput. 3 (April (1)), 42–55.
Soleymani, M., Pantic, M., Pun, T., 2012b. Multimodal emotion recognition in response to videos. IEEE Trans. Affect. Comput. 3 (2), 211–223.
Teatero, ML, Penney, AM, 2015. Fight-or-flight response. In: Milosevic, I., McCabe, RE. (Eds.), Phobias: The Psychology of Irrational Fear. Greenwood, Santa Barbara, CA

2015.
Valstar, M.F., Gunes, H., Pantic, M., 2007. How to distinguish posed from spontaneous smiles using geometric features. In: Proceedings of ACM International Conference on Multimodal Interfaces (ICMI'07). Nagoya, Japan. pp. 38–45 November.
Willenbockel, V., Sadr, J., Fiset, D., Horne, G.O., Gosselin, F., Tanaka, J.W., 2010. Controlling low-level image properties: the SHINE toolbox. Behav. Res. Methods 42 (3), 671–684.
Zheng, W.L., Dong, B.N., Lu, B.L., 2014. Multimodal emotion recognition using EEG and eye tracking data. In: 36th Annual International Conference of the IEEE Engineering on Medicine and Biology Society (EMBC), pp. 5040–5043 26-30 Aug.
Zhou, Feng, Qu, Xingda, Martin, G, 2011. Helander, and Jianxin (Roger) Jiao. "Affect prediction from physiological measures via visual stimuli. Int. J. Hum. Comput. Stud. 69 (12), 801–819.