2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)

Eye-based Continuous Affect Prediction

Jonny O'Dwyer, Niall Murray, Ronan Flynn
Deptartment of Computer & Software Engineering
Athlone Institute of Technology
Athlone, Ireland
j.odwyer@research.ait.ie, nmurray@research.ait.ie, rflynn@ait.ie

***Abstract*—Eye-based information channels include the pupils, gaze, saccades, fixational movements, and numerous forms of eye opening and closure. Pupil size variation indicates cognitive load and emotion, while a person's gaze direction is said to be congruent with the motivation to approach or avoid stimuli. The eyelids are involved in facial expressions that can encode basic emotions. Additionally, eye-based cues can have implications for human annotators of affect. Despite these facts, the use of eye-based cues in affective computing is in its infancy and this work is intended to start to address this. Eye-based feature sets, incorporating data from all of the aforementioned information channels, that can be estimated from video are proposed. Feature set refinement is provided by way of continuous arousal and valence learning and prediction experiments on the RECOLA validation set. The eye-based features are then combined with a speech feature set to provide confirmation of their usefulness and assess affect prediction performance compared with group-of-humans-level performance on the RECOLA test set. The core contribution of this paper, a refined eye-based feature set, is shown to provide benefits for affect prediction. It is hoped that this work stimulates further research into eye-based affective computing.**

***Index Terms*—eye gaze, pupillometry, eye closure, affective computing, feature engineering**

I. Introduction

A connection between the eyes and displays of emotion has been accepted for many years [1]. Eye-based cues are of increasing interest to the research community for automatic emotion classification and affect prediction. Such cues include eye gaze, eye saccades, pupillometry, fixational eye movements, and various forms of eye opening and closure events. The eyes can provide information on attention [2]–[4], perception [5], [6], social and emotional cues [5]–[10], cognitive load [11], [12], locomotion [2], and mental and physical pathology [3], [9], [10], [13].

While the pupil size is known to vary under environmental, pathological and pharmacological conditions [13], there is a body of evidence suggesting its efficacy for studies of neuropsychological and affective responses in healthy inidviduals. Pupillometry studies from psychology showed the pupils to be responsive during emotional arousal [5], [6] and monetary incentive or penalty [14] during a memory task. In neurophysiological literature, it was demonstrated that the pupils were responsive to autonomic nervous system stimulation [10], which is known to generate response output under numerous emotional states [15]. Evidence has also been provided for pupilliary responses to reward expectation [9].

Eye gaze, the line of sight between an individual and an object of fixation, has been referred to as central to social understanding [3]. The shared signal hypothesis [7], [8] postulates that one's gaze is congruent with their emotional display if the gaze signal matches their underlying motivation to approach or avoid stimuli, direct gaze anger being a congruent threat cue for example. Eye-based cues, in addition to allowing subjects outwardly display their attention and/or social signal, can also have effects on social and emotional decoders or perceivers. Hess [16] reported behavioural changes in subjects who viewed image stimuli that had different pupil sizes; images with larger pupils were perceived to be more attractive than those with smaller pupils. Direct gaze (subject gazing directly at you) observation has been shown to contribute to attentional blink [17], which is reduced or degraded attention at a later duration in time due to stimuli received at an earlier point in time. Therefore, it is important to understand eye-based cues from the perspective of both emotion encoders and decoders.

Blinks, winks, complete eye closure and partial eye opening or closure events are involved in certain eye gazes and saccades [18], and facial expressions of emotion [19]. Additionally, it was suggested that eye gaze and blink share common signalling pathways [4], which makes incorporation of eye closure and blink events important for a complete investigation of potential eye-based cues for affective computing.

Eye-based cue estimation from video frames is now possible due to advances in computer vision models and tools [20], which provides new opportunitites for researchers to model affect. However, some contextual factors must be in place for these cues to be available. In the case of OpenFace [20], for example, a face must be detected in a video frame for visual features to be detected, and, in addition, subjects must provide consent for recording capture and research use. Furthermore, if eye-based cues from video are shown to be efficacious for affect prediction, ethical issues must be considered. Paridoxically, due to the potential of physcially non-intrusive feature gathering in audio-video streams/recordings, some of the most personally intrusive affect prediction may be possible including covert affect prediction or unwanted affect mining and reporting. It is hoped that benefits of overt audio-video affect prediction can be realised in spite of potential misuse downsides of forthcoming technologies.

This work was supported by the Irish Research Council (Grant No. GOIPG/2018/2030).

Soceital benefits of positive uses of the proposed features, in combination with other audio-video cues include: more accessible and personalised remote psychoanalysis, improved psychopathology prediction and improved content or service delivery in domains like education, healthcare or home care.

Based on identified trends and research opportunities in the literature, this work presents a comprehensive study of eye-based cues for for arousal and valence prediction. The core contribution of this paper is a proposal for feature sets from an alterative modality that make use of the aforementioned eye-based information channels for the purpose of continuous affect prediction. A secondary contribution of this paper includes the investigation of a feature selection method that incorporates annotator delay into the selection process. Both contributions are validated by way of continuous arousal and valence prediction experiments on the RECOLA [21] corpus. A practical use analysis combining the proposed eye-based feature sets with speech is also carried out. Code and feature sets used in the experiments will be made available to the research community on GitHub[1].

The remainder of this paper is structured as follows. Related work is presented in Section II. Data and tools for the experiments are detailed in Section III. In Section IV, the methods for feature selection and evaluation are described. Experimental results, along with discussion, are provided in Section V and the paper is concluded in Section VI.

II. Related Work

In affective computing, eye-based features are being investigated for numerous tasks. An EyeLink 1000 eye tracking device was used in [24] to gather measurements from individuals observing emotion provoking images and a decision tree neural network was able to classify the individual's emotion responses correctly at a 53.6% rate. For affect-level recognition using support vector machine (SVR), eye-based features, including statistics and spectral power calculations from the descriptors: pupil diameter, gaze distance, eye blinking, and x and y gaze coordinates gathered using a Tobii X120 eye tracking device were used in [22]. Eye-based features performed best when compared to electroencephalogram (EEG) and peripheral physiology measures, while bimodal fusion of EEG and eye-based features performed best overall (arousal = 67.7%, valence = 76.1%). Eye gaze was combined with speech in [25], where a feature set similar to that of [22] was used. Additional statistics were gathered for eye scan paths and eye closure features were measured by frame counts. Results achieved in [25] showed that eye gaze, when combined with speech as part of a feature fusion SVR system, could improve arousal prediction compared to that of unimodal speech while, model fusion improved valence prediction most compared to unimodal speech. Psychopathological affective computing work incorporating eye-based features as part of multimodal approaches include post traumatic stress disorder estimation [26] and depression recognition [27], [28].

The affective computing community is acknowledging the potential for eye-based cues for system development. However, the full benefit of eye-based cues estimated from video for affective computing purposes is not yet known as several works use specialised eye recording devices [22], [24] for their experiments. Although research on eye-based continuous affect prediction does exist [25], a limitation of their work was a focus only on free-roaming eye gaze and eye closure-based features. This paper is the most comprehensive study of eye-based cues from video for continuous affect prediction to date. With inspiration from the literature, direct gaze knowledge and pupil wavelet cues, designed for continuous affect prediction, are investigated in addition to eye gaze and closure features, all of which are gathered non-intrusively from video and evaluated on the RECOLA [21] corpus.

III. Experiment Design

A. Data and Tools

The RECOLA [21] corpus is used as the experimental data set for this work. RECOLA is an affective data set comprised of audio-visual and physiological recordings of subjects cooperating on a task in French. Arousal and valence annotations, ranging from -1.0 to +1.0, are provided with the set at a rate of 25 values per second. Recordings of 23 subjects available in the set were paritioned into training, validation and test sets with the aim of matching the distributions used in [31]. Specifically, the training set is comprised of subjects [P16, P17, P19, P21, P23, P26, P30, P65], the validation set includes subjects [P25, P28, P34, P37, P41, P48, P56, P58], and the test set includes subjects [P39, P42, P43, P45, P46, P62, P64].

Key software for the experiments included: OpenFace (version 2.0.6) [20] for gaze, eye blink/closure and pupils estimation from video and the CUDA RecurREnt Neural Network Toolkit (version 0.2 rc1) [29] for BLSTM-RNN model training and prediction.

B. Initial Eye-based Features

The initial eye-based features were extracted from 6 binary [direct gaze, gaze approach, eyes fixated, eye closure/blink, pupil dilation and pupil constriction] and 7 numerical [eye blink intensity, pupil diameter, Δpupil diameter, x and y gaze angles, and Δx and Δy gaze angles] low level descriptors (LLDs). The LLDs were gathered frame-wise from each subject video recording. The pupil estimation was based on the left eye, due to OpenFace's model implementation, and gaze angles were measured in radians. All but eye closure/blink, blink intensity and x and y gaze angles LLD features required calculation in this work. The direct gaze binary variable required a human coder to view cropped images of subjects' faces and provide true/false ratings based on whether they thought the interacting subject was looking at the interlocutor (direct gazing) or not.

For the pupil modality, mid-level features were gathered using an 8 second time window (200 frames at 25 frames per second) moved forward at a rate of 1 frame per

[1] https://github.com/sri-ait-ie/Non-intrusive_affective_computing

interval. Specifically, 10-order Daubechies [33] discrete wavelet transform features were gathered at 7 levels of decomposition, the maximum possible decomposition for the time window used, based on [10]. Following the low- and mid-level descriptor feature extraction, statistics were applied on these descriptors to achieve an initial set of 292 features that comprised of 69 eye gaze features, 209 pupillometry features and 14 eye closure features. The initial feature list, which is computed using an 8 second time window, moved forward at a rate of 1 frame per interval included:

- Direct gaze, pupil dilation and pupil constriction (12 features):
ratio, time seconds: mean, max, total

- Gaze approach, eyes fixated and eye closure/blink (14 features):
ratio, time seconds: min (min not applied for gaze approach), median, mean, max

- Pupil diameter, Δpupil diameter, x, y gaze angles, Δx and Δy gaze angles (84 features):
min, max, mean, median, quartile 1, quartile 3, skewness, kurtosis, standard deviation, inter-quartile range (IQR) 1-2, IQR 2-3, IQR 1-3, linear regression slope, linear regression intercept

- Eye blink intensity (9 features):
max, mean, median, quartile 3, standard deviation, IQR 1-2, IQR 2-3, linear regression slope, linear regression intercept

- 10-order Daubechies scale and approximation wavelet coefficients at 7 levels of decomposition (173 features):
min, max, median, quartile 1, quartile 3, skewness, kurtosis (kurtosis not applied for final scale and approximation wavelet coefficients), standard deviation, IQR 1-2, IQR 2-3, IQR 1-3, RMS, zero crossing rate (ZCR) (ZCR not applied to scale coefficients)

C. *BLSTM-RNN Training and Evaluation Method*

In this work, BLSTM-RNN was used to train models for feature set appraisal. The training method used largely follows that of Ringeval et al. [31]. Single-task models were trained using BLSTM-RNN with 2 hidden layers, each with 40 and 30 nodes respectively, with a sum-of-squared-errors (SSE) objective function. All input features and regression targets were standardised using the parameters mean and standard deviation, computed on the training set. The network learning rates were set at 10^{-5} and a random seed of 1787452436 was used throughout the experiments. Gaussian noise with a standard deviation 0.1 was added to all input features prior to training. BLSTM-RNN models were trained for a maximum of 100 epochs, however, training was stopped when no performance increase (lower SSE) was observed on the validation set after 10 epochs.

Following the training phase, network models were evaluated and selected using concordance correlation coefficient (CCC) [38], [39], where higher CCC is better. The CCC measure penalises correlated time-series by applying a penalty of mean-squared error as in (1), where x represents predicted values, y represents ground-truth values, σ_{xy} is the covariance, σ^2 is the variance and μ is the mean.

$$CCC = \frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{1}$$

D. *Feature Selection*

The feature selection method used followed a simple approach of mutual information (MI) estimation to regression target-based filtering. MI is “the amount of information that one random variable contains about another random variable” [39, p.18]. An additional component that was considered for the feature selection process included a now common affect learning parameter, ground-truth backward time-shift. It is accepted in the literature that human annotators produce a delay when providing their ground-truth ratings [28], [32], [37]. To mitigate for this, researchers now incorporate time-shifting of ground-truth values back in time for continuous affect prediction purposes [28], [32], [37]–[39]. However, ground-truth time-shifts are evaluated on the validation set using model performance CCC and the ground-truth time-shifting may occur prior to achieving the best feature set. Therefore, the best inputs to models for ground-truth time-shift assessment may not be present. When carrying out feature selection in this work, MI was estimated between the input features and ground-truth regression targets *before*, *during*, and *after* ground-truth time-shifting has occurred.

The MI thresholds evaluated were 0.1, 0.15, 0.2. In total, 23 ground-truth time-shifts were applied ranging from 0 (not applied) to 4.4 seconds, altered in backward shifts of 0.2 seconds. Features that had a MI less than the threshold under test were removed as these features were deemed independent of arousal or valence, and therefore poor predictors. For the ground-truth time-shift iterated MI selection, the threshold that provided the best performing feature set *before* any ground-truth time-shift was again used *during* feature selection after each backward time-shift iteration. Ground-truth time-shift is referred to as D_s for the remainder of this work, where s is the value in seconds for the delay D applied.

E. *Practical Significance Evaluation*

In order to assess the practical significance of the eye-based feature sets following feature selection, two steps are taken. Firstly, group-of-humans-level performance estimates were calculated using the RECOLA training set arousal and valence annotations. These estimates consist of averages for annotator-to-annotator CCC for the training set arousal and valence ratings provided. The estimates are important for continuous arousal and valence prediction in order to have a practical baseline for automatic systems. The results of these calculations are group-of-humans-level arousal prediction CCC = 0.341, and valence prediction CCC = 0.383.

While comparisons between the group-of-humans-level performance and that of the eye-based cues validation set results can be made, this is an unfair comparision. The human coders have access to additional modalities to reach their decision while the eye-based models and features have been tuned on the validation set. It is clear that advantages from both comparison groups have been taken away, for example, both human generalisation and computing machine capabilities are not being taken advantage of. Therefore, the second step in the practical significance evaluation is to offer as much of a like-for-like comparison as possible between the group-of-humans-level and eye-based performances. Since the human annotators have the advantage of both audio and visual data for decoding and rating affect, it was decided that speech would be combined with the eye-based features for practical evaluation and comparison to that of the human coders. The eGeMAPS [40] speech feature set was used in the experiments and it was combined with the eye-based features using early feature fusion. Performances of bimodal systems were assessed versus unimodal speech on the validation set, and, if improvements were observed, a test set pass was carried out for the speech and eye-based systems, followed by group-of-humans-level performance comparisons.

F. Eye-based Arousal and Valence Feature Set Proposals

The experiments culminate with the results for eye-based affective computing sets, intended for the continuous prediction of arousal and valence affect dimensions based on video input. The final ground-truth time-shift parameter required for use with the sets, along with the proportions of retained features, are given along with discussion of the final feature sets. The top 20 performers from each of the arousal and valence sets, ranked in terms of both linear, absolute value PCC, and nonlinear, MI, metrics, are provided for the final, proposed feature sets. The full list of features will be made available in the repository for this work.

IV. Results and Discussion

A. Feature Selection

The results are given in Table I for feature selection carried out *before* application of D_s. Table I shows that this feature selection technique is effective. Performance increases in terms of validation set CCC along with feature set size reductions are always observed compared to when feature selection was not applied. These results mean that a more optimal sub-set of features can be used for further feature selection incorporating D_s. The sets gathered from this experiment include *before* condition feature sets for arousal and valence of sizes 147 and 152 respectively. Additionally, parameters resulting from this experiment include MI threshold values, 0.15 for arousal and 0.2 for valence, which are used for the D_s iterated feature selection from the feature sets at each iteration.

TABLE I: Validation Set Feature Selection BLSTM-RNN Results Before Ground-truth Time-shift

MI $<$	**Arousal**			**Valence**		
Filter	**SSE**	**CCC**	**Features**	**SSE**	**CCC**	**Features**
N/A	0.368	0.106	292	0.417	0.000	292
0.1	0.361	0.15	182	0.414	0.029	198
0.15	0.346	**0.188**	147	0.414	0.032	152
0.2	0.352	0.187	99	0.405	**0.058**	124

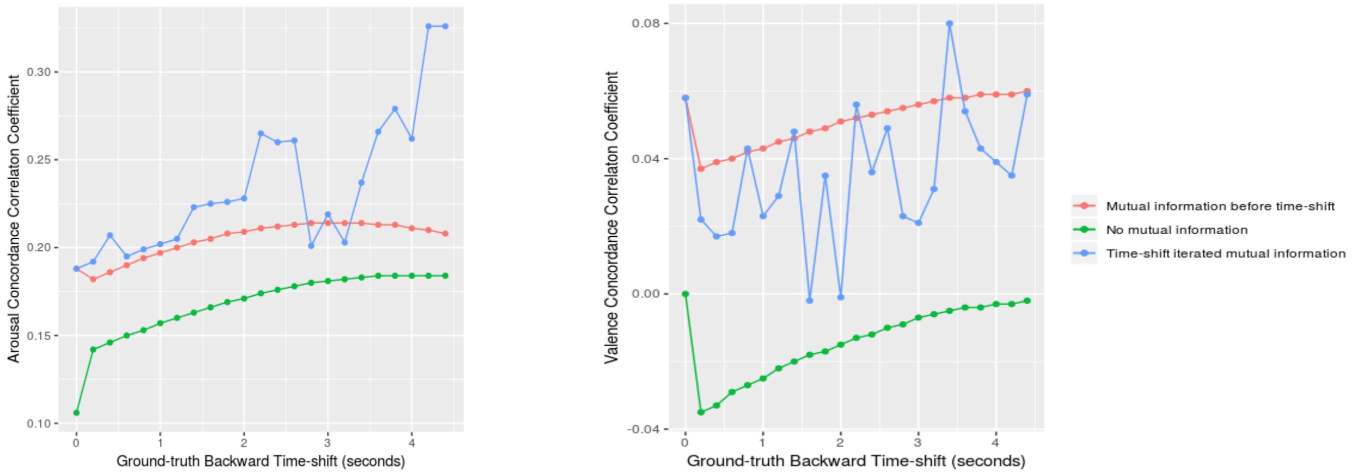
The results depicted in Fig. 1 (a) showed that the D_s iterated MI feature selection (blue line) for arousal performs better than the other methods at this experimental stage. Additionally, assessing how much greater the iterated MI feature selection was compared to the MI selection (red line) *before* D_s shifting was found to be statistically significant (Wilcoxon rank sum test, W = 378.5, p-value = 0.006). The highest performing D_s value for arousal was 4.4 seconds, with an eye-based feature count of 151 features that was achieved using the MI threshold = 0.15. The validation set arousal CCC for this system was 0.326. Another result from this experiment included the highest performing D_s for the group where no MI feature selection was applied, which was 4.4 seconds. Therefore, when MI is to be utilised *after* D_s, a 4.4 seconds D_s is applied to the arousal ratings first.

Fig. 1 (b) shows the D_s iterated MI feature selection for valence. It is clear from this graph that not utilising feature selection prior to, or during D_s shifting, model building and evaluation, can be detrimental. Both MI feature selection *before* and in the iterated fashion, performed better than when MI was not applied and D_s effects were evaluated. The top performer in terms of CCC came from the D_s iterated MI group, a value of 0.08, and this was achieved using D_s = 3.4 seconds applied to valence ground-truth. The top performing feature set size includes 128 of the originally proposed 292 features, gathered using an MI threshold of 0.2. Due to the difficultly in concluding which D_s performed best for the group where no MI feature selection was applied, it was decided against applying any feature selection *after* D_s for this group.

The final MI feature selection to be applied *after* D_s shifting could only be applied to the arousal set. The results for this feature selection are given in Table II. The poor performance of the valence feature set without MI applied prior to D_s shifting (performance always dropped) resulted in no plausible selection for D_s and MI selection *after* shifting for this affect dimension. The top performing result in Table II is the same as from the D_s iterated MI selection, with matching feature set size, SSE, and CCC. These early results indicate that, for arousal, applying feature selection *during*, or *after* D_s shifting makes no difference in terms of CCC performance. An interesting point to note from Table II is that the the top performing system, achieving a CCC of 0.326, fell just 4.4% short of the group-of-humans-level CCC performance estimate of 0.341, which provides early evidence of promise for this eye-based feature set for arousal prediction.

B. Practical Significance Evaluation

The results for the practical significance evaluation are given in Table III for the speech feature set alone and when



(a)

(b)

Fig. 1: Arousal (a) and valence (b) validation set CCC scores under different ground-truth backward time-shift (D_s) conditions. D_s shifts evaluated ranged from 0 (not applied) to 4.4 seconds, in increments of 0.2 seconds.

TABLE II: Validation Set Feature Selection BLSTM-RNN Results After Ground-truth Time-shift Applied

MI <	**Arousal**		
Filter	**SSE**	**CCC**	**Features**
N/A	0.347	0.184	292
0.1	0.327	0.277	187
0.15	0.313	**0.326**	151
0.2	0.334	0.205	111

combined with the final eye-based feature sets. The same D_s shift and MI feature selection values as used for the eye-based features were applied. The results show that when eye-based features are combined with speech, performance benefits can be achieved. A relative increase of 9.19% above that of unimodal speech is observed in Table III. Due to the increase in performance of the bimodal system for arousal, a test set pass was performed for this affect dimension. This result can be compared to other multimodal arousal prediction test set results from the literature that incorporate the eGeMAPS [40] speech feature set, which are given in the table. Unfortunately, the bimodal valence system did not perform well, there was a performance decrease found for this dimension. Further experimental work is required for valence prediction using eye-based cues from video, based on the results presented they appear unsuitable for valence prediction either in the presence or absence of speech. The test set result CCC of 0.72 for arousal achieved by the speech and eye-based system in Table III compares favourably to that of the group-of-humans-level CCC of 0.341, greatly outperforming this baseline. Additionally, the arousal result from this work is comparable to that of other published works. While the performance is lower than that of [37] and [41], it must be mentioned that the authors of these works had access to further audio and visual cues in addition to eGeMAPS [40] and physiological measures for prediction. The results achieved indicate promise for these eye-based cues for continuous arousal prediction, especially when considered in multimodal systems.

TABLE III: Final BLSTM-RNN Results For Systems Including Speech

System	**Arousal**		**Valence**	
(Evaluation)	**SSE**	**CCC**	**SSE**	**CCC**
Speech-based (Validation)	0.192	0.675	0.391	**0.103**
Speech & Eye-based (Validation)	0.17	**0.737**	0.402	0.059
Speech & Eye-based (Test)	-	0.72	-	-
He et al. [37] (Test)	-	0.747		
Brady et al. [41] (Test)	-	0.77		

C. Eye-based Arousal and Valence Feature Set Proposals

The final 151-dimensional eye-based arousal feature set produced from this work was gathered using the MI threshold set to 0.15 during D_s iterated feature selection, with D_s = 4.4 seconds. The retention proportion from each of the information chananels for the final set are as follows: eye gaze (55 of 69 features), pupillometry (85 of 209 features) and eye closure/blink (11 of 14 features). The retention proportions for the 128-dimensional eye-based valence feature set included: eye gaze (46 of 69 features), pupillometry (73 of 209 features) and eye closure/blink (9 of 14 features). This feature set was gathered using D_s = 3.4 seconds along with the MI threshold set to 0.2 which was applied during D_s iterated feature selection. For both arousal and valence feature sets, it is interesting to note that all 4 direct gaze binary-based features were retained in the final sets, and all of both pupil constriction and pupil dilation binary-based features were removed from the final sets using the MI feature selection.

TABLE IV: Top Features Ranked by Correlation and Mutual Information With Arousal and Valence

Arousal				**Valence**			
Feature	**PCC**	**Feature**	**MI**	**Feature**	**PCC**	**Feature**	**MI**
gaze x max	0.361	gaze x max	0.57	scale coeffs l1 max	0.321	Δ gaze y max	0.594
gaze x quartile 3	0.33	Δ gaze y max	0.547	scale coeffs l2 max	0.319	Δ gaze y min	0.568
gaze x mean	0.304	Δ gaze y min	0.539	pupil diameter mm max	0.315	gaze x max	0.548
gaze x median	0.299	gaze y min	0.514	gaze x max	0.306	Δ gaze x max	0.537
gaze y min	-0.266	eye blink intensity max	0.508	scale coeffs l3 max	0.295	Δ gaze x min	0.528
Δ gaze y inter-quartile range (IQR) 1-3	0.242	Δ gaze x min	0.506	gaze x quartile 3	0.287	gaze x min	0.498
Δ gaze y quartile 3	0.236	Δ gaze x max	0.505	gaze x median	0.281	gaze y max	0.498
gaze x quartile 1	0.236	gaze x min	0.496	gaze x mean	0.278	eye blink intensity max	0.495
Δ gaze y IQR 1-2	0.234	gaze y max	0.469	scale coeffs l4 max	0.273	gaze y min	0.493
Δ gaze y IQR 2-3	0.231	Δ pupil diameter mm max	0.421	wavelet coeffs l2 SD	0.267	Δ pupil diameter mm min	0.482
Δ gaze y quartile 1	-0.226	gaze y median	0.41	wavelet coeffs l2 RMS	0.267	Δ pupil diameter mm max	0.437
gaze x standard deviation (SD)	0.225	gaze x quartile 3	0.404	pupil diameter mm quartile 3	0.249	pupil diameter mm max	0.414
direct gaze time ratio	0.224	Δ pupil diameter mm min	0.402	scale coeffs l5 max	0.24	gaze y quartile 3	0.407
wavelet coeffs l3 RMS	0.222	gaze y quartile 3	0.4	scale coeffs l1 quartile 3	0.236	gaze y median	0.403
wavelet coeffs l3 SD	0.222	Δ gaze y SD	0.4	gaze x quartile 1	0.237	max gaze fixation time	0.399
pupil diameter mm max	0.21	gaze x median	0.389	gaze y min	-0.226	gaze x median	0.397
gaze y quartile 1	-0.207	gaze x quartile 1	0.388	scale coeffs l2 quartile 3	0.226	Δ gaze y SD	0.396
gaze y mean	-0.203	gaze y quartile 1	0.385	wavelet coeffs l2 max	0.224	gaze x quartile 3	0.395
gaze y SD	0.2	max gaze fixation time	0.373	Δ gaze y IQR 1-2	0.218	scale coeffs l1 max	0.392
wavelet coeffs l2 SD	0.195	max eyes closed time	0.36	scale coeffs l2 max	0.218	gaze y quartile 1	0.39

The top 20 performers for the final feature sets are given in Table IV. The top-ranked performer of the arousal features is *gaze x max*, which is highest both in terms of PCC (0.361) and MI (0.57). The top performing valence features include Daubechies *scale coefficients l1 max*, which achieved a PCC of 0.321, and Δ*gaze y max*, which achieved a MI of 0.594. The top performers for arousal contain eye gaze features for 16 of the top 20 in terms of PCC and 17 of the top 20 in terms of MI. For the valence dimension, 13 of the top 20 features ranked by PCC are provided by pupil-based features, with the majority of these features coming from Daubechies wavelet coefficients (11 of 13). In terms of MI with valence, 15 of the top 20 features shown in Table IV are provided by eye gaze features.

V. Conclusions and Future Work

In this paper, cues from gaze, pupillometry and eye closure are proposed. Feature vectors were refined for continuous arousal and valence prediction using ground-truth backward time-shifting, feature selection and evaluation using BLSTM-RNN. Performance comparable to that of group-of-humans-level arousal CCC was achieved on the validation set for the eye-based cues on their own. Additionally, the results obtained show the benefit of combining the eye-based cues with speech for arousal prediction; the CCC for the bimodal system was 0.72 compared to the group-of-humans baseline of 0.341 on the RECOLA test set. Eye gaze features were shown to be particularly salient for arousal prediction from the eye-based cues with the majority of top 20 performers as measured by both linear and nonlinear relationships with arousal provided from gaze. The validation set performance of the eye-based features for valence was poor when combined with speech, providing performance degradation compared to unimodal speech. This study shows that valence features from eye-based cues gathered from video require further investigation prior to practical application. Potential avenues for further investigation of eye-based features for valence may include more advanced feature fusion, automatic feature learning and different temporal windows for feature extraction. The majority of top performers for the final valence feature set comprised of Daubechies wavelet pupillometry features for linear relationships with valence and eye gaze features for nonlinear relationships with valence.

Some limitations of this study include the nonoptimal fusion and ground-truth time-shift of the speech and eye-based cues, the human coder required to provide direct gaze binary annotations and the lack of consideration for other visual descriptor features in contrast/combination with the proposed eye-based cues. Future work includes incorporating the proposed eye-based cues for arousal into multimodal systems that include speech, facial expression and head pose, for comparison and combination with other visual descriptors as well as addressing the optimal fusion strategy for these cues. The evaluation of the features on other corpora is also planned.

REFERENCES

[1] C. Darwin, “The expression of the emotions in man and animals”, London, England: John Murray, 1872.

[2] O. Lappi, “Eye movements in the wild: Oculomotor control, gaze behavior & frames of reference”, Neuroscience & Biobehavioral Reviews, vol. 69, pp. 4968, Oct. 2016.

[3] R. J. Itier and M. Batty, “Neural bases of eye and gaze processing: The core of social cognition”, Neuroscience & Biobehavioral Reviews, vol. 33, no. 6, pp. 843863, Jun. 2009.

[4] U. Engelke et al., “Psychophysiology-Based QoE Assessment: A Survey”, IEEE Journal of Selected Topics in Signal Processing, vol. PP, no. 99, pp. 11, 2016.

[5] E. H. Hess and J. M. Polt, “Pupil size as related to interest value of visual stimuli”, Science, vol. 132, pp. 349350, 1960.

[6] J. M. Polt and E. H. Hess, “Changes in pupil size to visually presented words”, Psychonomic Science, vol. 12, no. 8, pp. 389390, 1968.

[7] R. B. Adams and R. E. Kleck, “Effects of Direct and Averted Gaze on the Perception of Facially Communicated Emotion”, Emotion, vol. 5, no. 1, pp. 311, Mar. 2005.

[8] R. B. Adams Jr. and R. E. Kleck, “Perceived gaze direction and the processing of facial displays of emotion”, Psychological Science (0956-7976), vol. 14, no. 6, pp. 644647, Nov. 2003.

[9] M. Schneider, L. Leuchs, M. Czisch, P. G. Smann, and V. I. Spoormaker, “Disentangling reward anticipation with simultaneous pupillometry / fMRI”, NeuroImage, vol. 178, pp. 1122, Sep. 2018.

[10] A. Franco, C. M. Neves, C. Quinto, R. Vigrio, and P. Vieira, “Singular Spectrum Analysis of Pupillometry Data. Identification of the Sympathetic and Parasympathetic Activity”, Procedia Technology, vol. 17, pp. 273280, Jan. 2014.

[11] E. H. Hess and J. M. Polt, “Pupil Size in Relation to Mental Activity during Simple Problem-Solving”, Science, vol. 143, no. 3611, pp. 11901192, 1964.

[12] D. Kahneman and J. Beatty, “Pupil Diameter and Load on Memory”, Science, vol. 154, no. 3756, pp. 15831585, 1966.

[13] R. H. Spector, “The Pupils”, in Clinical Methods: The History, Physical, and Laboratory Examinations, 3rd ed., H. K. Walker, W. D. Hall, and J. W. Hurst, Eds. Boston: Butterworths, 1990.

[14] D. Kahneman, W. S. Peavler, and L. Onuska, “Effects of verbalization and incentive on the pupil response to mental activity”, Canadian Journal of Psychology/Revue canadienne de psychologie, vol. 22, no. 3, pp. 186196, 1968.

[15] S. D. Kreibig, “Autonomic nervous system activity in emotion: A review”, Biological Psychology, vol. 84, no. 3, pp. 394421, Jul. 2010.

[16] E. H. Hess, “The role of pupil size in communication”, Scientific American, vol. 233, no. 5, pp. 110-119, Nov. 1975.

[17] P. Ricciardelli, L. Lugli, A. Pellicano, C. Iani, and R. Nicoletti, “Interactive effects between gaze direction and facial expression on attentional resources deployment: the task instruction and context matter”, Scientific Reports, vol. 6, p. 21706, Feb. 2016.

[18] E. Wood, T. Baltruaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, “Rendering of Eyes for Eye-Shape Registration and Gaze Estimation”, in 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 37563764.

[19] P. Ekman, “What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)”, Oxford University Press, 1997.

[20] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency, “OpenFace 2.0: Facial Behavior Analysis Toolkit”, in 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), 2018, pp. 5966.

[21] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions”, in 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp. 18.

[22] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A Multimodal Database for Affect Recognition and Implicit Tagging”, IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 4255, Jan. 2012.

[23] Y. Zhao, X. Wang, and E. M. Petriu, “Facial expression anlysis using eye gaze information”, in 2011 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA) Proceedings, 2011, pp. 14.

[24] C. Aracena, S. Basterrech, V. Snel, and J. Velsquez, “Neural Networks for Emotion Recognition Based on Eye Tracking Data”, in 2015 IEEE International Conference on Systems, Man, and Cybernetics, 2015, pp. 26322637.

[25] J. ODwyer, R. Flynn, and N. Murray, “Continuous affect prediction using eye gaze and speech”, in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017, pp. 20012007.

[26] G. Stratou and L. P. Morency, “MultiSense - Context-Aware Nonverbal Behavior Analysis Framework: A Psychological Distress Use Case”, IEEE Transactions on Affective Computing, vol. 8, no. 2, pp. 190203, Apr. 2017.

[27] S. Alghowinem et al., “Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors”, IEEE Transactions on Affective Computing, vol. 9, no. 4, pp. 478490, Oct. 2018.

[28] F. Ringeval et al., “AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge”, in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, New York, NY, USA, 2017, pp. 39.

[29] F. Weninger, “Introducing CURRENNT: The Munich Open-Source CUDA RecurREnt Neural Network Toolkit”, Journal of Machine Learning Research, vol. 16, pp. 547551, 2015.

[30] R Core Team, “R: A language and environment for statistical computing”, R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

[31] F. Ringeval et al., “Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data”, Pattern Recognition Letters, vol. 66, pp. 2230, Nov. 2015.

[32] M. Valstar et al., “AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge”, in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, New York, NY, USA, 2016, pp. 310.

[33] I. Daubechies, “The wavelet transform, time-frequency localization and signal analysis”, IEEE Transactions on Information Theory, vol. 36, no. 5, pp. 9611005, Sep. 1990.

[34] L. I.-K. Lin, “A Concordance Correlation Coefficient to Evaluate Reproducibility”, Biometrics, vol. 45, no. 1, pp. 255268, 1989.

[35] L. Lin, A. S. Hedayat, B. Sinha, and M. Yang, “Statistical Methods in Assessing Agreement: Models, Issues, and Tools”, Journal of the American Statistical Association, vol. 97, no. 457, pp. 257270, 2002.

[36] T. Cover and J. Thomas, “Elements of Information Theory”, New York: John Wiley, 1991.

[37] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, “Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks”, in Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, New York, NY, USA, 2015, pp. 7380.

[38] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, “Feature selection in multimodal continuous emotion prediction”, in 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2017, pp. 3037.

[39] F. Ringeval, M. Valstar, R. Cowie, and M. Pantic, Eds., “AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition”, in Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge, AVEC18, co-located with the 26th ACM International Conference on Multimedia, MM 2018, Seoul, Korea, 2018.

[40] F. Eyben et al., “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing”, IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190202, Apr. 2016.

[41] K. Brady et al., “Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction”, in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, New York, NY, USA, 2016, pp. 97104.