Speech, Head, and Eye-based Cues for Continuous Affect Prediction

Jonny O'Dwyer
Department of Computer & Software Engineering
Athlone Institute of Technology
Athlone, Ireland
j.odwyer@research.ait.ie

***Abstract*—Continuous affect prediction involves the discrete time-continuous regression of affect dimensions. Dimensions to be predicted often include arousal and valence. Continuous affect prediction researchers are now embracing multimodal model input. This provides motivation for researchers to investigate previously unexplored affective cues. Speech-based cues have traditionally received the most attention for affect prediction, however, non-verbal inputs have significant potential to increase the performance of affective computing systems and in addition, allow affect modelling in the absence of speech. However, non-verbal inputs that have received little attention for continuous affect prediction include eye and head-based cues. The eyes are involved in emotion displays and perception while head-based cues have been shown to contribute to emotion conveyance and perception. Additionally, these cues can be estimated non-invasively from video, using modern computer vision tools. This work exploits this gap by comprehensively investigating head and eye-based features and their combination with speech for continuous affect prediction. Hand-crafted, automatically generated and CNN-learned features from these modalities will be investigated for continuous affect prediction. The highest performing feature sets and feature set combinations will answer how effective these features are for the prediction of an individual's affective state.**

***Index Terms*—speech, head pose, eyes, affective computing, feature engineering**

I. Introduction

Affective computing can be thought of as the use of computers related to human or human-like feelings. Such use of computers can include human emotion or psychopathology recognition based on audio-video data, or synthesising human-like emotional speech for robotics applications. Speech-based affective computing is now well developed. There is nearly thirty years of research related to speech-based affective computing [1] and 66% of the world's native language speaking populations are represented by affective speech data sets [2]. Despite a large body of evidence linking eye and head-based cues to emotion and motivational state conveyance [3]–[12], the use of these cues is underdeveloped for affective computing purposes.

Based on the identified research opportunity, answering the following research question is the aim of this work: How significant an improvement in the prediction of an individual's affective state can be achieved by processing the combined cues gathered from an individual's speech, head and eyes? Therefore, in this paper, work towards finding appropriate features and showing the usefulness of head and eye-based cues estimated from video and combined with speech for continuous affect prediction is described. The layout of the remainder of this paper is as follows. Related work that informed and inspired this project is provided in Section II. This is followed by presentation of the methodology employed in Section III. Initial results are given along with brief discussion of these results in Section IV. Section V concludes this paper in the form of some final remarks and a future work plan.

II. Related Work

Within affective computing, continuous affect prediction involves the discrete time-continuous regression of the affective state of individuals. Features extracted from different modalities such as speech or facial expression are used as input to this process. Advantages of continuous affect prediction include the time-dependent nature of prediction and the complex N-dimensional affect representation provided. These advantages can allow temporal gradients of affect to be predicted and enable affect or emotion capture that may be outside that of human verbal description [13]. Common affect dimensions include arousal and valence, which may be plotted in a 2D circumplex model such as that proposed by Scholsberg [14] and later refined and demonstrated to resemble a cognitive structure of affect by Russell [15].

Speech-based affective computing is now a well developed field. There are corpora [13], [16]–[18], feature sets [18]–[21] tools [22] and repositories [23] to enable speech-based input for affective computing systems. Including speech in affective computing systems, is a good idea whenever possible, as the performance benefits of using this modality are well established, particularly for continuous arousal prediction for example. Additionally, if a baseline of human-level performance is to be considered, including speech in multimodal systems provides for fairer comparisons as humans clearly have access to both verbal and non-verbal output when forming their affect annotations. An interesting new direction in speech affective computing is end-to-end learning. End-to-end learning was performed in [24] with concordance correlation coefficient (CCC) scores of 0.686 for arousal and 0.261 for

This work was supported by the Irish Research Council (Grant Nos. GOIPG/2016/1572, GOIPG/2018/2030.

J. O'Dwyer, "Speech, Head, and Eye-based Cues for Continuous Affect Prediction," in 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), IEEE, Sep. 2019, pp. 16–20, isbn: 978-1-7281-3891-6. doi: 10.1109/ACIIW.2019.8925042

valence on the RECOLA [17] test set. The input features for [24] were learned directly from raw speech data using a convolutional neural network (CNN) algorithm and the feature vectors were passed to a bidirectional long short-term memory recurrent neural network (BLSTM-RNN) for continuous affect prediction.

Busso et al. [10] showed head pose and speech prosody to be strongly linked by objective measures in their work. They additionally carried out a subjective experiment that showed emotion perception changing in the presence of different head motion patterns. While in [11], the authors concluded that vocalists' head movements encode emotional information during speech and song, and that observers could identify emotion based on head movement alone in their subjective experiment. Objective and subjective experiments were carried out in [12] for the purpose of understanding how head motions contribute to the perception of emotion in an utterance. For the objective experiment, 45 dynamic features based on the discrete Fourier transform of yaw, pitch and roll angular head movements, gathered using a head-mounted device, along with a static measure of head pitch, were used as input to machine learning algorithms. The best results achieved were 94% for neutral, 79% for sad, 57% for happiness and 72% for angry classification. These studies clearly indicate that head motion is important for emotion signalling and discrimination, despite this, no studies exist which have investigated head-based features for continuous affect prediction.

Eye-based cues have been linked with various emotional displays for a long time [25] and numerous works suggest various eye signals to elucidate one's affective state. The shared signal hypothesis [5], [6] suggests that eye gaze is congruent with emotion expressions when the gaze direction matches the underlying motivation to approach or avoid stimuli while pupil size variation occurs during monetary reward or penalty [26], while viewing emotionally arousing stimuli [27], [28], during reward expectation [29] and during autonomic nervous system stimulation [30]. In affective computing, eye-based cues have been employed for emotion recognition [31], [32]. Decision tree neural network was used in [32] .affect-level recognition [16] and psychopathology applications [34], [35]. Soleymani et al. [16] achieved good performance using eye features comprised of statistics and power spectral density calculations based on eye gaze, eye blink and pupillometry low-level descriptors (LLDs) as input to Support Vector Machine (SVM). The eye-based features performed best for unimodal affect-level recognition when compared to compared to other physiological measures in their experiments. However, specialised equipment is required for gathering the proposed features in [16], in the form of an external Tobii eye gaze recorder, which provides a barrier to researchers in accepting and enhancing the proposed features/approaches.

Following the review of related work it is clear that there is evidence supporting the use of head and eye-based cues as inputs to affective computing systems. Furthermore, modern computer vision tools can estimate these cues from video [33], [35] which can serve to maximise the impact of these cues for affective computing, if their usefullness can be shown. This serves as motivation for this work. It is hypothesised that head and eye-based cues, estimated from video, can improve the performance of affective computing systems, providing performance benefits when combined with speech and providing affect prediction capabilities in the absence of speech. Therefore, the technical scope of this work includes comprehensively investigating these cues combined with speech for continuous affect prediction.

III. Methodology

The methodology employed for this work includes gathering LLDs from video followed by BLSTM-RNN model creation and evaluation based on proposed features and feature combinations input. As part of the feature engineering process, some LLDs are captured from raw head pose, eye-based or visual features, while others must be calculated based on differences in the raw data measurement. For example, pupil dilation is calculated as being equal to 1 based on whether the pupil diameter measurement gathered using OpenFace is larger for a frame when compared to an immediately preceding frame measurement whereas a feature such as the raw eye gaze x coordinate is taken directly as a LLD. The LLD extraction processes is then followed by higher-level features extraction, which are gathered under varying temporal feature windows. This is followed by altering the now commonly accepted affect learning parameter, ground-truth backward time-shift, on the validation set in order to provide better performance of extracted features by taking human annotator lag in providing ratings into account. Confirmatory analysis of features' or feature sets' efficacies are provided by way of test set evaluation. Further details on the methodology are given in this section.

A. Corpus and Training, Validation and Test Partitioning

The RECOLA [17] 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 and communicating in French. Arousal and valence annotations, ranging from -1.0 to +1.0, are provided with the set in discrete-continuous-time at a rate of 25 values per second. Each recording in the set is 5 minutes in length. 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 [32]. 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].

B. Gathering LLDs

LLDs comprised of raw head pose and eye-based data are gathered using OpenFace 2.0.6 [35]. The initial set of raw head-based data includes head location x, y and z in camera coordinate millimeters, and head rotation yaw, pitch and roll in world coordinate radians with camera origin. The initial

set of eye-based data includes pupil diameter, eye gaze x, y radians, eye gaze distance, logical eye blink/closure and eye blink intensity. All of the eye-based raw data are based on world coordinates except eye gaze distance, the one camera coordinate LLD. Briefly, world coordinates are independent of the camera whereas camera coordinates depend on camera location.

A number of additional LLD features are calculated on a frame-wise basis to capture more detailed information on low-level feature dynamics. These features include: displacement (deltas) of the all head features and eye gaze x, y radians, and binary true/false features for eye fixation, eye gaze approach, direct gaze, pupil dilation and pupil constriction. All of the additional LLDs are calculated using software resulting from this research while the direct gaze features were annotated by a human observer who judged OpenFace output frames as either direct gaze = 1 (looking at the screen/camera) or 0 (averted gaze away from the screen/camera) as in Fig. 1.

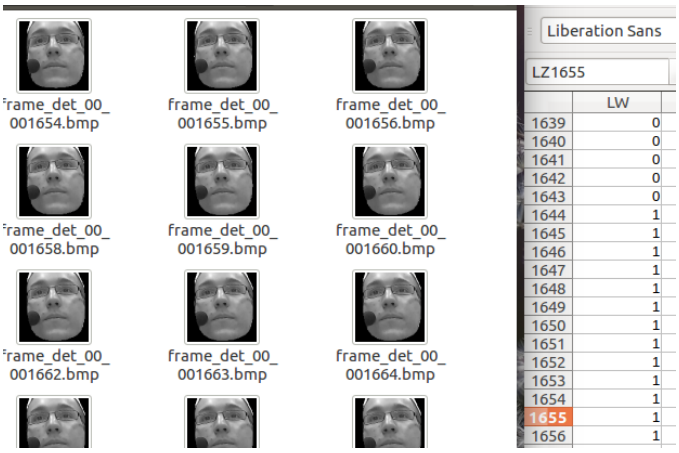

Fig. 1: Screen capture of human annotation of direct eye gaze.

C. High-level Feature Extraction

High-level features are extracted using 4, 6 and 8 second temporal windows, moved forward at a rate of 1 frame per interval. Each temporal window is tested using each modality, or combination of modalities, as input to the BLSTM-RNN for training and validation. The high-level features are sometimes preceded by mid-level feature extraction in the form of scale and detail wavelet coefficient features, where 10-order Daubechies wavelet [36] features are extracted using discrete wavelet transform for as many decomposition levels as possible for a given temporal window. The calculations for the final high-level features include, where appropriate: ratio, time in seconds[min, mean, max, total], min, max, mean, median, quartile 1, quartile 3, skewness, kurtosis, standard deviation, numerous inter-quartile range measurements, linear regression slope, linear regression intercept, RMS, and zero-crossing rate. The high-level features are calculated from static-time (one frame of data) and dynamic-time (frame-wise lagged difference data) measurements of LLDs and the mid-level time-frequency wavelet features for each temporal feature sample window to enhance the data prior to machine learning. Temporal feature windows are denoted W_s for the remainder of this work, where s indicates the window size in seconds per interval. After gathering feature sets, exploratory data analysis is carried out using feature-to-target relationship calculations.

D. Ground-truth Backward Time-shift

Following from recent works [37], [38], ground-truth annotations provided with the RECOLA [17] corpus are shifted back-in-time to account for annotator rating time delay. The ground-truth backward time-shift sizes for the experiments range from 0 to 4.4 seconds in steps of 0.2 seconds. These are referred to as D_s for the remainder of this work, where s indicates the delay in seconds applied to ground-truth annotations prior to concatenation with input features.

E. Feature Selection

In order to try to achieve the best features sets from the modalities, feature selection is applied in order to remove redundant or weak features inadvertently generated during early feature engineering. The feature selection approach taken in this work follows 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" [40, p.18] and it provides information on the nonlinear relationship between input features and target variables. Features under MI thresholds of 0.1, 0.15 or 0.2 are removed from features sets for experimental evaluation as these features are deemed independent of arousal or valence and therefore poor predictors.

F. BLSTM-RNN Training and Evaluation

BLSTM-RNN is used in this work to train models for feature set appraisal. The training method largely follows that of Ringeval et al. [40]. Single-task models are trained using BLSTM-RNN with 2 hidden layers, each with 40 and 30 nodes respectively, with a sum-of-squared-errors (SSE) objective function using the CUDA RecurREnt Neural Network Toolkit [41]. All input features and regression targets are standardised using the parameters mean and standard deviation, computed on the training set. The network learning rates are set at 10^{-5} and a random seed of 1787452436 is used throughout the experiments. Gaussian noise with a standard deviation 0.1 is added to all input features prior to training. BLSTM-RNN models are trained for a maximum of 100 epochs, however, early stopping is employed where training is stopped when no performance increase (lower SSE) is observed on the validation set after 10 epochs.

Following the training phase, network models are evaluated and selected using validation set CCC [42], where higher CCC is better, which is acheived using a forward-pass of the validation set data into the trained network. 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. Models that perform best during validation set experiments are selected for a final test set pass.

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

IV. Initial Results and Discussion

A. Speech, Head Pose and Eye Gaze Affect Prediction

This experiment included speech, head pose and eye gaze modalities. New head pose and eye gaze feature sets were generated and these were compared against and combined with speech for continuous affect prediction on the RECOLA [17] corpus. The head pose and eye gaze feature sets' extraction scripts for this experiment are made available in an accompanying GitHub repository [1]. During initial LLD feature exploration, the strongest correlations with both arousal and valence resulted from averages of *x* head location (arousal *r* = -0.26, valence *r* = -0.149) for head pose and *x* gaze angle (arousal *r* = 0.273, valence *r* = 0.245) for eye gaze, both using W_s = 8.

The results for the top performing modalities and systems from this experiment are given in Table I. The top performing arousal system included speech & head pose while the top performing valence system included speech & head pose & eye gaze. There was always a performance increase observed when head pose and/or eye gaze was added to speech for continuous valence prediction and unimodal head pose was shown to perform well for arousal prediction from the non-verbal channel (validation set CCC = 0.535). This experiment demonstrated the early promise of the proposed features as performance improvements above that of unimodal speech were shown to be possible with the proposed sets.

TABLE I: BLSTM-RNN Results For Top Performing Systems

System (Evaluation)	Arousal		Valence	
	SSE	CCC	SSE	CCC
Speech & Head Pose (Validation)	0.152	**0.771**	0.352	0.418
Speech & Head Pose & Eye Gaze (Validation)	0.177	0.744	0.355	**0.444**
Top Performing Validation Systems (Test)	-	0.779	-	0.326

B. Eye-based Continuous Affect Prediction

Based on the initial experiment, it was considered that the performance of the eye gaze feature set could be improved with further research. Therefore, a logical next step in this work dealt with further exploration of eye-based cues in order to more fully explore this modality for continuous affect prediction. This experiment was intended to provide further evidence for the use or omission of eye-based input for continuous affect prediction. Pupillometry and direct gaze knowledge measures in addition to eye gaze features previously investigated were evaluated combined and compared with speech on the RECOLA [17] corpus. An interesting result for the exploration of the final eye-based feature sets for both arousal and valence from these experiments includes the consistent removal of pupil dilation and constriction during feature selection. Also of note is the consistent retention of direct gaze-based features, namely, direct gaze ratio and time in seconds [mean, max, total], across both arousal and valence feature sets after selection.

The final results for this experiment can be seen in Table II. The results show that eye-based cues, considered with speech, provide performance benefits for continuous affect prediction. This result is an improvement on the previous experiment incorporating eye-based cues in the form of gaze as these results show that there can be a benefit of include eye-based cues with speech for arousal prediction. Unfortunately, adding the eye-based cues to speech for valence prediction did not outperform unimodal speech which indicates that these cues on their own, estimated form video, are ineffective when considered with speech for continuous valence prediction.

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

System (Evaluation)	Arousal		Valence	
	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	-	-

V. Conclusions and Future Work

The contributions of this work thus far include hand-crafted feature sets based on head and eye-based cues. Software to extract these sets are provided publicy, for use by the affective computing community. Good results have been achieved for unimodal head pose, and head pose and/or eye-based cues when combined with speech, which shows the benefits of considering these features. A tentative future work plan involves generating speech, head and eye-based cues using automated[2] and CNN-based feature generation on the raw data in order to ensure full exploration of these modalities. These automatically generated/learned features sets will be compared against and combined with the hand-crafted feature sets. Feature selection and fusion will be further investigated, in order to find the optimal feature sets and fusion strategy for said cues. Evaluation of the features on the SEMAINE [13] data set is also planned to assess feature generalisability. This work will provide the affective computing community with new knowledge and tools for feature extraction from video sequences that can provide for transparency of future results, ease of use for researchers and provide ground-work towards shared standard feature sets in a fashion similar to that of speech [21].

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

[2] http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/index.html

REFERENCES

[1] R. Fernandez and R. Picard, "Recognizing affect from speech prosody using hierarchical graphical models", Speech Communication, vol. 53, no. 9, pp. 1088–1103, Nov. 2011.

[2] S. M. Feraru, D. Schuller, and B. Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies", in 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 2015, pp. 125–131.

[3] E. H. Hess and J. M. Polt, "Pupil size as related to interest value of visual stimuli", Science, vol. 132, pp. 349–350, 1960.

[4] J. M. Polt and E. H. Hess, "Changes in pupil size to visually presented words", Psychonomic Science, vol. 12, no. 8, pp. 389–390, 1968.

[5] 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. 3–11, Mar. 2005.

[6] 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. 644–647, Nov. 2003.

[7] M. Schneider, L. Leuchs, M. Czisch, P. G. Sämann, and V. I. Spoormaker, "Disentangling reward anticipation with simultaneous pupillometry / fMRI", NeuroImage, vol. 178, pp. 11–22, Sep. 2018.

[8] A. Franco, C. M. Neves, C. Quintão, R. Vigário, and P. Vieira, "Singular Spectrum Analysis of Pupillometry Data. Identification of the Sympathetic and Parasympathetic Activity", Procedia Technology, vol. 17, pp. 273–280, Jan. 2014.

[9] 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.

[10] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis", IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 3, pp. 1075–1086, Mar. 2007.

[11] S. R. Livingstone and C. Palmer, "Head movements encode emotions during speech and song", Emotion, vol. 16, no. 3, pp. 365–380, Apr. 2016.

[12] Y. Ding, L. Shi, and Z. Deng, "Low-level Characterization of Expressive Head Motion through Frequency Domain Analysis", IEEE Transactions on Affective Computing, pp. 1–1, 2018.

[13] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent", IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 5–17, Jan. 2012.

[14] H. Schlosberg, "The description of facial expressions in terms of two dimensions", Journal of Experimental Psychology, vol. 44, no. 4, pp. 229–237, Oct. 1952.

[15] J. A. Russell, "A circumplex model of affect", Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[16] 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. 42–55, Jan. 2012.

[17] 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. 1–8.

[18] M. Valstar et al., "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge", in Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, New York, NY, USA, 2014, pp. 3–10.

[19] B. Schuller et al., "The interspeech 2010 paralinguistic challenge", in Eleventh Annual Conference of the International Speech Communication Association, pages 2794–2797, 2010b.

[20] B. Schuller et al., "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity and Native Language", 2016, pp. 2001–2005.

[21] 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. 190–202, Apr. 2016.

[22] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor", in Proceedings of the 21st ACM International Conference on Multimedia, New York, NY, USA, 2013, pp. 835–838.

[23] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - 2014; A collaborative voice analysis repository for speech technologies", in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 960–964.

[24] G. Trigeorgis et al., 'Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network', in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5200–5204.

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

[26] 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. 186–196, 1968.

[27] E. H. Hess and J. M. Polt, "Pupil size as related to interest value of visual stimuli", Science, vol. 132, pp. 349–350, 1960.

[28] J. M. Polt and E. H. Hess, "Changes in pupil size to visually presented words", Psychonomic Science, vol. 12, no. 8, pp. 389–390, 1968.

[29] M. Schneider, L. Leuchs, M. Czisch, P. G. Sämann, and V. I. Spoormaker, "Disentangling reward anticipation with simultaneous pupillometry / fMRI", NeuroImage, vol. 178, pp. 11–22, Sep. 2018.

[30] A. Franco, C. M. Neves, C. Quintão, R. Vigário, and P. Vieira, "Singular Spectrum Analysis of Pupillometry Data. Identification of the Sympathetic and Parasympathetic Activity", Procedia Technology, vol. 17, pp. 273–280, Jan. 2014.

[31] 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. 1–4.

[32] C. Aracena, S. Basterrech, V. Snáel, and J. Velásquez, "Neural Networks for Emotion Recognition Based on Eye Tracking Data", in 2015 IEEE International Conference on Systems, Man, and Cybernetics, 2015, pp. 2632–2637.

[33] 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. 190–203, Apr. 2017.

[34] 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. 478–490, Oct. 2018.

[35] 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. 59–66.

[36] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis", IEEE Transactions on Information Theory, vol. 36, no. 5, pp. 961–1005, Sep. 1990.

[37] 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. 3–10.

[38] 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. 73–80.

[39] T. Cover and J. Thomas, "Elements of Information Theory", New York: John Wiley, 1991.

[40] F. Ringeval et al., "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data", Pattern Recognition Letters, vol. 66, pp. 22–30, Nov. 2015.

[41] F. Weninger, "Introducing CURRENNT: The Munich Open-Source CUDA RecurREnt Neural Network Toolkit", Journal of Machine Learning Research, vol. 16, pp. 547–551, 2015.

[42] L. I.-K. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility", Biometrics, vol. 45, no. 1, pp. 255–268, 1989.