Decoupling Temporal Dynamics for Naturalistic Affect Recognition in a Two-Stage Regression Framework

Yona Falinie A. Gaus, Hongying Meng, Asim Jan
Department of Electronic and Computer Engineering, Brunel University London,
Email: yonafalinie.abdgaus@brunel.ac.uk, Hongying.Meng@brunel.ac.uk, Asim.Jan@brunel.ac.uk

***Abstract*—Automatic continuous affect recognition from multiple modalities is one of the most active research areas in affective computing. In addressing this regression problem, the advantages of a model, such as Support Vector Regression (SVR), or a model that can capture temporal dependencies within a predefined time window, such as Time Delay Neural Network (TDNN), Long Short-Term Memory (LSTM) or Kalman Filter (KF), have been frequently explored, but in an isolated way. The motivation is towards decoupling temporal information from its features at the semantic level, in order to exploit the slow-changing emotional property at decision level. This paper explores and proposes 2-stage regression framework where SVR, that has been regarded as the baseline approach on affective recognition task, is concatenated together with subsequent models. Extensive experiments have been carried out on a naturalistic emotion dataset, using eight modalities present in RECOLA database. The results shows the proposed framework can capture temporal information at the prediction level, and outperform state-of-the-art approaches in continuous affective recognition.**

I. Introduction

In recent years, considerable amount of research in automatic affect recognition has undergone to enable natural, intuitive and friendly human-machine interaction. Early works have focused on the recognition of basic discrete emotion, with the data being collected in a laboratory setting, where speakers act in specific emotional states [1], [2]. Recently, a considerable amount of literature focus on naturalistic emotional behaviours in continuous dimensions (e. g., arousal and valence) [3], [4], One possible explanation is that a single label may not reflect the complexity of the affective state conveyed by such a rich source of information. Hence a number of research areas have started to model, analyze and interpret the continuity of affective behaviors in terms of latent dimensions, rather than discrete emotion categories. With the advancement of devices such as the camera and microphone, multimodality has been widely implemented for emotion recognition [5]. The combination of various modalities can be more useful for identifying and classifying emotions, which can boost emotion accuracy, since each modality can provide complementary information [6], [7].

In order to learn the relationship between the feature from various modalities and the multi-dimensional affective space, a variety of machine learning models have been investigated, such as k-Nearest Neighbor [8], continuous conditional random fields (CCRF) [9] and Relevance Vector Machine (RVM) [10]. Support Vector Regression (SVR) is a popular technique that has been frequently employed and is regarded as the baseline regression approach for many continuous affective computing tasks [11], [7]. Recently, Long Short-Term Memory Recurrent Neural Networks (LSTM- RNN) [12] has become one of the state-of-the-art modeling technique in continuous emotion recognition, thanks to its ability to incorporate knowledge about how emotions typically evolve over time so that the inferred emotion estimates are produced under consideration of an optimal amount of context [13]. LSTM-RNN was first applied on acoustic features [14], and consecutively it has been successfully employed for other modalities (e. g., video, and physiological signals) [15], [11].

There is a large volume of published studies covering wide range of machine learning techniques for continuous affect recognition [11]. However the methods discussed above make use of modeling techniques that are still tied at the features level. The aspect of decision level has been an uncommon approach when compared to feature level. The approach employed in [16] tackles this issue by proposing multistage classification by taking into account temporal information at the decision level. Motivated by this initial promising results, this paper plans to extend this approach by capturing the temporal relationship at decision level by proposing a two-stage regression framework for automatic continuous emotion recognition from given modalities. This method will try decoupling the modeling of the temporal dynamics of an emotional state from the high variability of the feature level by using a two-stage regression framework. In this approach, the SVR model, which captures the strength of the features, represents the initial prediction. This further progresses as inputs for regression analysis in a subsequent model. By using this approach, it will allow the network to easily exploit the slow changing dynamic between emotional state.

The remainder of the paper is organized as follows: Section 2 discusses the related works; Section 3 presents the methodology of two-stage regression framework; Section 4 describes the databases and its corresponding features; Section 5 offers an experimental evaluation towards proposed approach; and finally, Section 6 concludes this work and discusses potential avenues for future work.

II. Related Works

Automatic naturalistic affect recognition typically comprises of two systems: feature extraction, which provides a low-level representation from given modality such as audio, visual and/or physiological recordings; and modeling approaches that translate the low-level descriptors into high-level affect-related features.

Then, all features will be compose to form a feature vector, which in turn feeds a regular classifier or a regressor. As a result, given a sufficient set of features of the time series, it permits any classical supervised learning algorithm (linear classifier, k-NN, Support Vector Machine, etc) to be applied. The most widely used regression method, which become the baseline for continuous affect is SVR [11] [7]. Other than SVR, Relevance Vector Regression (RVR), or linear regression models [18] are effective in predicting naturalistic affect.

However, a major disadvantages of feeding a set of descriptive features into regular classifier is that, even though features summarizing the whole sequence can provide a meaningful description for the purpose of classification/regression, the temporal structure of the sequence (of emotions) is neglected and not being taken into consideration. For example, a randomly shuffled version of a sequence would result in the same statistical feature representation, but it would correspond to a completely different temporal pattern. These regression models are effective in predicting continuous affect, but are insufficient in capturing the temporal information of the affective dimensions. To overcome this limitation, LSTM-RNN has been successfully applied to continuous emotion prediction [11], since it has the ability to incorporate knowledge about how emotions typically evolve over time so that the inferred emotion estimates are produced under consideration of an optimal amount of context [13] Apart from LSTM-RNN, continuous state-model approaches such as Kalman Filter (KF) [19] is employed to allow a deeper insight into emotional prediction. In [20] the authors leverage a KF based approach by treating emotional state (x) as a function of time of the features information (z) from the respective modality using the standard state space framework. This model is being used to fuse each of the modality/sensor measurement per time step, and at the same time it models the time-varying nature of the model to further improve system performance. In [21], good performance were achieved simply by modeling acoustic feature from music using KF. A possible rationale behind the success cases of KF in affect recognition is that its ability to propagate the emotion predictions mean and covariance of the current state in time. On the other hand, only a few parameters are needed to be estimated with a small number of observations for KF modeling. A somewhat less explored NN method is Time Delay Neural Network (TDNN) [22]. TDNN has the ability to capture dynamic relationship between consecutive observations, due to its delay property in the TDN nodes. Emotional expressions which occur in a particular moment are classified by taking into account not only the input features describing that moment, but also the input features describe the moment before. This sentence is parallel with [14], whose amongst the first to apply LSTM-RNN in continuous affect recognition, where emotions are typically evolve over time. The delay property in TDNN nodes can be set as the number of past instance of emotional expression, making it a perfect fit for a modeling continuous emotion recognition. However, only [23] fully exploits TDNN structure when modeling the temporal relationship at the semantic level. For each and every regression model stated above, despite having the ability to learn useful features directly from each and every modality, they completely ignore temporal information present in continuous affect recognition. Therefore, in this paper, a two-stage regression framework is proposed to try decoupling emotional state dynamics by modeling using the emotional state prediction step based on the input features. The first stage regression model, which is based on SVR, generates the original training prediction based on training a feature vector. Then, the training prediction will become an input to the subsequent model, in order to learn the temporal information on decision label. The subsequent model, which consists of TDNN, LSTM and KF is chosen because it has the ability to incorporate the temporal information of the decision labels, and propagate this information from one time point to the next (see arrow in second-stage regression). In second stage, it will learn the expected prediction using the development prediction from the first stage regression approach.

III. Methodology

The proposed two-stage regression framework for continuous affect prediction is depicted in Fig. 1. SVR is being chosen as the first regression model, generates the initial prediction $\hat{x_n}$ based on the feature vector from each modality. Then, subsequent model is trained with initial prediction $\hat{x_n}$ to learn the expected prediction x_n. To implement the two-stage regression framework, SVR and subsequent model are trained in order. In other words, subsequent model takes the predictive ability of SVR in account for training. These two-stage models are both built using the same training dataset. Once the two-stage models have been trained, it can be treated as two layer system where the expected prediction x_t is produced continuously once a unit of sequences of features are received.

A. First-stage regression model

1) Support Vector Regression (SVR): SVR has been chosen as model to perform first stage regression task, by fully utilizing liblinear library [24]. SVR is the most prominent method in the context of machine learning, particularly in continuous emotion recognition[11] [7]. In this experiment, the hyperparameter is chosen empirically to balance the emphasis on the error and generalization performance. A more in-depth explanation of the SVR is in [25].

B. Subsequent model : Second-stage regression model

1) Time Delay Neural Network (TDNN): Given TDNN network has *M* inputs $(x^1(t), x^2(t), ..., x^M(t))$ and one output

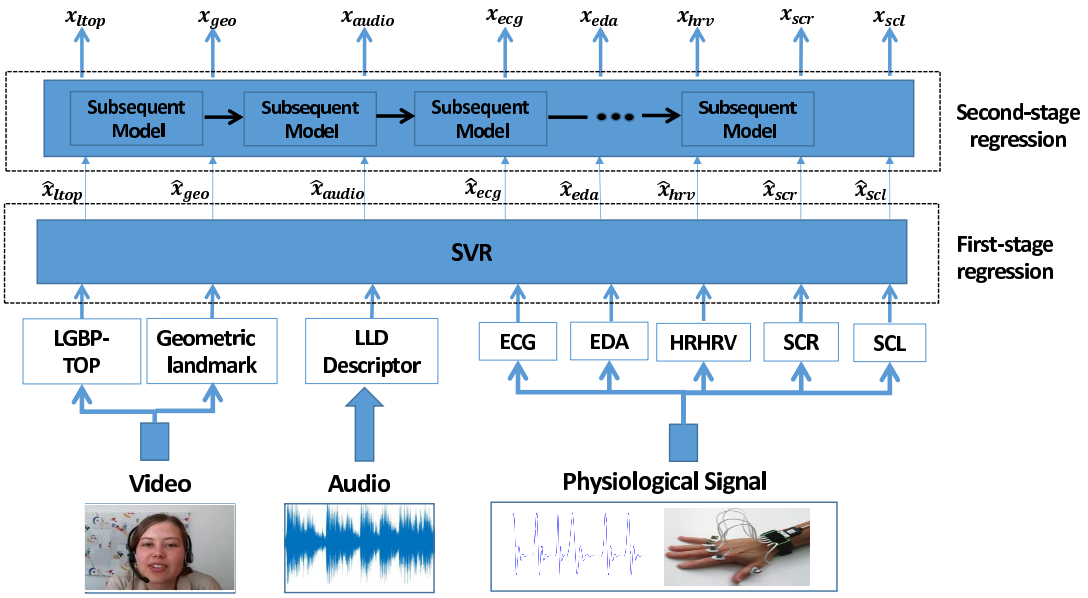

Fig. 1: Architecture of Two-stage Regression Modeling of Naturalistic Affect Recognition using RECOLA dataset [7]

$(z(t))$ where these inputs are time series with time step t. At each input $x^i(t)$, given $i = 1, 2, ..., M$ each of these input has bias, b, delays N, and unknown weights w. N delays, indicate as $D_1^i, D_2^i, .., D_N^M$ act as a function of storing the previous input $I^i(t-d)$ with $d = 1, 2, .., N$ and unknown weights ,$w = w_{i1}, w_{i2}, .., w_{iN}$. In this experiments, sigmoid function (σ) is being chosen as transfer function due to its convenient of mathematical properties. A single TDNN node can be represented as in Equation 1

$$z(t) = f \sum_{i=1}^{M} [\sum_{d=0}^{N} x^i(t-d) * w_i d + b_i] \tag{1}$$

As previously known, features in affective dimensions may represent different information at a different time point. When it is fed into a traditional neural network, the neural network may only treat them as different information in different neurons, without taking the accountability that the emotions may have dynamic relationships between each consecutive time sample. In TDNN, the presence of delay tackle this matter, since it has the ability to relate and compare current input to the past history of events. In Equation 1, both the input at current time steps t and previous time step $t-d$ contribute to the overall outcome of the neuron.

2) Long Short Term Memory-Recurrent Neural Network (LSTM-RNN): An LSTM is consist of a series of cells, each of which has internal state (c_t) that is updated based on the current input (x_t) and previous cell state (c_{t-1}). Using gates, the network then determines how much the previous cell state and current input contribute to the new cell. The forget gate f_t calculates the value between 0 and 1 using a sigmoid function (σ), which detemines the contribution of the previous cell state (c_{t-1}). The input gate (i_t) performs same operation on current input (x_t). The equation for these operation are as shown below:

$$i_t = \sigma(w_{xi}x_t + w_{hi}h_{t-1} + w_{ci}c_{t-1} + b_i) \tag{2}$$
$$f_t = \sigma(w_{xf}x_t + w_{hf}h_{t-1} + w_{cf}c_{t-1} + b_f) \tag{3}$$
$$\tilde{c_t} = tanh(w_{xc}x_t + w_{hc}h_{t-1} + b_c) \tag{4}$$
$$c_t = f_t * c_{t-1} + i_t * \tilde{c_t} \tag{5}$$

The model then uses the cell state (c_t) to compute its output representation for time t, that is h_t. The current cell state contribution is determined by an output gate o_t

$$o_t = \sigma(w_{xo}x_t + w_{ho}h_{t-1} + w_{co}c_t + b_o) \tag{6}$$
$$z_t = o_t * tanh(c_t) \tag{7}$$

3) Kalman Filter (KF): The Kalman Filter approach is used to exploit the time series nature of the emotional label data by estimating the emotional state as a function of time (x_t), from the respective modality using the standard state-space framework. In this paper, the prediction of naturalistic affect trajectory is considered as a time series and apply methods for time series. It is assumed that continuous affect trajectory can be modeled by the state transition equation:

$$x_{t+1} = F_t x_t + w_t \tag{8}$$

where F_t is state transition matrix, x_t is the state of time t and w_t is process noise. To further up the assumption, the observation of the state can be made through a measurement system which can be represented by a linear equation in the form:

$$z_t = H_t x_t + v_t \tag{9}$$

where z_t is the observation or the measurement made at time t. x_t is the state of time t, H_t is the observation matrix and v_t is additive measurement noise.

To simplify the equation, F_t and H_t is being assumed as constant 1, $F_t = 1$; $H_t = 1$, the process and measurement noise random processes, w_t and v_t are uncorrelated, zero-mean, white noise processes with known covariance matrices, $w_t \sim N(0, Q_t)$ and $v_t \sim N(0, R_t)$. At each lookback window N, a new estimation of the process covariance Q_t and noise covariance R_t is produced using Equation below

$$Q_t = cov(x_{t=2,N} - F x_{t=1,N-1}) \tag{10}$$

$$R_t = cov(z_{t=1,N} - H x_{t=1,N}) \tag{11}$$

Noted that, Q_t and R_t is produced from training prediction, therefore it can be incorporated as $predict$ and $update$ of the Kalman Filter iteration, where the input is the continuous affect prediction from development data of 1-stage approach.

IV. DATASET AND FEATURES

In this experiment - RECOLA [26] has been adopted as standard dataset for the Audio-Visual Emotion Challenge in 2015/2016 [7] [11]. In this database, it was recorded in in spontaneous interaction mode, during resolving of a collaborative task remotely through video conference. The corpus consists of multimodal data, such as; audio, video, and physiological signal. The data is labeled in two affective dimensions, namely arousal and valence, and was manually annotated using a slider-based label tool. A combination of these individual ratings is used as gold standard label.

A. Audio Features

Audio features are computed with openSMILE [27] and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [28]. LLD descriptors cover spectral, cepstral, prosodic and voice quality information are extracted. Overall, the acoustic baseline features set contains 88 features.

B. Video Features

The RECOLA dataset provides a set of video features consisting of appearance features, namely Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [29], as well as geometric features computed from 49 landmarks tracked in the video sequences using a Supervised Descent Method (SDM) [30].

LGBP-TOP is a dynamic appearance descriptor that is robust to illumination changes and misalignment [29]. Principal Component Analysis (PCA) is applied, obtained 84 features representing 98% of the total variance. As for video geometric features, 49 facial landmark were tracked.

C. Physiological Signal

The physiology features which consists of ECG signals, EDA signals, HRHRV signals, SCR and SCL signals were adopted exactly as in [7], with a features size depends on each modality respectively.

The first physiological features is ECG, recorded the electrical activity of the heart. In this signal, it consists of total of 19 features, zero-crossing rate, the four first statistical moments, the normalized length density, the non-stationary index, the spectral entropy, slope, mean frequency plus 6 spectral coefficients, the power in low frequency (LF, 0.04-0.15Hz), high frequency (HF, 0.15-0.4Hz) and the LF/HF power ratio.

The second physiological features is EDA, derived from the property of the human body that causes continuous variation in the electrical characteristics of the skin. From this signal, 8 features has been extracted including the four first statistical moments from the original time-series and its first order derivative.

The third physiological features is HRHRV, which is extracted from the heart rate and its measure of variability. Its being derived from filtered ECG signal by applying zero-delay bandpass filter (3-27Hz) on the signal. 10 features has been extracted including the two first statistical moments, the arithmetic mean of rising and falling slope, and the percentage of rising values for each of those two descriptors.

The fourth physiological features is SCR. SCR is the phenomenon that the skin momentarily becomes a better conductor of electricity when either external or internal stimuli occur that are physiologically arousing. 8 features were extracted, including the four first statistical moments from the original time-series and its first order derivative.

The fifth physiological features is SCL, where its directly controlled by the sympathetic nervous system and indicates the activity of the sweat glands in the skin. 8 features were extracted, including the four first statistical moments from the original time-series and its first order derivative.

V. EXPERIMENTAL RESULTS

Separate continuous affect recognition, arousal and valence predictions are obtained from individual modalities as described in the AVEC2016 paper [7]. The regression task is performed using linear SVR provided with the liblinear library [24] for the continuous affect recognition in the unimodal setting. Then, second stage prediction is incorporated using subsequent model, i.e, TDNN, LSTM-RNN and KF, as shown in Fig. 1.

The results of the proposed approach is reported in terms of *Concordance Correlation Coefficient* [11] metric:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{12}$$

where ρ is the *Pearsons Correlation Coefficient* between two time series (e.g: prediction and gold-standard); μ_x and μ_y are the means of each time series; and σ_x^2 and σ_y^2 are the corresponding variance. Thus, CCC fits better in continuous

affect recognition and its being utilized as an official scoring metric for the last two edition of AVEC challenge.

The CCC values for unimodal performance in predicting arousal and valence separately are shown in Table I. Results show comparable unimodal performance in audio modality, where no significant improvement in performance is observed here due to the audio recordings being mostly clean speech. Second-stage regression approach perform somewhat better on video appearance, video geometric and physiological signal. As it can be seen, the second-stage regression set-ups either match or outperform their corresponding individual baseline SVR models in majority of the cases. The improvement could be due to the fact that the baseline initial models such as SVR prediction actually helps the subsequent model to avoid its local minima.

In the arousal dimension, closer inspection of Table I indicates that the second-stage regression approach predominantly gives higher performance compared to the baseline in all modalities except audio. More specific, in each subsequent model, SVR-TDNN gives higher performance compared to SVR-LSTM and SVR-KF, in video-appearance, video-geometric, ECG and HRHRV modality. These results further support the idea of TDNN, where the model can capture dynamic behaviour between consecutive affective states through delay. It is noted that the delay is determined by experimentally setting the parameter. Detailed analysis of the number of delay chosen would be the part of future work. However, in contrast with earlier statements, SVR-KF gives superior results on EDA, SCL and SCR modality, compared with baseline, SVR-TDNN and SVR-LSTM. While the noisy observations or measurements in this context are the unimodal predictions acquired in the 1-stage approach, KF is able to model the process and measurement noise as a Gaussian efficiently, thus giving the optimal solution towards the performance, compared with the other two subsequent model.

Similarly, the second-stage approach brought additional performance improvements when compared to baseline results on the valence dimension, although not as obvious as for the arousal dimension. From Table I, it shows that video geometric modality appear more informative than audio modality. SVR-KF is the only models that contribute to the higher results across video-appearance, video-geometric, HRHRV, SCL and SCR modality, but not as significance as in arousal dimensions.

The relatively lower performance of predicting valence when compared to arousal dimensions in second-stage approach is likely due to the subsequent model which cannot easily capture the temporal aspect of the emotional expressions, when mapping it to the valence dimensions. Taking a deeper look into its gold standard, it appears that, with only small time interval, valence dimensional has a lot of start, peak, and end points, making its harder for the model to capture temporal dynamics, and leads to lower performance of improvement when compared to arousal dimensions. It is possible to hypothesise that having a high level feature extraction method can possibly lead to further improvements of the results, but it is outside the scope of this paper. The future works would involve additional experiments about high level features which can verify these assumptions.

TABLE I: Comparison of baseline results with two-stage regression approach of unimodal performance on development sets

Modality	Arousal			
	Baseline-SVR [7]	SVR−TDNN	SVR−LSTM	SVR−KF
D-Audio	**0.796**	0.794	0.769	0.780
D-VA	0.483	0.519	0.504	**0.524**
D-VG	0.379	**0.430**	0.402	0.415
D-ECG	0.288	**0.356**	0.325	0.337
D-HRHRV	0.382	**0.427**	0.417	0.423
D-EDA	0.077	0.125	0.130	**0.163**
D-SCL	0.101	0.105	0.161	**0.247**
D-SCR	0.071	0.157	0.175	**0.214**
Modality	Valence			
	Baseline-SVR [7]	SVR−TDNN	SVR−LSTM	SVR−KF
D-Audio	**0.455**	0.381	0.428	0.432
D-VA	0.474	0.462	0.467	**0.481**
D-VG	0.612	0.624	0.630	**0.638**
D-ECG	0.153	0.130	0.153	**0.153**
D-HRHRV	0.293	0.274	0.293	**0.298**
D-EDA	0.104	0.188	0.194	**0.197**
D-SCL	0.124	0.156	0.166	**0.170**
D-SCR	**0.110**	0.066	0.085	0.086

To further highlight advantages of 2-stage regression architecture, Table II and Table III shows the comparison of performance on online tracking by Kalman Filter [31] with SVR-KF and LSTM on physiological sensors with SVR-LSTM [20], respectively. From Table II, in arousal dimension, one can observe that SVR-KF achieves a higher CCC value than online tracking by Kalman Filter [31] for video appearance, video geometric, ECG, EDA and HRHRV modality. As for the valence dimension, it has achieved a balanced performance, SVR-KF performs better in video appearance, video geometric and EDA modality, whilst online affect tracking using KF is superior on audio and ECG modality. Surprisingly, it achieved similar performances of CCC for both SVR-KF and online affect tracking in the HRHRV modality. From Table III, similar observations were seen where SVR-LSTM gives superior results on arousal dimensions for both HRHRV and EDA modality. In valence dimensions, applying LSTM directly on physiological sensor features [20] gives superior results on HRHRV modality, however SVR-LSTM gives comparable results on EDA modality.

TABLE II: Comparison of unimodal performance by Somandepalli et al and SVR-KF on the development set.

Modality	Arousal		Valence	
	[31]	SVR-KF	[31]	SVR-KF
D-Audio	**0.800**	0.780	**0.448**	0.432
D-VA	0.481	**0.524**	0.474	**0.481**
D-VG	0.297	**0.415**	0.612	**0.638**
D-ECG	0.272	**0.337**	**0.159**	0.153
D-EDA	0.080	**0.163**	0.178	**0.197**
D-HRHRV	0.383	**0.423**	0.298	0.298

TABLE III: Comparison of unimodal performance by Brady et al. and SVR-LSTM on the development set.

Modality	Arousal		Valence	
	[20]	SVR-LSTM	[20]	SVR-LSTM
D-HRHRV	0.357	**0.417**	**0.364**	0.293
D-EDA	0.082	**0.130**	0.177	**0.194**

VI. CONCLUSION

This paper proposed and investigated a 2-stage regression approach, for naturalistic affective emotion recognition from multiple modalities.

Results gained from RECOLA database indicate that the 2-stage regression approach can match or outperform the corresponding conventional SVR models when performing naturalistic affect recognition. An interesting observation was, apart from the audio modality, that all three subsequent models outperform baseline results of arousal dimension, and give competitive results towards baseline results of valence dimension in unimodal setting. This demonstrates the effectiveness of proposed framework, in terms of capturing temporal information in prediction of emotion recognition, and being able to work with different combination of regression strategies.

For future works, the combination/fusion of all individual predictions will be investigated.We also intend to evaluate the performance on other affective dimension regression tasks in order to generalize the promising advantage offered by 2-stage regression framework.

REFERENCES

[1] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 1, pp. 64–84, 2009.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[3] S. Petridis and M. Pantic, "Prediction-Based Audiovisual Fusion for Classification of Non-Linguistic Vocalisations," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 45–58, 2016.

[4] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016*. New York City, NY: IJCAI/AAAI, July 2016, 7 pages.

[5] A. Metallinou, A. Katsamanis, M. Wollmer, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification (Extended abstract)," in *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, 2015, pp. 463–469.

[6] M. Soleymani, S. Asghari Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 3045, no. c, pp. 1–1, 2015.

[7] M. Valstar, J. Gratch, F. Ringeval, M. T. Torres, S. Scherer, and R. Cowie, "AVEC 2016 Depression , Mood , and Emotion Recognition Workshop and Challenge," 2016.

[8] R. Prakash Gadhe, R. R. Deshmukh, and V. B. Waghmare, "KNN based emotion recognition system for isolated Marathi speech," *International Journal of Computer Science Engineering (IJCSE)*, 2015.

[9] V. Imbrasaite, T. Baltrusaitis, and P. Robinson, "CCNF for continuous emotion tracking in music: Comparison with CCRF and relative feature representation," *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, 2014.

[10] Z. Huang, B. Stasak, T. Dang, K. Wataraka Gamage, L. Phu, V. Sethu, and J. Epps, "Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016," *Proceedings of the 6th International Workshop on AVEC, ACM MM*, pp. 19–26, 2016.

[11] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J. P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, 2015.

[12] S. Hochreiter and J. Urgen Schmidhuber, "LONG SHORT-TERM MEMORY," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, feb 2013.

[14] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - Towards continuous emotion recognition with modelling of long-range dependencies," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2008, pp. 597–600.

[15] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long Short Term Memory Recurrent Neural Network based Multimodal Dimensional Emotion Recognition," *AVEC workshop*, pp. 65–72, 2015.

[16] H. Meng and N. Bianchi-Berthouze, "Affective state level recognition in naturalistic facial and vocal expressions," *IEEE Transactions on Cybernetics*, vol. 44, no. 3, pp. 315–318, 2014.

[17] H. Gunes, M. Nicolaou, and M. Pantic, "Continuous Analysis of Affect from Voice and Face," in *Computer Analysis of Human Behavior SE - 10*, 2011, pp. 255–291.

[18] L. van der Maaten, "Audio-visual emotion challenge 2012: a simple approach," *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 473–476, 2012.

[19] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, no. 1, p. 35, 1960.

[20] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: ACM, 2016, pp. 97–104.

[21] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions using Kalman filtering," in *Proceedings - 9th International Conference on Machine Learning and Applications, ICMLA 2010*, 2010, pp. 655–660.

[22] a. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[23] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, "Time-Delay Neural Network for Continuous Emotional Dimension Prediction From Facial Expression Sequences," *IEEE Transactions on Cybernetics*, vol. 46, no. 4, pp. 916–929, 2016.

[24] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[25] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, no. x, pp. 155–161, 1997.

[26] 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*, 2013.

[27] F. Eyben, F. Weninger, F. Groß, B. Schuller, F. Gross, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," *Proceedings of the 21st ACM International Conference on Multimedia (MM 2013)*, no. May, pp. 835–838, 2013.

[28] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "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, 2016.

[29] T. R. Almaev and M. F. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 2013, pp. 356–361.

[30] X. Xiong and F. De La Torre, "Supervised descent method and its applications to face alignment," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 532–539, 2013.

[31] K. Somandepalli, R. Gupta, M. Nasir, B. M. Booth, S. Lee, and S. S. Narayanan, "Online affect tracking with multimodal kalman filters," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: ACM, 2016, pp. 59–66.