Automatic Affective Dimension Recognition from Naturalistic Facial Expressions Based on Wavelet Filtering and PLS Regression

Yona Falinie A. Gaus, Hongying Meng, Asim Jan, Fan Zhang, and Saeed Turabzadeh
Department of Electronic and Computer Engineering, Brunel University London, UK

Abstract— **Automatic affective dimension recognition from facial expression continuously in naturalistic contexts is a very challenging research topic but very important in human-computer interaction. In this paper, an automatic recognition system was proposed to predict the affective dimensions such as Arousal, Valence and Dominance continuously in naturalistic facial expression videos. Firstly, visual and vocal features are extracted from image frames and audio segments in facial expression videos. Secondly, a wavelet transform based digital filtering method is applied to remove the irrelevant noise information in the feature space. Thirdly, Partial Least Squares regression is used to predict the affective dimensions from both video and audio modalities. Finally, two modalities are combined to boost overall performance in the decision fusion process. The proposed method is tested in the fourth international Audio/Visual Emotion Recognition Challenge (AVEC2014) dataset and compared to other state-of-the-art methods in the affect recognition sub-challenge with a good performance.**

I. INTRODUCTION

Human face provides an essential, spontaneous channel for the communication of mental states. In addition to functioning as a conversation enhancer, facial expressions directly communicate feelings, cognitive mental states, and attitude toward other people.

In the affective computing field [18], various studies have been carried out to create systems that can recognize the affective states of their user by analyzing their vocal [2], [21] and facial expressions [19], [14]. Most of that work has been done on acted or stereotypical expressions. More recently, there has been a shift towards using naturalistic expressions to create systems that can interact with people in their everyday life (e.g. [7], [22], [12]).

Naturalistic expressions present a big challenge to the research community because they are less stereotypical and not always fully-fledged expressions. Furthermore, the dynamic of these expressions is more complex, leading to a larger variability in the way affect is expressed. Finally, human affective states tend to change much slower than the typical rate at which video or audio is recorded in naturalistic expressions as analyzed in [12]. The focus of this paper is to utilize this property in the naturalistic expressions in an efficient way and build a better automatic affective dimension recognition system.

The proposed automatic affective dimension recognition system includes four steps: (1) Advanced feature extraction methods are introduced to capture the characteristics of both video frames and audio segments; (2) Wavelet Transform (WT) based digital filtering method is applied on time line on feature space to remove the irrelevant noise; (3)Partial Least Squares (PLS) regression is employed to predict the measurement of three dimensional affects for both video and audio modalities; (4) Decision fusion is done to combine video and audio modalities together. It was tested in the AVEC2014 Challenge [28] dataset and good performance is achieved.

The rest of the paper is organized as follows. Section 2 briefly reviews related work in this area. Section 3 provides a detailed description of the proposed method. Section 4 displays and discusses the experimental results on the AVEC2014 dataset [28]. Section 5 is the conclusion and discussion.

II. RELATED WORK

Continuous emotional state detection and prediction has been a challenging research topic in recent years within the affective computing research community. Many research focused on building automatic systems that can recognize preselected instances of expressions in acted or non-acted videos. An important challenge is to create systems that can continuously (i.e. over time) monitor and classify affective expressions into either discrete affective states or continuous affective dimensions [19]. The initial approaches treated the videos as sequences of independent facial expression frames and aimed at improving the classification performances for each independent expression at frame level. For example, the baseline method of AVEC2013 challenge [29] treats each unit of expression (e.g. a video frame, a audio segment) independently and makes it a standard regression problem at frame or word level. Firstly, Local Phase Quantization (LPQ) feature was extracted from each frame to describe the characteristic of the face image. Then, feature selection process was carried out to remove less relevant feature. Furthermore, optimized kernel based Support Vector Regression (SVR) method was used for regression. Finally, a fusion process combining video and audio together was used to boost the overall performance. In the baseline method of AVEC2014 challenge [28], Local Gabor Binary Patterns from Three Orthogonal Plane (LGBP-TOP) was extracted from videos and an epsilon-SVR with intersection kernel was employed. The results from these works miss the opportunity to exploit temporal relations that exist between consecutive instances of an expression in naturalistic facial expressions.

Recent years, the research work on emotion recognition from naturalist expressions has been seen a significant progress with the great support of naturalist expression datasets and competitions(e.g. AVEC2011 [23],

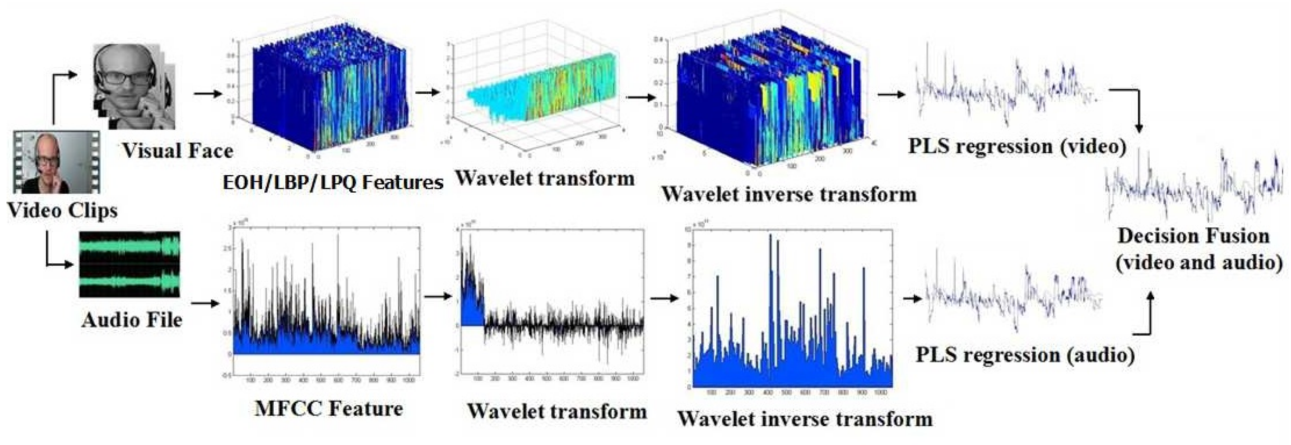


Fig. 1. Overview of the proposed automatic affective dimension recognition system.

AVEC2012 [24], AVEC2013 [29], AVEC2014 [28]). A few works tried to use the temporal relations in the naturalistic expressions by modeling. Meng and Berthouze [12] proposed a multi-stage automatic affective expression recognition system to use HMMs to take into account this temporal relationship and finalize the classification process. The hidden states of the HMMs are associated with the levels of affective dimensions to convert the classification problem into a best path finding problem in HMM. The system achieved the best performance on the audio data of AVEC2011 dataset. Nicolle et. al. [15] proposed a method to use log-magnitude Fourier spectra to extract multiscale dynamic descriptions of signals characterizing global and local face appearance as well as head movements and voice. It performs a kernel regression with very few representative samples selected via a supervised weighted-distance-based clustering, that leads to a high generalization power. Then a particularly fast regressors level fusion framework was used to merge systems based on different modalities. Savran et. al. [20] use temporal statistics of texture descriptors extracted from facial videos, a combination of various acoustic features, and lexical features to create regression based affect estimators for each modality. The single modality regressors are then combined using particle filtering, by treating these independent regression outputs as measurements of the affective states in a Bayesian filtering framework, where previous observations provide prediction about the current state by means of learned affect dynamics. At AVEC2014 affect recognition sub-challenge, the temporal relations in naturalistic expressions was used to boost the performance in decision level filtering [10] [9]. Kachele et al. [10] proposed an approach based on abstract meta information about individual subjects and also prototypical task and label dependent templates to infer the respective emotional states and achieved the best performance. Gupta et al. [9] proposed a 4-stage system including individual training, prediction processing, system fusion and temporal regression. Chao et. al [4] utilized deep belief network based models to recognize emotion states from audio and visual modalities by combining the predicted results from different modalities and emotion temporal context information simultaneously.

These approaches have produced very good performances in the challenge competition. They used the relationship between consecutive expression unit (e.g. frame or audio segment) on the decision level to boost the overall performance. However, in terms of temporal relations in the feature level, only [4] did it by temporal pooling functions in the deep neutral network. In this paper, we will investigate how to use this temporal relations in the feature space further. We designed a wavelet transform based digital filtering technique on feature vector to remove their high frequency component and then integrate it in our affective dimension recognition system.

III. AFFECTIVE DIMENSION RECOGNITION

In affective dimensional space, it represents more details on how we deal with emotional states other than six basic emotion categories (e.g. happy, sad, disgust, anger, surprise and fear). There are Arousal, Dominance and Valence. Arousal is the individual's global feeling of dynamism or lethargy. It subsumes mental activity, and physical preparedness to act as well as overt activity. Dominance is an individual's sense of how much they feel to be in control of their current situation. Valence is an individual's overall sense of feeling positive or negative about the things. In this sub-challenge, the main task was to classify the scales of arousal, valence and dominance from video and audio database. Therefore, the automatic affective dimension recognition system needs to comprehensively model the variations in visual and vocal clues and automatically predict the scale of each Arousal, Dominance and Valence from video and audio. From the machine learning point of view, it is a regression problem, not a classification problem, on each individual frame in a image sequence because the predicted values are real numbers.

A. System Overview

Figure 1 depicts the framework of the proposed approach to automatic affective dimension recognition. For each video clip, we deal with the channel of the video and audio signals separately. In the step of video process, feature extraction is implemented towards each of the video frame, producing features for texture representation. Then wavelet transform based digital filtering technique is implied towards the features. The main reason is that we believe the affect related features should also change slowly as the affective dimensions. The high frequency components might be noise that is irrelevant to the affective dimensions. The Partial Least Square (PLS) regression is then applied to predict the scales of the affective dimensions. While in the step of audio process, a set of Mel-frequency cepstral coefficients (MFCCs) are employed to encode the characteristic of the audio. Then wavelet transform based digital filtering is again used to remove irrelevant information of the vocal expression, followed by PLS based regression as does in video process. The prediction labels from each video and audio is smoothed using low pass filtering to enhance the results. Finally, the video and audio modalities was combined to improve the overall performance.

B. Image Feature Extraction

Based on the modality of facial expression and vocal expression recorded in video data, three dynamic features are extracted respectively, which are detailed and presented in the subsequent.

1) EOH: Edge Orientation Histogram (EOH), an efficient and powerful operator, is regarded as a simpler version of Histogram of Oriented Gradients (HOG) [5] that captures the edge or the local shape information of an image. For an image frame, firstly, the edge image is captured using Sobel edge detection algorithm from each frames. Secondly, the angle and intensity of the gradient function on each pixel is calculated and arranged into a polar coordinate system. Finally, the histogram from each block is normalized and concatenated into a feature vector. If the whole image is divided into 4×4 blocks and each polar coordinate system has 24 bins, the feature vector will have 384 components.

2) LBP: Local Binary Pattern (LBP), a non parametric descriptor summarizes local texture structures of images into a set of patterns. It is highly discriminative and its key advantages, namely invariance to monotonic gray level changes and computational efficiency, make it successful and popular in many topics, e.g. texture classification [16] and face recognition [1]. The basic LBP operator labels the pixels of an image with decimal numbers, called LBP codes, which encode the local structure around each pixel.

3) LPQ: Local Phase Quantization (LPQ) operator was originally proposed by Ojansivu and Heikkila [17] as a texture descriptor and further used in emotion recognition in [8]. LPQ is based on the blur invariance property of the Fourier phase spectrum. It uses the local phase information extracted using the 2-D short-term Fourier transform (STFT) computed over a rectangular neighborhood at each pixel position of the image. Only four complex coefficients are considered, corresponding to 2-D frequencies of an images.

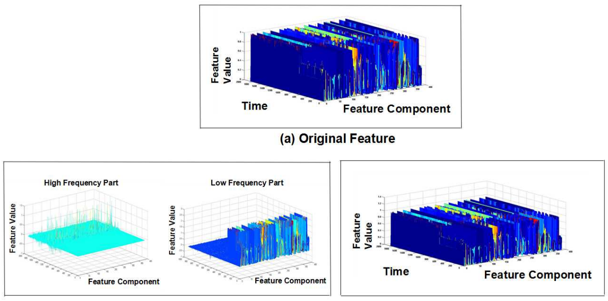
C. Audio Feature Extraction

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up the mel-frequency cepstrum of an audio signal. They are derived from a type of cepstral representation of the audio clip (a nonlinear spectrum-of-a-spectrum) and widely used in automatic speech and speaker recognition. For each MFCCs, we fully utilized long, short and valid-segmented baseline acoustic feature sets. For detailed information, please see the baseline paper [28].

D. Wavelet Filtering

Digital filters are generally used for two purposes: (1) separation of signals that have been combined, and (2) restoration of signals that have been distorted in some way [26]. In image processing, digital filters are mainly used to suppress either the high frequencies in the image, i.e. smoothing the image, or the low frequencies, i.e. enhancing or detecting edges in the image. Wavelet transform is a special digital filter which has been used as image compression, edge detection, noise removal and so on.

Human's emotional state changes much slower than the video frame/word recording rate in naturalistic facial and vocal expressions [12]. That means there are no much difference within the frames in a short time period. However, in the feature space, there are significant differences between nearby frames due to the powerful feature extraction methods. That means that there exist some irrelevant information in the features that will decrease the overall pattern recognition performance. Here, wavelet transform is used to remove this irrelevant noise in the features, by compressing high frequency components in feature space without distortion. Moreover, Haar wavelet transform is one of a easiest tool to determine where the low and high frequency area are.



(a) Original Feature

(b) Haar Wavelet Decomposition

(c) Reconstruction Feature

Fig. 2. Haar Wavelet Transform filtering on selected feature space. 1840 frames, selected components and decomposition level=4. (a). Original feature vectors (b). Low and high frequency parts of the feature vectors in wavelet space (c). Filtered feature vectors.

1) Haar Wavelet Transform: For a signal $X = (x_1, x_2, \cdots, x_N)$ with N values, it can be decomposed into two parts s and d with the length of $N/2$ each based on Haar wavelet transform as the following equations:

$$s_k = \frac{x_{2k-1} + x_{2k}}{2}, \quad k = 1, 2, \cdots, N/2 \tag{1}$$

$$d_k = \frac{x_{2k-1} - x_{2k}}{2}, \quad k = 1, 2, \cdots, N/2 \tag{2}$$

s_k is called approximation of the signal that represents the low frequency part of the signal while d_k is called details of the signal that represents the high frequency of the signal. After this first level Haar wavelet transform decomposition, the signal $X = (x_1, x_2, \cdots, x_N)$ will be transformed to

$$(s_1, s_2, \cdots, s_{N/2} | d_1, d_2, \cdots, d_{N/2}), \tag{3}$$

The original signal X can be fully reconstructed from s_k and d_k easily by using:

$$s_k + d_k = \frac{x_{2k-1} + x_{2k}}{2} + \frac{x_{2k-1} - x_{2k}}{2} = x_{2k-1}, \quad k = 1, 2, \cdots, N/2 \tag{4}$$

and

$$s_k - d_k = \frac{x_{2k-1} + x_{2k}}{2} - \frac{x_{2k-1} - x_{2k}}{2} = x_{2k}, \quad k = 1, 2, \cdots, N/2 \tag{5}$$

2) Filtering: In order to remove the high frequency components of the signal, after performing wavelet transform, low frequency part s will be kept and high frequency part d will be replaced by zeros. In this way, the reconstructed signal will lose the high frequency components. The low frequency part s can be decomposed further by Haar wavelet transform to remove more high frequencies. It was called level 2 wavelet transform. It can be carried this way further for level 3, level 4 decomposition. Figure 2 shows the wavelet transform process on EOH feature with 4 level decomposition. It can be seen, through reconstruction step by step, final reconstructed signal generated is smoother along the time line. For affective dimensions prediction, smooth and simple features are matching the slow change property of the real affective dimensions.

E. PLS Regression

Partial Least Squares (PLS) regression [6] is a statistical algorithm that bears some relation to principal components regression. Instead of finding hyperplanes of minimum variance between the response and independent variables, it builds a linear regression model by projecting the response and independent variables to another common space. Since both the response and independent variables are projected to a new space, the approaches in the PLS family are known as bilinear factor models. PLS is used here because it achieved better performance than SVR in emotion recognition [13].

F. Filtering on Decision Label

After performing Haar wavelet transform on the features, and use PLS regression for training and testing data, it will give prediction label as the output. Since these challenge requires the prediction of continuous affect labels per frame, we carry out smoothing over the prediction labels using simple low pass filtering. Low pass filtering is carried out on prediction of each development and testing frames to further enhance the results.

G. Decision Fusion

The decision fusion stage aims to combine multiple decisions into a single and consensus one [25]. The linear opinion pool method is used in this case due to its simplicity [3], and a weighted sum rule [27] is defined to combine the predicted values from each decision as shown in Equation 6.

$$D_{\texttt{linear}}(\hat{x}) = \sum_{i=1}^{K} \alpha(i) D_i(\hat{x}) \tag{6}$$

where $\hat{x}$ is a testing sample and $D_i(\hat{x})$ is its i_{th} decision value ($i = 1, 2, ..., K$) while $alpha(i)$ is its corresponding weight which should satisfy $\sum_{i=1}^{K} \alpha(i) = 1$.

IV. EXPERIMENTAL RESULTS

A. AVEC2014 Dataset

The proposed approach is evaluated on the Audio/Visual Emotion Challenge (AVEC 2014) dataset, which includes recording of subjects performing a Human-Computer Interaction task while being recorded by a webcam and microphone. The video and audio recordings were split into three partition: training, development, and testing set of 150 Northwind-Freeform pairs, totaling 300 task recordings. The detailed information about this dataset can be found in the AVEC2014 baseline paper [28].

B. Experimental Setting

The proposed system was trained on training set and tested on development and testing sets for the affect recognition challenge, where the level of affect has to be predicted for each frame of the recording. The Pearson's correlation coefficients (CORR), and Root Mean Square Error (RMSE) over all M sessions are both used as measurements in challenge competition, as shown in Equation 7 and Equation 8.

$$CORR = \frac{1}{M} \sum_{i=1}^{M} \frac{\sum_{j=1}^{N_i} (y_i^j - \overline{y_i})(\hat{y}_i^j - \overline{\hat{y}_i})}{\sqrt{\sum_{j=1}^{N_i} (y_i^j - \overline{y_i})^2} \sqrt{\sum_{j=1}^{N_i} (\hat{y}_i^j - \overline{\hat{y}_i})^2}} \tag{7}$$

$$RMSE = \frac{1}{M} \sum_{i=1}^{M} \sqrt{\frac{1}{N_i} \sum_{j=1}^{N_i} (y_i^j - \hat{y}_i^j)^2} \tag{8}$$

where N_i is the number of frames in session $i(i = 1, 2, \cdots, M)$. y_i^j and $\hat{y}_i^j$ are the true and predicted values for the frame $j(j = 1, 2, \cdots, N_i)$ in session i. $\overline{y_i}$ and $\overline{\hat{y}_i}$ are the mean values of y_i^j and $\hat{y}_i^j$ for session $i(i = 1, 2, \cdots, M)$.

C. Performance Comparison

1) AVEC2014 Development Dataset: Table I shows the performance of different features on each affective dimensions in terms of CORR and RMSE in the development dataset. It can be seen that there is small difference between video features EOH, LBP and LPQ while each one achieved best performance in one of three dimensions. However, for audio features ('long', 'short', and 'valid segmented'), 'valid segmented' achieved best performance in Dominance and Valence while 'long' achieved best performance in Arousal. For the video and audio fusion, there is no significant difference. The reason is that only very simple fusion rule was used here. In comparison with the baseline results, the proposed method outperform in every modality and every dimension.

TABLE I

PERFORMANCE OF AFFECTIVE DIMENSION RECOGNITION MEASURED IN CROSS-CORRELATION (CORR) AND RMSE AVERAGED ALL OVER SEQUENCES IN THE DEVELOPMENT SET. BOLD INDICATE HIGHEST CORRELATION VALUE ACHIEVED ACROSS EACH DIMENSION AND MODALITY.

Modality	Affect Dimension	Feature	CORR	RMSE
Video	Arousal	EOH	0.5669	0.0879
		LBP	**0.5868**	0.0956
		LPQ	0.5624	0.0987
	Dominance	EOH	0.6021	0.1026
		LBP	0.5135	0.1049
		LPQ	**0.6023**	0.1015
	Valence	EOH	**0.5523**	0.0670
		LBP	0.5480	0.0706
		LPQ	0.5211	0.0665
Audio	Arousal	Long	**0.6136**	0.0992
		Short	0.5911	0.0981
		Vad_seg	0.5954	0.1002
	Dominance	Long	0.5866	0.0989
		Short	0.5902	0.0988
		Vad_seg	**0.6054**	0.0987
	Valence	Long	0.5773	0.0659
		Short	0.5509	0.0659
		Vad_seg	**0.5798**	0.0661
Video + Audio	Arousal	EOH_Long	0.5668	0.0894
		LBP_Long	**0.5873**	0.0944
		LPQ_Long	0.5165	0.0955
	Dominance	EOH_Vad_seg	**0.6021**	0.0998
		LBP_Vad_seg	0.5891	0.1005
		LPQ_Vad_seg	0.5788	0.1011
	Valence	EOH_Vad_seg	**0.5525**	0.0654
		LBP_Vad_seg	0.5479	0.0676
		LPQ_Vad_seg	0.5199	0.0660
Baseline (video)	Arousal	LGBP_TOP	0.412	–
	Dominance		0.319	–
	Valence		0.355	–
Baseline (audio)	Arousal	LLDs+MFCC	0.517	–
	Dominance		0.439	–
	Valence		0.347	–
Baseline (video +audio)	Arousal	LGBP_TOP+ LLDs+MFCC	0.421	–
	Dominance		0.348	–
	Valence		0.236	–

2) AVEC2014 Testing Dataset: The results for the official AVEC2014 challenge test set can be seen in Table II. All parameters, such as filter window size, number of component in PLS are identical to the previous set of experiments on the development set. From Table II, it shows that the results are better than the baseline results [28] in every modality and dimension.

TABLE II

PERFORMANCE OF AFFECTIVE DIMENSION RECOGNITION MEASURED IN CROSS-CORRELATION (CORR) AND RMSE AVERAGED ALL OVER SEQUENCES IN THE TESTING SET. BOLD INDICATE HIGHEST CORRELATION VALUE ACHIEVED ACROSS EACH DIMENSION AND MODALITY.

Modality	Affect Dimension	Feature	CORR	RMSE
Video	Arousal	EOH	**0.5713**	0.0921
		LBP	0.5597	0.0961
		LPQ	0.5711	0.1017
	Dominance	EOH	0.4916	0.1009
		LBP	**0.5179**	0.0597
		LPQ	0.4835	0.0993
	Valence	EOH	0.5032	0.0570
		LBP	0.5183	0.0597
		LPQ	**0.5319**	0.0560
Audio	Arousal	Long	**0.5277**	0.0951
		Short	0.4913	0.0954
		Vad_seg	0.5081	0.0953
	Dominance	Long	0.4750	0.0907
		Short	0.4892	0.1797
		Vad_seg	**0.4913**	0.0901
	Valence	Long	0.4987	0.0552
		Short	0.4468	0.0553
		Vad_seg	**0.5355**	0.0548
Video + Audio	Arousal	EOH_Long	0.5721	0.0950
		LBP_Long	0.5586	0.0935
		LPQ_Long	**0.5760**	0.0968
	Dominance	EOH_Vad_seg	0.4913	0.0953
		LBP_Vad_seg	**0.5182**	0.0915
		LPQ_Vad_seg	0.4842	0.0945
	Valence	EOH_Vad_seg	0.5030	0.0542
		LBP_Vad_seg	0.5184	0.0559
		LPQ_Vad_seg	**0.5354**	0.0549
Baseline (video)	Arousal	LGBP_TOP	0.2062	–
	Dominance		0.1959	–
	Valence		0.1879	–
Baseline (audio)	Arousal	LLDs+MFCC	0.540	–
	Dominance		0.360	–
	Valence		0.355	–
Baseline (video +audio)	Arousal	LGBP_TOP+ LLDs+MFCC	0.478	–
	Dominance		0.324	–
	Valence		0.282	–

In addition, our method was compared to state-of-the-art results in the AVEC2014 affect challenge competition as shown in Table III, it can be seen that our method achieved competitive performance in term of CORR and best performance in term of RMSE.

V. CONCLUSION AND DISCUSSION

In this paper, an automatic affective dimension recognition system is proposed based on wavelet filtering and PLS regression for naturalistic facial expressions. Instead of using the temporal relations in the decision level like other methods, Haar wavelet transform based digital filtering method was used to remove any irrelevant noise in the feature space. The reconstructed features were input to PLS regression and final fusion process was used for combining video and audio modalities.

TABLE III

Performance comparison with state-of-the-art methods in AVEC2014 affect recognition sub-challenge in term of average CORR and RMSE values.

Team	Method	CORR	RMSE
Baseline [28]	SVR+Fusion	0.4185	0.2090
Ulm [10]	Subjects+Label Inference	**0.5946**	0.1009
NLPR [4]	Deep Learning+Fusion	0.5499	0.1630
SAIL [9]	Fusion+Temporal Regression	0.5219	0.0831
BU-CMPE [11]	CCA ensemble	0.3932	0.0928
Our method	Wavelet Filtering+PLS+Fusion	0.5432	**0.0810**

The system was trained on the AVEC2014 training, tested on both development and testing set, and compared with baseline method. It was also compared with all the state-of-the-art methods in the AVEC2014 affect recognition sub-challenge with fairly good performance. NLPR [4], SAIL [9], BU-CMPE [11] and our method achieve better performance than baseline method, while Ulm [10] achieves best performance. However, it utilized extra information on subjects and annotation process that is not comparable with other methods.

The performance of the proposed system can be enhanced by improving the fusion rule on video and audio modalities. In addition, other wavelet transform filters might achieve better results than Haar wavelet here because Haar wavelet is the simplest one in the wavelet family. For the future work, the proposed method can be tested on other naturalistic expression datasets.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] M. E. Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572 – 587, 2011.

[3] I. Bloch. Information combination operators for data fusion: a comparative review with classification. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 26(1):52–67, 1996.

[4] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen. Multi-scale temporal modeling for dimensional emotion recognition in video. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, pages 11–18, New York, NY, USA, 2014. ACM.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[6] S. de Jong. Simpls: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993.

[7] L. Devillers, C. Vaudable, and C. Chastagnol. Real-life emotion-related states detection in call centers: a cross-corpora study. In *INTERSPEECH'10*, pages 2350–2353, 2010.

[8] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using PHOG and LPQ features. In *Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21-25 March 2011*, pages 878–883, 2011.

[9] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan. Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, pages 33–40, New York, NY, USA, 2014. ACM.

[10] M. Kächele, M. Schels, and F. Schwenker. Inferring depression and affect from application dependent meta knowledge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, pages 41–48, New York, NY, USA, 2014. ACM.

[11] H. Kaya, F. Çilli, and A. A. Salah. Ensemble CCA for continuous emotion prediction. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, pages 19–26, New York, NY, USA, 2014. ACM.

[12] H. Meng and N. Bianchi-Berthouze. Affective state level recognition in naturalistic facial and vocal expressions. *IEEE Transactions on Cybernetics*, 44(3):315–328, 2014.

[13] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13, pages 21–30, New York, NY, USA, 2013. ACM.

[14] H. Meng, B. Romera-Paredes, and N. Bianchi-Berthouze. Emotion recognition by two view SVM_2K classifier on dynamic facial expression features. In *FG*, pages 854–859, 2011.

[15] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 501–508, New York, NY, USA, 2012. ACM.

[16] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[17] V. Ojansivu and J. Heikkila. Blur insensitive texture classification using local phase quantization. In A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, editors, *Image and Signal Processing*, volume 5099 of *Lecture Notes in Computer Science*, pages 236–243. Springer Berlin Heidelberg, 2008.

[18] R. W. Picard. *Affective Computing*. The MIT Press, 1997.

[19] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2014.

[20] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 485–492, New York, NY, USA, 2012. ACM.

[21] B. Schuller, A. Batliner, S. Steidl, and D. Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087, 2011.

[22] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu. Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image Vision Comput.*, 27(12):1760–1774, nov 2009.

[23] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. AVEC 2011: The first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction (ACII 2011)*, 2011.

[24] B. Schuller, M. F. Valstar, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge - an introduction. In *ICMI*, pages 361–362, 2012.

[25] A. Sinha, H. Chen, D. G. Danu, T. Kirubarajan, and M. Farooq. Estimation and decision fusion: A survey. In *IEEE International Conference on Engineering of Intelligent Systems*, pages 1–6, 2006.

[26] S. W. Smith. The scientist and engineer's guide to digital signal processing. California Technical Publishing, San Diego, California.

[27] N. Ueda. Optimal linear combination of neural networks for improving classification performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(2):207–215, 2000.

[28] M. F. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. AVEC 2014 - 3d dimensional affect and depression recognition challenge. In *International Conference on ACM Multimedia - Audio/Visual Emotion Challenge and Workshop*, 2014.

[29] M. F. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. AVEC 2013 - the continuous audio/visual emotion and depression recognition challenge. In *International Conference on ACM Multimedia - Audio/Visual Emotion Challenge and Workshop*, 2013.