2018 13th IEEE International Conference on Automatic Face & Gesture Recognition

Linear and Non-Linear Multimodal Fusion for Continuous Affect Estimation in-the-Wild

Yona Falinie A. Gaus, Hongying Meng
Department of Electronic and Computer Engineering
Brunel University London
London, UK
{yonafalinie.abdgaus, hongying.meng}@brunel.ac.uk

***Abstract*—Automatic continuous affect recognition from multiple modality in the wild is arguably one of the most challenging research areas in affective computing. In addressing this regression problem, the advantages of the each modality, such as audio, video and text, have been frequently explored but in an isolated way. Little attention has been paid so far to quantify the relationship within these modalities. Motivated to leverage the individual advantages of each modality, this study investigates behavioral modeling of continuous affect estimation, in multimodal fusion approaches, using Linear Regression, Exponent Weighted Decision Fusion and Multi-Gene Genetic Programming. The capabilities of each fusion approach are illustrated by applying it to the formulation of affect estimation generated from multiple modality using classical Support Vector Regression. The proposed fusion methods were applied in the public Sentiment Analysis in the Wild (SEWA) multimodal dataset and the experimental results indicate that employing proper fusion can deliver a significant performance improvement for all affect estimation. The results further show that the proposed systems is competitive or outperform the other state-of-the-art approaches.**

***Keywords*-linear; affect; non-linear; fusion; GP; linear regression**

I. Introduction

Automatic continuous emotion estimation aims to enable intelligent systems to recognize, feel, infer and interpret human emotions. Recent developments of sensors like camera and microphone have led to a renewed interest in emotion recognition, from recognizing discrete basic emotion to recognizing continuous emotion, or continuous affect estimation, in terms of Arousal and Valence [1] [2].

Numerous studies have been performed to compare the advantages offered by a wide range of modeling techniques for continuous affect recognition [3] [4]. AVEC challenge aim to create a benchmarks to evaluate modeling systems that are capable of recognizing affect recognition beyond laboratory conditions.

Therefore, this paper describes a multimodal approach on SEWA dataset, by leveraging the individual advantages of each modality, then quantifying the relationship between each modality. Here, we apply decision fusion on initial prediction by employing linear and non linear fusion approach. Some researchers advocate that combined multiple modalities will contribute to the recognition accuracy, and it can be achieved in numerous way. Method from simple mapping such as averaging [5] to complex method such as linear regression [6], SVR [7] or Kalman filters based [8] [9] has been used to combine prediction from multiple modalities. However, a systematic understanding of the relationship between modalities contribute to the higher recognition accuracy is not fully explored. Few methods assumed that the continuous affect label are linear in time. Looking closely at the gold standard affect label in [5], potential nonlinearities behavior may occur in continuous affect label. In summary, the contributions of this paper are two-folds:

- We investigate linear and non-linear multimodal fusion approach to predict each affect dimension.
- We examine the possibility of constructing affect estimation prediction equation from initial prediction result. These modeling equation can provide convenient way to express the relationship between each modality and affect estimation in multimodal fusion manner.

The rest of this paper is organized as follows. In the next section, we discuss related work on affect estimation in-the-wild settings from audio, video and text. Section 3 describes the proposed approach on affect estimation system. In Section 4, we elaborate more on experimental results as well as the discussion. Finally, Section 5 concludes the paper and summarizes our findings.

II. Related Works

The evolution of continuous affect estimation usually comprises of two system: 1) classical features extraction methods which are grounded on statistical/mathematical notions, and 2) modern machine learning which is based on algorithms from artificial intelligence field. In the literature of continuous affect recognition, typically there are two modality present to estimate affect, audio and visual modality [10]. Audio modality, usually represent by audio features such as acoustic low-level descriptors (LLD), include a wide range of features that cover spectral, cepstral, prosodic and voice quality information. As for video modality, it typically referred as video features which consists of appearance feature and geometric feature. Noted that, the video modality

978-1-5386-2335-0/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00079

capture the change and intensity of facial expressions over time. For appearance feature, the most popular example would be local binary patterns (LBP) and histogram of gradients (HOG) modeled using bag of words (BOW). A robust variant of LBP, called Local Gabor Binary Patterns from Three Orthogonal Planes (LGBPTOP) is incorporated in spatio-temporal volumes of the video after convolving with 2D Gabor filter-bank. LGBPTOP has been used as baseline feature in automatic affect recognition challenge [3] [11]. Video geometric features include identifying landmarks on the face [11] or shoulder [12] or the whole body [13].

Experimenting with text modality is quite new approach in continuous affect recognition. The semantic of the words used can be an important aspect in emotion detection. It is because, the words chosen can say a lot on the current state of emotion of the person. In previous AVEC challenge, only Povolny et al. [14] addressing text feature by exploring automatic speech recognition, lexicon-based approach and word embedding technique, in order to create a dictionary for each utterance. In this paper, we will go deeper on text modality by incorporating a *bag-of-text-words* (BOTW) feature representation generated based on the transcription of the speech.

Affect estimation is usually performed with human-annotated emotional dimension such as Arousal for emotion activation, Valence for emotion positiveness and for the first time; Likability which presents the users preference to the commercial product, for gold standard ratings.

Modeling approaches here are generally supervised and regression based method is the approach of estimating affect. Support Vector Machine (SVR) is perhaps the most widely used regression method for affect estimation and has been regarded as baseline approach for affect estimation [3] [15] [11]. Recent literature takes into account short term temporal correlation such as Continuous Conditional Random Fields (CCRF) on top of SVRs [16] and various type of neural network including Time Delay Neural Networks [17], Recurrent Neural Networks (RNN) [18] and Long-Short Term Memory RNN (LSTM-RNN) in [19] [20]. Another study [12], employed a bidirectional LSTM model with an output-associative framework to achieve improved performance in affect prediction. Following this trend, a deep bidirectional LSTM was proposed [21] in which was gives the highest results in [3].

When dealing with several modality and modeling technique the question of how to fuse them arises. Feature level fusion and decision level fusion is the most well known approach for assessing continuous affect estimation. Feature level fusion is undertaken simply by concatenating each features from multiple modality then a single classifier is trained on the concatenated features [18] [22]. However, feature-level fusion is plagued by several challenges. Generally, this approach tends to create very high dimensional feature vectors and lead to overfit. Secondly, features from multiple modalities are collected at different time scales. For example, HRV features from physiological modality typically extracted in minutes [23] while LLD features from audio modality can be in the order of milliseconds [11].

The second fusion approach, decision fusion is the process of first generating separate estimations fusing them into one final estimation. Each estimation from multiple modalities can be independently generated using separate models and the results are joined using a multitude of possible methods. In this case, the fusion of prediction obtained from various modalities becomes easy compared to feature-level fusion, since the prediction resulting from multiple modalities usually have the same form of data. Another advantage is that, each of every modality can utilize its best suitable model to learn its corresponding features. Among the notable decision-level fusion methods in continuous affect recognition is linear regression [11] [6] has been implemented in several AVEC challenge to fuse the estimation from each modality. Other than linear regression, method such as SVR [7], random forests [24] or Kalman filters based [8] [9] has been used to combine prediction in decision fusion process.

However, although such feature and/or modelling approach successfully predicting affect in a continuous way, a systematic understanding on what is the relationship between each modality in multimodal fusion is still less frequently explored. Each of the modeling approach reviewed usually does not give a definite function for the fusion rule. On top of that, it is not always possible to design a model that suits each modality because of the complexity. Therefore, the need to develop a model that can approximate the relationship between the predictions based on a measured set of data without a need of prior knowledge about the modality that produced the experimental data is desired.

III. AFFECT ESTIMATION SYSTEM

Figure 1 show the overview of the proposed system. We first perform SVR modelling for the continuous affect recognition in unimodal setting (audio, video and text modality) using different features. Once the unimodal estimation of each affect are optimized, we then incorporate it with linear regression, multi gene GP fusion as well as exponent weighted decision fusion strategies to investigate its robustness in the multimodal setting settings. In order to evaluate the proposed approach, three fusion rule is evaluated by comparing it to the widely-used decision fusion rules in affect regression methods.

A. Unimodal affect estimation

For the transparency of this experiments, we utilized SEWA dataset [6], the first and only audio-visual behaviour in-the-wild [25]. SEWA, stands for *Sentiment Analysis in the Wild* consists of audio-visual recordings of subjects showing spontaneous and natural behaviors. Audiovisual were recorded during dyadic interactions, 32 pairs in total, using

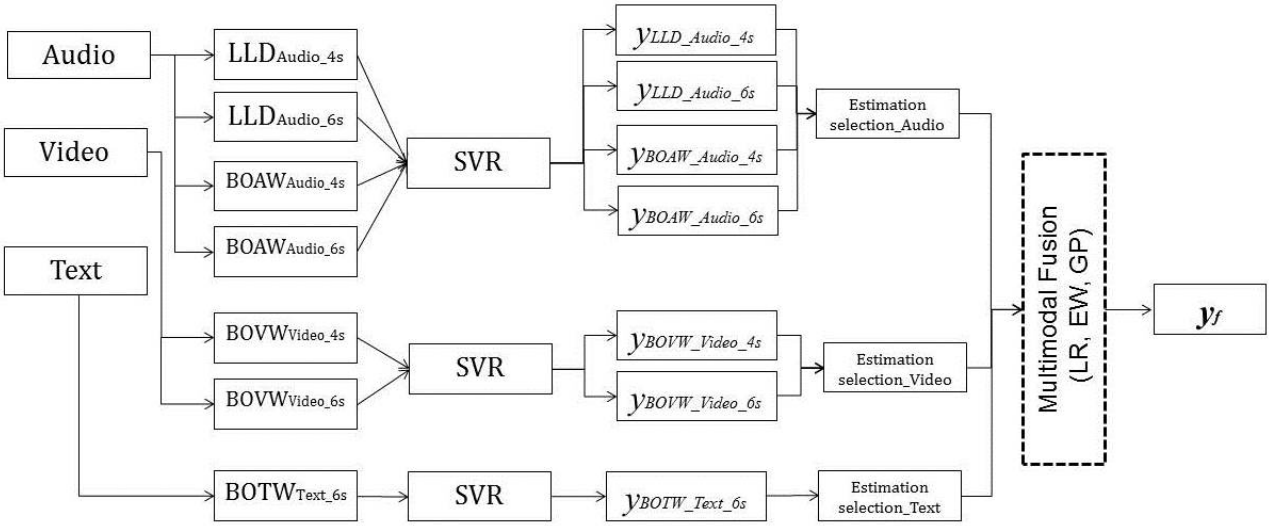

Figure 1. Overview of the proposed system. Fusion of the predictions of the three modalities: audio, video and text. Subscript indicate computation of features over second. For example, *Audio_4s* means audio feature were computed over segments of 4 seconds.

standard webcams and microphones from the computers in the subjects offices or homes, without any intervention of specific speakers, headphone, or sensors. The data is provided in three partitions (Training, Development, and Test), where both partners of one video chat appear in the same partition. The data is labeled in three affective label, namely Arousal, Valence and Likability, manually annotated by 6 annotators (3 female, 3 male), all were German native speakers, using a joystick. The dataset is provided together with a set of pre-calculated features which will be incorporated into the model. To avoid repetition, we refer to for details [6] on the feature extraction procedures for all features in the next subsection.

1) Audio: For the audio modality, the database provide two sets of audio features, namely Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) LLDs: *functionals* extracted using openSMILE toolkit [26] and *bag-of-audio-words* (BOAW): extracted using openXBOW toolkit [27]. The latter features, BOAW is inspired by text mining research area and commonly used in document classification (bag-of-words). Using bag-of-words principle, LLD on certain segment is quantised using a codebook of $'$*audio words*$'$, then histogram of audio words is produced on a corresponding segment. The important parameter that need to be taken into consideration are the codebook size, i. e., the number of audio words set into the framework. In the baseline features, the codebook size is set to 1000, then standardised to zero mean and unit variance prior to vector quantisation. Both segment-level eGeMAPS LLDs and BOAW types were computed over segments of 6 seconds. In total, the audio baseline feature sets with functionals contain 88 features, while the BoAW features contain 1 000 features.

2) Video: As for video modality, the database provide two sets of video features: facial features and *bag-of-video-words* (BOVW) features. The facial features include face orientation (pitch, yaw, and roll - 3 features), pixel coordinates for 10 eye point (20 features) and pixel coordinates for 49 facial landmarks (98 features). In total, facial has feature value of 121. Then, each of the features is standardised to zero mean and unit variance on frame level. The latter features, BOVW features were computed on top of standardised facial features with a codebook size of 1 000. The facial features have been extracted for each video frame using the Chehra face tracker [28] while BOVW features is extracted using openXBOW toolkit [27].

3) Text: Experimenting with text based features is quite new approach in continuous emotion recognition. In this paper, a *bag-of-text-words* (BOTW) feature representation is generated based on the transcription of the speech. By taking into account only the terms with at least two occurence, the results in a dictionary contained 521 words. Therefore, openXBOW toolbox with a codebook size of 521 is used, resulting 521 features of BOTW.

4) Regression models: Separate Arousal, Valence and Likability predictions are obtained from individual modalities as described in the last subsection. The regression task is performed using linear SVR provided with the liblinear library [29]. Unimodal predictions are first obtained from the five feature sets provided in SEWA dataset (LLD, BOAW, facial landmark video, BOVW and BOTW). We conducted additional experiments by scaling and shifting the unimodal estimation according to the training label in order to correct the bias and scaling issues. These unimodal estimations are used as an input in multimodal estimation in the proposed

late fusion approaches.

B. Multi-modal affect estimation

In this section, we leverage the individual advantage of each modality by combining them in a multimodal fashion manner. It is also to the examine the possibility to construct prediction equation of each affect. Each of the initial prediction from audio, video and text is denoted as y_A, y_V, y_T, and become and input in the following subsequent multimodal fusion.

1) Linear Regression (LR): LR attempts to model the relationship between two variables by fitting a linear equation to observed data. In the case of continuous affect estimation, regression coefficients γ need to be weighted separately according to contribution of each modality towards affects. Equation 1 is the linear regression formula where γ and ϵ_m are the regression coefficients and bias term computed in development sets, and $\boldsymbol{y_f}$ is the final fused prediction.

$$\boldsymbol{y_f} = \gamma_A(y_A) + \gamma_V(y_V) + \gamma_T(y_T) + \epsilon_m \tag{1}$$

2) Exponent Weighted Decision Fusion: In this paper, we leverage the exponent weighted decision fusion approach by Kim et al. [30] in regression manner, where its validation accuracy represent by the correlation from development dataset. Suppose an SVR model with a best correlation, C where C_A is the best correlation for audio, C_V is for the best correlation for video and C_T is the best correlation for text, will provide an initial prediction for each modality. Then, the final ensemble of our initial prediction from each of the features in the exponent weighted decision fusion become:

$$\boldsymbol{y_f} = (C_A)^q(y_A) + (C_V)^q(y_V) + (C_T)^q(y_T) \tag{2}$$

where a decision weight in terms of $(C)^q$ reflects the significance of initial prediction according to each modality and an exponent q is a hyper-parameter tuning. Here, the value of q is found by a simple uniform search: scanned over [-50:0.1:150] then selected to provide the maximum correlation after the fusion. The scanning procedure and the corresponding correlation values for the selected q are illustrated in Figure 2.

3) Genetic Programming (GP): GP is inspired from by biological evolution in nature. In order to improve their genomes, the evolution begins by iteratively process randomly generated solutions (individuals). The objective function are the individual fitness. Iteratively, the reproduction generation is constructed by *survival-of-the-fittest* individuals, by employing *crossover* and *mutation*. In brief, *crossover* is the recombination of parent genome to produce child genome while *mutation* is a possible modification that happens to child genome. The iterative process stop when the maximum number of generations is reached or the best fitness is visited.

Multi gene GP is the results of combination of GP, multiple gene and linear regression. In other words, each solution is formed by a linear combination of one or more such functions, called genes. A graphical representation of formulation with three input variable, x_1, x_2, x_3 as shown in Figure 3. As can be seen, the structure of this model contain nonlinear terms such as $\sin$, $\exp$, $\cos$, and the overall model is a weighted linear combination with respect to each coefficient.

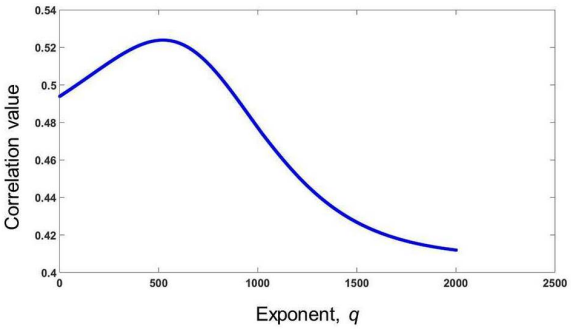

Figure 2. C_{corr} values as an exponent q is scanned in the exponentially weighted decision fusion. Noted that when proper q was selected, it gives maximum C_{corr} in the development sets.

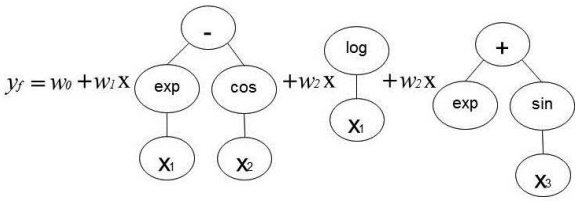

Figure 3. Graphical formula with three input variable.

From Figure 3, the solution is in the from of:

$$y_f = w_0 + w_1 g_1(x) + w_2 g_2(x) + ... + w_n g_n(x) \tag{3}$$

where n is the number of gene. Each gene is applied to the feature matrix, producing $N \times 1$ vector where:

$$\boldsymbol{y_f} = [\mathbf{1} \boldsymbol{g_1 g_2 ... g_n}] \cdot \boldsymbol{w} \tag{4}$$

with $\mathbf{1}$ being $N \times 1$ vector of ones. The output y of the whole solution is then given by formula:

$$\boldsymbol{y_f} = \boldsymbol{G} \cdot \boldsymbol{w} \tag{5}$$

The optimal coefficient vector w^* can then be found using the least-squares estimation with respect to the true target vector y

$$\boldsymbol{w}^* = (\boldsymbol{G}^\intercal \boldsymbol{G})^{-1} \boldsymbol{G}^\intercal \boldsymbol{y} \tag{6}$$

IV. EXPERIMENTAL RESULTS

This section empirically evaluates the proposed algorithm in SEWA dataset.

A. Experimental Set-ups and Evaluation Metrics

We reported the performance of our proposed architecture based on C_{corr} [6] metric:

$$C_{corr} = \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \tag{7}$$

where ρ is the P_{corr} between two time series (e.g: prediction and gold-standard); μ_{haty} and μ_y are the means of each time series; and $\sigma_{\hat{y}}^2 x$ and σ_y^2 are the corresponding variance. Here, the value of C_{corr} is within the range of $[-1, 1]$, where ± 1 represents perfect concordance and discordance while 0 means no concordance between two time series.

B. Affect Estimation in Unimodal Modality

Table I displays the results in terms of C_{corr} obtained from unimodal modality of SVR on the development sets of SEWA. On Arousal, the best performance is achieved with video modality, more specifically on BOVW features. In Valence, the highest results of C_{corr} is taken from audio modality, more specifically on BOAW features. Whereas in Likability, the highest is from text modality, more specifically on BOTW features.

C. Affect Estimation in Mutimodal Modality

1) Linear Regression: In the first experiment, we build fusion model by a simple linear regression of the predictions obtained on the development partition, using Equation 1 in Weka 3.7 [31] on top of MATLAB with the same setting as mentioned above. Equation 8 9 10 shows the final equation according to each affect, respectively.

$$\boldsymbol{y_{fAR}} = 0.956y_A + 0.425y_V + 0.404y_T - 0.0915 \tag{8}$$

$$\boldsymbol{y_{fVA}} = 0.299y_A + 0.302y_V + 0.249y_T - 0.0116 \tag{9}$$

$$\boldsymbol{y_{fLI}} = 0.144y_A + 0.202y_V + 0.348y_T - 0.0069 \tag{10}$$

2) EW: For the second experiment using EW, the best exponent q is obtained from the first section, then the same q is applied in the second section. Each of the q is scanned in the range of [-50:0.1:150] and validated by using Equation 2 thus selected to provide the maximum performance after the fusion. Equation 11 12 13 shows the final equation according to each affect, respectively.

$$\begin{aligned}\boldsymbol{y_{fAR}} = (0.328)^{6.6}(y_A) + (0.455)^{6.6}(y_V) \\ +(0.407)^{6.6}(y_T)\end{aligned} \tag{11}$$

$$\begin{aligned}\boldsymbol{y_{fVA}} = (0.401)^{2.1}(y_A) + (0.389)^{2.1}(y_V) \\ +(0.386)^{2.1}(y_T)\end{aligned} \tag{12}$$

$$\begin{aligned}\boldsymbol{y_{fLI}} = (0.175)^{1.5}(y_T) + (0.249)^{1.5}(y_V) \\ +(0.390)^{1.5}(y_T)\end{aligned} \tag{13}$$

3) GP Modelling: Three multi gene GP models are established in this paper for predicting the continuous affect for each of affect dimension, respectively. GPTIPS2 developed by Searson et al., [32] was used for model development. The parameters that were set in the multi gene GP algorithms include: a population size of 250, a tournament size of 20, maximum number of genes allowed in an individual 8, function set {+,-,x,/,sin,cos,exp} and terminal sets $\{y_A, y_V, y_T\}$. The resulting prediction equation discovered by a multi gene GP model according to each affect is reported as follows:

$$\begin{aligned}\boldsymbol{y_{fAR}} = 5.5e^{-4}\sin(27y_V^3) + 0.31e^{y_T} - 200y_V^3 y_T^9 \\ +3.4y_A(y_A^3 + y_V y_A^2 + y_V) \\ -0.025e^{(-3y_V)}\sin(9.5y_T) \\ +0.1y_A^{1/4} - 0.1y_T^3 - 0.33\end{aligned} \tag{14}$$

$$\begin{aligned}\boldsymbol{y_{fVA}} = 0.057\sin(16y_V y_T) - 0.32\sin(y_A y_V y_T) \\ +0.13\sin(y_V^2(y_A + 7.8)) \\ +0.12y_T^2 e^{-y_T}(y_A + 7.8) \\ +0.16y_A(e^{-y_T})^{1/2}(y_V + 7.5)^{y_A} + 4.5e^{(-3)}\end{aligned} \tag{15}$$

$$\begin{aligned}\boldsymbol{y_{fLI}} = 0.15y_V + 0.15y_T + 0.15\sin(\sin(y_A)) \\ -0.18|y_T| + 9.3y_V^4 y_T - 3.4e^3 y_V^7 y_T \\ +0.36y_A^2 - 6.5y_V^3 + 79y_V^5 \\ +399y_A y_V^3 y_T + 5.6e^3 y_A y_V^5 y_T - 6.9e^{(-3)}\end{aligned} \tag{16}$$

Table I
UNIMODAL PERFORMANCE USING C_{corr} ON THE DEVELOPMENT SET

Modality	Features	$\mathbf{C_{corr}}$		
		Arousal	*Valence*	*Likability*
Audio	LLD_4s	.380	.338	.062
	LLD_6s	.342	.274	.089
	BOAW_4s	.325	.390	.032
	BOAW_6s	.327	**.392**	.104
Video	BOVW_4s	**.453**	.384	.172
	BOVW_6s	.370	.340	.132
Text	BOTW_6s	.364	.382	**.317**

4) Performance Comparison: Closer inspection on Table II shows that in most cases, decision fusion gives better results than feature fusion method. We suspect that, given the fact that features are extracted in the same manner, there are tendency of the features have similar or nearly similar distribution, which makes one of them is redundant, when performing feature fusion. Our finding confirms that in Arousal and Valence dimension, the multimodal system in Table II performs better than the best unimodal system in Table I. The new dimension, Likability however performs the best result on unimodal system on text modality. In LR, overall we have achieved a better performance for estimating Arousal than Valence and Likability consistent with existing linear modeling frameworks, as shown in Equation 8. From this Equation, it shows that audio modality gives the highest

weighting factors which contribute significantly to the higher performance in Arousal. However, when it comes to Valence and Likability dimensions, there seems to be relatively lower performance in estimating those two affect, most likely due to non-linearities in the relationship between the features and those two affect ratings. We further investigate those non-linearity behavior on those two affect ratings using EW and multi gene GP approach. By using EW, the system performance is further improved upon using non-linearity behavior in estimating Valence and Likability. By having proper q selection in EW approach gives significant gain in C_{corr} results for Valence and Likability, from 0.507 to 0.549 and 0.215 to 0.231 respectively. However, when we compare the results of Likability with the baseline approach, the baseline approach has slightly higher performance than our proposed multi gene GP approach. This may be due to the fact that the SVR models in the first stage have already fit well for the Likability with the original feature vector. Notably, the formula produced by multi gene GP seems to be more compact than yielded by LR and EW, which yields the best results on C_{corr} in Valence and Likability dimensions, by 0.559 and 0.257 respectively. Looking at the performance increase, we can conclude that a model with simple structure is incapable of describing such complex functional mapping in a satisfactory manner. A lower C_{corr} on multi gene GP and EW instead of LR confirms the assumption that evolution of Arousal dimension are linear in time, consistent with the assumption in [9].

Recent published results by Chen et al. [33] in validation set achieved achieved higher results where additional features and multitask learning were used. However, it is not strictly comparable because 2-fold cross-validation protocol was used in our results. Our focus is on the relationship between multiple modalities where proposed methods can give the mathematical expressions.

Table II
MULTIMODAL PERFORMANCE USING C_{corr} ON 2-FOLD CROSS VALIDATION

Fusion Type	**Fusion Method**	**C$_{corr}$**		
		Arousal	*Valence*	*Likability*
Feature [6]	Concatenate	.525	.507	.235
Decision	LR	.592	.507	.215
	EW	.440	.549	.231
	multi gene GP	.572	.562	.258
Chen et al. [33]	Multi-task learning	.750	.776	.579

V. CONCLUSION

This work investigates the possibility of employing different modeling approach, including LR, EW and multi gene GP, for constructing prediction fusion rules at the decision level in continuous affect estimation in-the-wild. To train and verify these multimodal fusion approaches, a dataset containing text and audiovisual recording is used. LLD, BOAW, BOVW and BOTW features are extracted respectively from audio, video and text modality. Then SVR have been employed to estimate the initial prediction of each affect. In fusion stage, the best initial prediction from unimodal modality is selected, and LR, EW and multi gene GP is being employed to construct the prediction rules. Experimental results shows that the prediction equation of multi gene GP shows better modeling outcome than the benchmark results, outperform the baseline approach in all affect dimension. Result comparison with other benchmark method such as LR shows that multi gene GP significantly improve the performance in Valence and Likability dimension. It confirms our initial assumption that there exists non-linearity behavior in those two affect dimension. As for Arousal dimension, LR perform better than baseline, EW and multi gene GP fusion approach. It shows that Arousal dimension is generally linear in time.

It should be mentioned here that the conclusion might not completely correct due to the use of the dataset. Although it is a very good dataset, however, the total number of samples is still limited and the features and first baseline regression method is very basic. In our future work, we would like to use more multimodal datasets and features to improve the system and verify these assumptions.

REFERENCES

[1] 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.

[2] 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.

[3] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2015 – The 5th International Audio/Visual Emotion Challenge and Workshop," *Proceedings of the 23rd ACM International Conference on Multimedia, MM 2015*, pp. 1335–1336, 2015.

[4] F. Ringeval, E. Marchi, C. Grossard, J. Xavier, M. Chetouani, D. Cohen, and B. Schuller, "Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-September-2016, 2016, pp. 1210–1214.

[5] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels, "Ensemble methods for continuous affect recognition: Multimodality, temporality, and challenges," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '15. New York, NY, USA: ACM, 2015, pp. 9–16.

[6] F. Ringeval, "AVEC 2017 Real-life Depression, and Affect Recognition Workshop and Challenge," *AVEC workshop*, vol. 38, no. 1, pp. 4–5, 2017.

[7] 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.

[8] 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.

[9] 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.

[10] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[11] 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.

[12] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[13] A. Metallinou, Z. Yang, C. chun Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations," *Language Resources and Evaluation*, vol. 50, no. 3, pp. 497–521, 2016.

[14] F. Povolny, P. Matejka, M. Hradis, A. Popková, L. Otrusina, P. Smrz, I. Wood, C. Robin, and L. Lamel, "Multimodal emotion recognition for avec 2016 challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: ACM, 2016, pp. 75–82.

[15] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012: the continuous audio/visual emotion challenge," *Proc. 14th Int'l Conf. Multimodal Interaction Workshops*, pp. 449–456, 2012.

[16] T. Baltrusaitis, N. Banda, and P. Robinson, "Dimensional Affect Recognition using Continuous Conditional Random Fields," *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, 2013.

[17] 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.

[18] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '15. New York, NY, USA: ACM, 2015, pp. 49–56.

[19] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, April 2011.

[20] E. Pei, L. Yang, D. Jiang, and H. Sahli, "Multimodal dimensional affect recognition using deep bidirectional long short-term memory recurrent neural networks," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sept 2015, pp. 208–214.

[21] 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," *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge - AVEC '15*, no. October 2015, pp. 73–80, 2015.

[22] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, and J. Epps, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '15. New York, NY, USA: ACM, 2015, pp. 41–48.

[23] H. A. Osman, H. Dong, and A. E. Saddik, "Ubiquitous biofeedback serious game for stress management," *IEEE Access*, vol. 4, pp. 1274–1286, 2016.

[24] P. Cardinal, N. Dehak, A. L. Koerich, J. Alam, and P. Boucher, "Ets system for av+ec 2015 challenge," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '15. New York, NY, USA: ACM, 2015, pp. 17–23.

[25] D. Collection, "SEWA Database: Data Collection, Annotation and Release."

[26] 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.

[27] M. Schmitt and B. W. Schuller, "openxbow - introducing the passau open-source crossmodal bag-of-words toolkit," *CoRR*, vol. abs/1605.06778, 2016.

[28] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1859–1866.

[29] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[30] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee, "Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 427–434.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[32] D. Searson, D. Leahy, and M. Willis, "GPTIPS: An open source genetic programming toolbox for multigene symbolic regression," *Proceedings of the International of the Multi-Conference of Engineers and Computer Scientists*, vol. I, pp. 17–20, 2010.

[33] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '17. New York, NY, USA: ACM, 2017, pp. 19–26.