INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Multimodal Prediction of Affective Dimensions via Fusing Multiple Regression Techniques

D.-Y. Huang[1], *Wan Ding*[2], *Mingyu Xu*[1], *Huaiping Ming*[1], *Minghui Dong*[1], *Xinguo Yu*[2], *Haizhou Li*[3]

[1]Human Language Technology Department, Institute for Infocomm Research, Singapore
[2]National Engineering Research Center for E-Learning, Central China Normal University
[3]ECE Department, National University of Singapore, Singapore

(huang,mingyu,minghp,mhdong)@i2r.a-star.edu.sg, (dingwan,xgyu)@mails.ccnu.edu.cn, haizhou.li@nus.edu.sg

Abstract

This paper presents a multimodal approach to predict affective dimensions, that makes full use of features from audio, video, Electrodermal Activity (EDA) and Electrocardiogram (ECG) using three regression techniques such as support vector regression (SVR), partial least squares regression (PLS), and a deep bidirectional long short-term memory recurrent neural network (DBLSTM-RNN) regression. Each of the three regression techniques performs multimodal affective dimension prediction followed by a fusion of different models on features of four modalities using a support vector regression. A support vector regression is also applied for a final fusion of the three regression systems. Experiments show that our proposed approach obtains promising results on the AVEC 2015 benchmark dataset for prediction of multimodal affective dimensions. For the development set, the concordance correlation coefficient (CCC) achieves results of 0.856 for arousal and 0.720 for valence, which increases 3.88 % and 4.66 % of the top-performer of AVEC 2015 in arousal and valence, respectively.

Index Terms: emotion recognition, arousal-valence, multimodal prediction, recurrent neural network

1. Introduction

Emotional signals such as facial expression, speech and physiological status (heart beat rate, skin conductance, etc) are becoming more and more easily available due to technological advancements in recent years. Research for multimodal affect dimension (arousal/valence [1]) prediction has been very active. Most studies focus on two directions: multimodal feature extraction and predictor design based on regression techniques. For feature extraction, audio, visual and physiology are the three modalities that people are most interested in [2–6]. Both handcrafted rule-based and deep learning methods are applied to extract emotion features. The studies show that the performance for emotion recognition usually can be improved by score level fusion among multimodal features [7, 8]. This suggests that the features of different modalities are complementary.

For regression predictor improvement, there are two basic ideas: fine tuning of regression models, and decision fusion of multiple regression techniques. An example of modeling improvement is Chao et al.'s [9] two new techniques to the long short term memory - recurrent neural network (LSTM-RNN) model for continuous emotion recognition. First, they modified the loss function to increase error tolerance, thus making the model more robust to label noise and better able to make stronger correlation between predicted values and labels. Second, the temporal pooling layer was added to increase the diversity of features presented to a forward prediction architectures. Partial Least Squares (PLS) have also been applied for different tasks of speaker state detection [11,12]. An example of decision fusion of multiple regression techniques is Ringeval et al.'s [10] study of Linear support vertor regression (SVR) and neural networks (NN) (feed-forward network, LSTM-RNN, and so on). The fusion of NN prediction and SVR prediction demonstrates the complementary nature of these two predictors, i.e., SVR performs best on some mono-modal subtasks whereas NN performs best on others.

In this paper, we studied the fusion of SVR, PLS and DBLSTM-RNN as predictors for multimodal prediction of affect dimensions. The organization of the paper is follows: an overview of system is given in Section 2; A proposed method will be presented in details in Section 3; The experimental results will be shown to support our proposed method in Section 4; Finally, a conclusion and perspective will be given in Section 5.

2. Overview of System

We select SVR, PLS, DBLSTM-RNN methods not only because they are three types of widely used regression techniques for emotion recognition systems, e.g., Partial Least Squares (PLS) [11], linear kernel Support Vector Regression (linear SVR) [8, 10], Deep Bidirectional Long Shot Term Memory Recurrent Neural Network (DBLSTM-RNN) [5–7, 9, 10]), but also based on an empirical assumption. We assume that there exists three types of emotional reactions driven by different biological and psychological mechanisms.

The first type is the instant strong action (e.g. impulsive response). Emotional expression is immediate and directly related to emotional status. We apply linear kernel support vector regression method to model this relationship. The second type is the instant moderate action (e.g. impulsive but brief response). In this second type, the emotions are not expressed as obviously nor sustained for as long as the first type, and actions are affected much more by the subject-specific factors such as social ability. We presume that the subject-specific information can be filtered by mapping data to new spaces and applying partial least squares regression to model this type of relationship [12]. The third type relates to the action on context situation (e.g. impact of historic emotions on future emotion). We apply the Long-Short Term Memory Recurrent Neural Network (LSTM-RNN) as it can capture the cues at long intervals (more than 100 time steps) without suffering from the deep learning network vanishing gradient problem [13].

http://dx.doi.org/10.21437/Interspeech.2017-1088

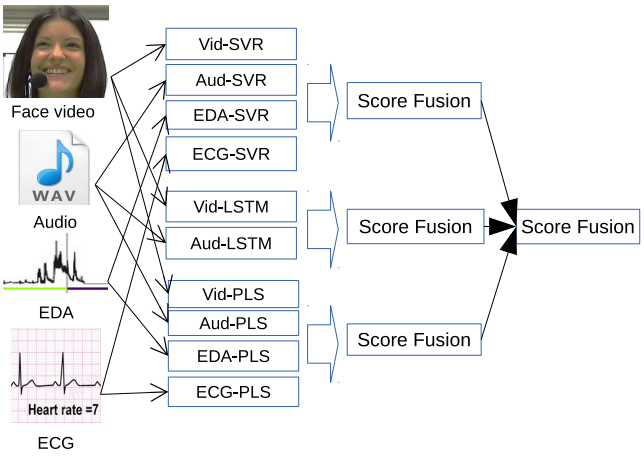

Figure 1: *Overview of the system architecture. We did not apply DBLSTM-RNN to EDA and ECG because the RNN architectures for these two modalities needs further study.*

Table 1: *The input feature sets from multi-modalities*

Modality	**Feature set**
Audio	eGeMAPs [2]
Video	LGBP-TOP [3] Facial Geometry Shapes [10]
Electrocardiography	Original signal [4] HRHRV [4]
Electrodermal activity	Original signal [4] Skin Conductance Response (SCR) [4] Skin Conductance Level (SCL) [4]

We first build three subsystems by applying each regression technique to perform multimodal affect dimension prediction with a post-processed score level fusion of four modalities (audio, video, electrodermal activity and electrocardiogram). We then use another linear support vector regression for a final fusion of the three subsystems. We conduct our approach to AVEC 2015 benchmark dataset for prediction of multimodal affective dimensions. For the development set, the concordance correlation coefficient (CCC) achieves results of 0.856 for arousal and 0.720 for valence, which are better than the top-performer of AVEC 2015 obtaining 0.824 for arousal and 0.688 for valence [7].

3. Proposed Method

3.1. Multimodality Features

We applied eight feature sets extracted from four modalities for multimodal affect dimension prediction. The four modalities and eight feature sets are listed in Table 1.

3.1.1. Audio Features

Audio features are obtained from the extended version of Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [2]. eGeMAPS is a small expert-knowledge crafted set of only 88 features that have shown high robustness for modeling speech emotion [8, 10]. The acoustic low-level descriptors (LLD) of eGeMAPS covers spectral, cepstral, prosodic and voice quality information. Besides LLDs, eGeMAPS also provides other statistics such as arithmetic mean and coefficient of variation.

3.1.2. Video Features

Two types of facial descriptors are used. The first is the Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [3]. The LGBP-TOP are computed by splitting the video into spatio-temporal video volumes. Each slice of the video volume extracted along 3 orthogonal planes is first convolved with a bank of 2D Gabor filters. The resulting Gabor pictures in three directions are then divided into blocks and the LBP operator is applied to each block followed by the concatenation of the resulting LBP histograms from all the blocks. For geometric features 49 facial landmarks are tracked and aligned based on stable points such as eye corners and nose points. The shape descriptors of facial features (eyes, mouth, etc) such as Euclidean distances and angles between points are then calculated and the concatenation is taken as geometric features.

3.1.3. Electrocardiography Features

Electrocardiography (ECG) records the electrical activity of the heart. The original ECG signals computes 19 features: the spectral entropy, slope, mean frequency and 6 spectral coefficients; the four first statistical moments; the zero-crossing rate; the non-stationary index; the normalized length density; the power in high frequency (HF, 0.15-0.4 Hz) and low frequency (LF, 0.04-0.15 Hz); and the LF/HF power ratio. Besides heart rate (HR), its measure of variability (HRV) is also extracted; this includes two first statistical moments, the arithmetic mean of rising and falling slope, and the percentage of rising values.

3.1.4. Electrodermal Activity Features

Electrodermal activity (EDA) refers to the variation of the electrical characteristics of the skin of the human body. It reflects a rapid, transient response named skin conductance response (SCR), and also a slower, basal drift named skin conductance level (SCL) [4]. For feature extraction, a 3rd order Butterworth filter may be used to estimated SCL (0.0-0.5 Hz) and SCR (0.5 -1.0 Hz). Eight features are then computed for each descriptor: the four first statistical moments from the original time-series and their first order derivatives.

3.2. Regression Methods

We used three regression methods for dimensional affect emotion recognition. The first is Linear Support Vector Regression [14]. Given training dataset x and y, the linear SVR computes the linear regression model W and b as

$$y - \epsilon \leq Wx + b \leq y + \epsilon \tag{1}$$

where b denotes the factors that can not be modeled as linear, ϵ denotes the error tolerance, which is defined before training.

The second is partial least squares regression. PLS models the correlation between emotion status and emotional actions as

$$x = Qr + \epsilon_x \tag{2}$$

$$y = Ps + \epsilon_y \tag{3}$$

where Q and P are subject-specific transform matrices, r and s are subject-independent latent variables and ϵ_x and ϵ_y are residual terms which cannot be modeled by the linear model. Solving Q and P leads to the regression model

$$y = Bx + \epsilon \tag{4}$$

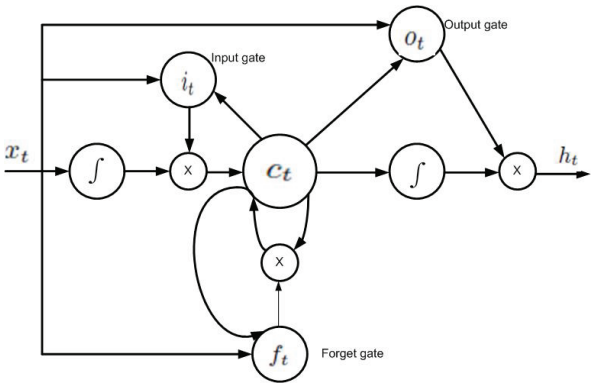

Figure 2: *The architecture of LSTM memory block.*

where B is the regression matrix that is a function of Q and P, and ϵ is the regression residual.

The third is deep BLSTM-RNN. Here deep denotes the number of hidden layers is more than one. Instead of the sigmoid-shaped active functions, the LSTM uses a set of activation functions called memory blocks to avoid the vanishing gradient problem. For one memory block in a hidden layer, the input at time step t is the output of the previous layer x_t, h_{t-1}, and the output of current layer at time t-1. The input interacts with four sub-components and the activation outputs determine the final memory block output. These four main elements are: an input gate, an output gate, a memory cell, and a forget gate. The self-recurrent connection of weight 1.0 is used in memory cell to ensure that, blocking any outside interference, the state of a memory cell can keep constant from one time step to another.

The input gate is able to let incoming signals to change the state of the memory cell or block it. The output gate can make the state of the memory cell to have an influence on other neurons or completely block it. The forget gate can modify the memory cells self-recurrent connection, letting the cell to remember or forget its previous state, if necessary. The activation and output of memory blocks are as follows:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{5}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{6}$$

$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{7}$$

$$c_t = i_t * \tilde{c}_t + f_t * c_{t-1} \tag{8}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{9}$$

$$h_t = \sigma * tanh(c_t) \tag{10}$$

where x_t is the input; W, U and V are weight matrices; and b terms are the bias vectors. σ is the sigmoid function; $\tilde{c}$ terms are the candidate states of cells at time step t; i_t, f_t, c_t, and o_t denote activation coefficients of the input gate, forget gate, memory cell and output gate. h_t is the final memory block output. In our approach we use bidirectional LSTMs for emotion recognition. The BLSTM network has two parallel layers propagating in forward and backward directions.

3.3. Post Processing

A post-processing chain as in [6] is applied to compensate for delays time in the ratings, scaling and bias in predictions, and data noise. The post-processing chain has four steps: median filtering, centering, scaling and time-shifting. The best value of parameters (filtering window size, time delay, etc) were optimized by maximizing the measure of performance - Concordance Correlation Coefficient (CCC) - on the development partition with the model, learned on the training partition.

4. Experiments

4.1. System Setup

4.1.1. The AVEC 2015 Dataset

We evaluate the performance of our approach on the AVEC2015 dataset, which adopts the RECOLA corpus [15]. In the context of remote collaborative work, this corpus was collected to study socio-affective behaviours from multimodal data. Multimodal signals were synchronously recorded from 27 French-speaking subjects, with each video a length of 5 minutes. For the annotation of the dataset, six gender balanced evaluators are asked to perform time-continuous ratings (40 ms binned frames) of emotional arousal and valence. Golden standard is then computed based on the intra class correlation coefficients. Finally, the dataset was partitioned into speaker independent subsets for training, development (validation) and testing. The feature sets such as eGeMAPS and LGBP-TOP are also provided in the AVEC dataset as baseline features.

4.1.2. Regression Methods Implementation

We used the Liblinear toolkit [16] for linear SVR implementation, the Matlab toolbox for PLS implementation and CURRENNT toolkit [17] for DBLSTM-RNN implementation. The complexity of linear SVR is 1e-5, and the number of components of PLS is 8. Other parameters of these two regression systems use default values. For DBLSTM-RNN, we tested four different architectures for each modality. The four architectures are 96-96, 128-128, 192-192, 256-256. A learning rate of 1e-6 and batch size of 2 sentences (each sentence denotes the feature sequence extracted from one video clip) were used to train the models. Early stopping was used to avoid overfitting. To find the best LSTM model, we ran training on each architecture ten times with network weights randomly initialized each time. The best results on the dev dataset is taken for the next fusion step.

4.2. Performance Evaluation

The AVEC 2015 challenge measure is the Concordance Correlation Coefficient (CCC), which incorporates the Pearsons correlation coefficient (CC) with the square difference between the mean of the two compared time series.

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{11}$$

where ρ refer to the Pearson correlation coefficient between two time series (e. g., gold-standard and prediction), μ_x and μ_y are the mean value of each time series, and σ_x^2 and σ_y^2 are the variance of each . Therefore, predictions that are well correlated with the gold standard but shifted in value are penalized in proportion to the deviation. The experimental results are presented in Table 2 and Table 3. Table 3 summarized the results of fusion across different features and regression techniques. The results suggest that each of systems represented by different regression techniques contributes to the overall performance, with the LSTM based system contributing the most.

We note that the concordance correlation coefficient (CCC) of the proposed system achieves results of 0.856 for arousal and 0.720 for valence, which outperforms the top-performer of

Table 2: *The CCC of mono-modality sub-systems*

	Arousal			**Valence**		
	SVR	PLS	LSTM	SVR	PLS	LSTM
Audio	0.796	0.776	**0.800**	0.455	0.447	**0.512**
Video-Ap	**0.483**	0.452	0.454	0.474	0.457	**0.547**
Video-Ge	**0.379**	0.371	0.377	0.612	**0.619**	0.554
ECG	0.288	0.265	-	0.153	0.140	-
HRHRV	0.382	0.371	-	0.293	0.256	-
EDA	0.073	0.065	-	0.194	0.134	-
SCL	0.068	0.057	-	0.166	0.145	-
SCR	0.073	0.067	-	0.085	0.075	-

Table 3: *The CCC of system fusion across different features and regression techniques.*

	SVR(S)	PLS(P)	LSTM(L)	S+L	P+L	S+P	S+P+L
Arousal	0.821	0.801	0.846	0.855	0.852	0.822	**0.856**
Valence	0.683	0.687	0.716	0.716	0.719	0.692	**0.720**

AVEC 2015 by 3.88% and 4.66% in arousal in arousal and valence [7], respectively.

5. Conclusions

In this paper we present a multimodal approach for affect dimension prediction based on decision fusion of three regression techniques (Linear SVR, PLS and DBLSTM-RNN). Improvements were observed from monomodal to multimodal fusion results, supporting the assumption that extracted features of audio, visual and physiological modalities are complementary. Results further show that the fusion of Linear SVR, PLS and DBLSTM-RNN regression models all contribute to the final model performance, with DBLSTM-RNN contributing the most. Our future work will be conducted in two directions. The first is to further study the decision fusion of different regression techniques. The second is to focus on developing DBLSTM-RNN models that can achieve better performances for multimodal emotion recognition.

6. References

[1] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 03, pp. 715–734, 2005.

[2] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

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

[4] 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.

[5] P. Khorrami, T. L. Paine, K. Brady, C. Dagli, and T. S. Huang, "How deep neural networks can improve emotion recognition on video data," *arXiv preprint arXiv:1602.07377*, 2016.

[6] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, S. Zafeiriou *et al.*, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[7] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 73–80.

[8] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016-depression, mood, and emotion recognition workshop and challenge," *arXiv preprint arXiv:1605.01600*, 2016.

[9] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 65–72.

[10] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "Av+ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 3–8.

[11] 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*. ACM, 2013, pp. 21–30.

[12] D.-Y. Huang, S. S. Ge, and Z. Zhang, "Speaker state classification based on fusion of asymmetric simpls and support vector machines." in *INTERSPEECH*, 2011, pp. 3301–3304.

[13] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks." *ICML (3)*, vol. 28, pp. 1310–1318, 2013.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[15] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[16] 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, no. Aug, pp. 1871–1874, 2008.

[17] F. Weninger, J. Bergmann, and B. Schuller, "Introducing currennt–the munich open-source cuda recurrent neural network toolkit," *Journal of Machine Learning Research*, vol. 16, no. 3, pp. 547–551, 2015.