Mutual Information-Based Selection of Audiovisual Affective Features to Predict Instantaneous Emotional State

Sudipta Paul†, Nurani Saoda†, S M Mahbubur Rahman†, *Member*, *IEEE*, Dimitrios Hatzinakos‡, *Fellow*, *IEEE*
†Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology, Dhaka-1205, Bangladesh
‡Identity, Privacy and Security Institute, Department of Electrical and Computer Engineering
University of Toronto, Toronto, ON, Canada, M5S 2E4
E-mail: {sudiptapaul.paul, saodacynthia}@gmail.com, mahbubur@eee.buet.ac.bd, dimitris@comm.utoronto.ca

***Abstract*—Automatic prediction of continuous level emotional state requires selection of suitable affective features to develop a regression system based on supervised machine learning. This paper investigates the performance of low-level dynamic features for predicting two common dimensions of emotional state, namely, valence and arousal instantaneously. Low-complexity features are extracted from audio and visual modalities independently and fused in the feature level. Features with minimum redundancy and maximum relevancy are chosen by using the mutual information-based selection process. The performance of frame-by-frame prediction of emotional state using the moderate length features as proposed in this paper is evaluated on spontaneous and naturalistic human-human conversation of SEMAINE database. Experimental results show that the proposed features selected by mutual information can be used for instantaneous prediction of emotional state with an accuracy higher than traditional audio or visual features that are used for affective computation.**

I. Introduction

Human can easily understand the instantaneous affective information conveyed by speakers during human-human conversation from multimodal cues. Automatic retrieval of affective information helps to develop artificial listener agents and emerging user-oriented technologies. For example, machines will be able to provide flexible performance under uncertain conditions, movie directors can change affective content of a video aiming at certain viewer groups, emotional state of drivers can be monitored to engage safety measures, and impact of television commercials or programs can be measured by judging the emotional state of the viewers remotely. Thus, automatic prediction of instantaneous affective state is becoming increasingly important in the recent years [1].

A. Related Works

Analysis of affective content is an interdisciplinary field involving research areas that includes computer vision, speech analysis, and psychology. To relate between measurable low-level features with corresponding affective state, certain models of emotion are required. Psychologists have used two major approaches, viz., categorical and dimensional to quantify the emotional states [2]. According to the categorical approach, the model of emotion was defined by Ekman [3], who grouped emotion into six basic categories including the happiness, sadness, anger, disgust, fear, and surprise. Situation arises where a small number of discrete categories may not reflect the complexity of emotional states. In this context, continuous emotional model reflect more subtle and context specific emotions avoiding boundaries. As a result, research in area of affective computing is shifting from categorical approach to dimensional approach. Wundt [2] introduced dimensional approach, wherein the emotion is divided into 3D continuous space - valence, arousal, and dominance. Valence represents the degree of pleasure, ranging from pleasant to unpleasant feelings. Arousal illustrates the activation level ranging from global feeling of dynamism to lethargy of an individual. Dominance characterizes the range of emotion from controlling sentiment to the controlled or submissive feelings. Dietz and Lang [4] have shown that the effect of the dominance dimension becomes visible only at points with

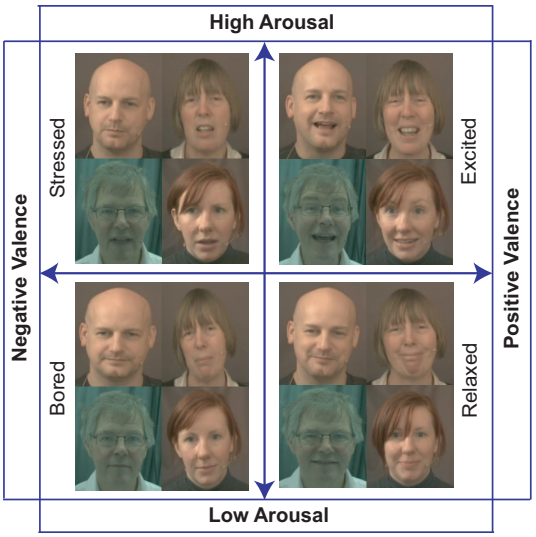

Fig. 1. Examples of facial appearances in the valence-arousal plane showing typical emotional states.

978-1-5090-4090-2/16/$31.00 ©2016 IEEE

ISBN 978-1-5090-4089-6

distinctly high absolute valence values. Greenwald *et al.* [5] have shown that the valence and arousal account for most of the independent variance in emotional responses. Fig. 1 shows a few examples of facial appearances in the valance-arousal plane. It is seen from this figure that different emotional states such as stressed, excited, bored or relaxed feelings can be recognized independent of the subjects by the ratings of the dimensions valance and arousal.

In general, the affective states of humans are estimated by extracting affective features from suitable sensors, and then relating between the low-level descriptors and high-level semantic meanings. Initial researches on affective content analysis confined in recognition of exaggerated expressions of prototypical emotions that are recorded in constrained environments. For example, studies carried out by of Chen and Huang [6], De Silva and Ng [7] involve the investigation audio and video signals to recognize six basic expressions introduced by Ekman [3]. Facial action coding system [8] is an well known approach of expression recognition that codify the muscle movements of faces. Spontaneous expressions are also recognized using the differentially expressive components of the facial images represented by orthogonal 2D Gaussian-Hermite moments [9]. Automatic detection of posed or spontaneous expressions with an acceptable accuracy is possible using monomodal static descriptors and static classifiers. But, challenges appear to solve for the problem of recognizing continuous level naturalistic emotions instantaneously those are displayed in our day-to-day life.

Recent trends for predicting continuous-level emotional state apply the process of feature extraction from different modalities, feature selection, and mapping of the selected feature to the semantic meanings using regression technique based on a suitable machine learning algorithm. Although features extracted from audio, visual or physiological signals have been attempted for prediction of emotional states, the former two-types of signals are preferred to the last-type of signal due to their easy accessibility and widespread availability. In order to analyze the dynamics of human emotion over a period of time, the regression algorithm requires the inclusion of dynamic features extracted from data of each modalities instead of commonly used static features. For example, Wollmer *et al.* [10] prescribed the use of low-level dynamic features extracted from the optical flow of facial video signals for automatically rating the emotional states. Dahmane and Meunier [11] used the Gabor energy texture descriptor of visual signals as the features of emotion. Another example of low-level dynamics is the use of quantized local-phase of texture patterns obtained from the three orthogonal planes (TOPs) of a video [12]. Nicolle *et al.* [13] used mid-level visual dynamic features including the head movement, face deformation, and local variations of face appearance. The performance of high-level static features such as the gaze direction, head tilt, smile intensity to obtain the continuous-level emotional meanings from a single image has been investigated in [14]. To remove redundant and irrelevant information in the regression process, suitable selection techniques such as those employ the correlation coefficients [13] and mutual information [15] are applied to certain features. In the multimodal analysis, results obtained from individual modality are required to be fused to predict the emotional states. In case the emotional state is predicted using the synchronized modality the feature-level fusion is suitable. On the other hand, the decision-level fusion is recommended for features extracted from asynchronous modalities. Various machine learning regression algorithms such as the long short-term memory network [10] and hidden Markov model [16] are applied to predict the continuous-level emotional state from the extracted features or the fused set of features.

B. Scope of Analysis

Existing methods of continuous-level affective computation use a relatively high number of features as well as the computational complexity of their extraction processes is relatively high. Prediction results of these features have been reported only for the entire set of test data, with performance comparisons are given only in terms of the root mean square error, concordance of correlation coefficient or mean absolute error. To the best of our knowledge, the deep investigation of prediction performance of instantaneous affective state using commonly-referred audiovisual features is absent in the literatures. Thus, there remains a scope of developing new algorithm such that the frame-by-frame emotional state can be predicted in real-time by the use of sufficiently low number of effective features chosen by a suitable selection process.

C. Specific Contributions

In this paper, we focus on instantaneous prediction of continuous-level emotional states using a suitable selection of audiovisual features. In particular, our specific contributions are as follows:

- The frame-by-frame instantaneous prediction of emotional states using a small set of audio-visual features. The extraction process of features involves low-level of computational complexity such that the method can be applied in real-time application.
- The comparison of the performance of different features for prediction of two emotional dimensions, namely, arousal and valence. We show that improvement of prediction result is obtained when only relevant features are taken care of and redundant features are eliminated.

This paper is organized as follows. Features extracted from audio and visual signals are presented in Section II. Sections III and IV detail the feature selection and regression processes, respectively. The experimental setup and results obtained are given in Section V. Finally, concluding remarks are provided in Section VI.

II. Feature Extraction

In order to extract relevant information from audio and video signals for emotion prediction, we perform different processing steps on these signals. In particular, we extract the mel frequency cepstral coefficient (MFCC) from audio signals and local binary pattern (LBP) texture descriptor from

978-1-5090-4090-2/16/$31.00 ©2016 IEEE

ISBN 978-1-5090-4089-6

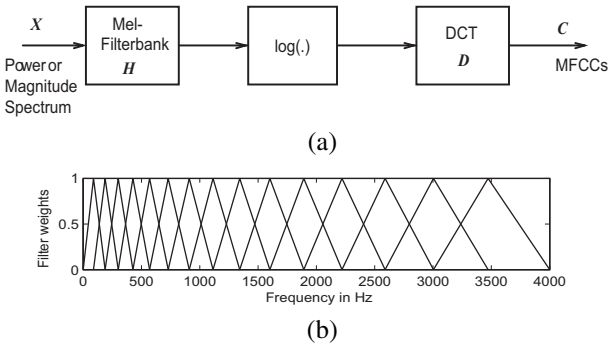

Fig. 2. The extraction process of feature MFCC from speech signals. (a) A simplified block diagram of the process. (b) The bank of 15 mel-scale filters with sampling frequency 8 kHz.

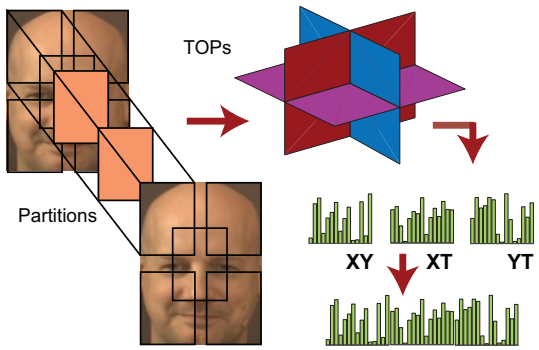

Fig. 3. The process of extraction of LBP-TOP features [19] for prediction of continuous-level emotional states from facial video data.

visual signals as features of emotion. The processes of feature extraction from two-types of signals and their normalization are presented in separate sections.

A. Audio Features

Relative importance of the MFCC in speech emotion recognition is higher than speech energy, pitch, and speech duration [17]. The MFCCs are obtained by discrete cosine transform(DCT) of log power spectrum of short-time speech signal on mel-scale frequency. A simplified block diagram of standard MFCC-based feature extraction process and the mel-scale filter bank is given in Fig. 2. Let $\boldsymbol{X}$ be the power spectrum obtained from windowed speech frame using the discrete Fourier transform. If $\boldsymbol{H}$ is the matrix of mel-scale filter bank and $\boldsymbol{D}$ is the matrix of DCT, then the MFCCs can be estimated as [18]

$$\boldsymbol{C} = \boldsymbol{D} \log (\boldsymbol{H} \boldsymbol{X}) \tag{1}$$

B. Visual Features

The dynamic features represented in terms of the LBPs obtained from TOPs of facial video data are chosen for prediction of continuous-level emotional state motivated by the fact that such features are shown to perform well for recognizing micro-level expressions in the case of spontaneous scenario [20]. Let $I(x, y; t)$ denote the spatial intensity at position (x, y) and time t in a video data of size (X, Y, T). The descriptor LBP-TOP is obtained by concatenating LBP on three planes: XY, XT, and YT, and considering only the co-occurrence statistics in these three directions. In estimating the LBPs in a plane, each pixel is represented by a binary value calculated by thresholding the graylevel of the pixel from that of the neighboring pixels. The value of uniform LBP code in terms of the signs of joint differences is defined as [19]

$$LBP = \sum_{q=0}^{3P+1} u(g_q - g_c) 2^q \tag{2}$$

where g_c and g_q $(q \in 0, 1, 2, \cdots, 3P + 1)$ are the graylevel intensities of the center and neighboring pixels of current, previous, posterior frames with P number of circularly symmetric neighboring points in a frame, respectively, and the function

$$u(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases} \tag{3}$$

Let a set of dynamic textures in terms of uniform LBP of size $X_d \times Y_d \times T_d$ $(x_c \in \{1, 2, \cdots, X_d\}, y_c \in \{1, 2, \cdots, Y_d\}, t_c \in \{1, 2, \cdots, T_d\})$ be estimated by considering only the center part of the neighborhood from a segment of video clip. The histogram of the dynamic feature is estimated as

$$H_{i\ell} = \sum_{x,y,t} I\{\Phi_\ell^c(x, y, t) = i\} \quad i = 1, 2, \cdots, n_\ell \quad \ell = 1, 2, 3 \tag{4}$$

where $\Phi_\ell^c(x, y, t)$ represents the uniform LBP code of central pixel (x_c, y_c, t_c), n_ℓ is the number of different labels produced by the LBP operator in the ℓth plane ($\ell = 1$: XY, 2 : XT, 3 : YT) and

$$I\{A\} = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false} \end{cases} \tag{5}$$

The labels from the XY-plane contain information about the appearance, whereas that in the XT- and YT-planes represent the co-occurrence statistics of motion in horizontal and vertical directions. The three histograms are concatenated to build a global description that represents dynamic feature with the spatial and temporal characteristics of the data, which is often referred to as the LBP-TOP. Fig. 3 shows a schematic diagram of the computation process of the LBP-TOP of a facial video data using five overlapping blocks. It is noted that the LBP-TOPs of five blocks are concatenated to obtain the dynamic feature sets of emotional states.

C. Feature Normalization

In the proposed method, we adopt a simple but effective normalization process of audio and visual features. The MFCCs are obtained from DCT coefficients, hence, the audio features are real-valued data. On the other hand, the video features are non-negative integers, since the LBP-TOPs are obtained in terms of histograms of texture patterns. In order to combine the two different-types of features, first the audio and video feature sets are scaled to the range of $(-1, 1)$ and $(0, 1)$, respectively,

and then concatenated. We refer to the feature set $\mathcal{F}$ of length L with normalized elements as $\{f_i\}$ $(i \in 1, 2, \cdots, L)$.

III. Feature Selection

In order to select the relevant features and to eliminate the redundant features, the minimum redundancy maximum relevance (mRMR) feature selection technique is used [21]. The feature selection process also reduces the length of feature vectors and thus lowers the effective computational complexity. To calculate mRMR ranking of a dynamic feature $f_i(t)$ $(i \in 1, 2, \cdots, L)$, we employ the difference between the maximum relevance and minimum redundancy criteria expressed in terms of their dynamic random variables $F_i(t)$ $(i \in 1, 2, \cdots, L)$ as [21]

$$\Psi_{R\mathcal{F}}(t) = \max_{F_i \in \mathcal{F}} \left[\frac{1}{|\mathcal{F}|} \sum_{F_i \in \mathcal{F}} \mathcal{M}(R(t), F_i(t)) - \frac{1}{|\mathcal{F}|^2} \sum_{F_i, F_j \in \mathcal{F}} \mathcal{M}(F_i(t), F_j(t)) \right] \quad (6)$$

where $R(t)$ is the dynamic random variable of the ground truth of instantaneous emotional rating $r(t)$, $\mathcal{M}(\cdot)$ is the mutual information of two random variables F_i and F_j given by

$$\mathcal{M}(F_i, F_j) = \int \int p(F_i, F_j) \log \frac{p(F_i, F_j)}{p(F_i)p(F_j)} \mathrm{d}F_i \mathrm{d}F_j \quad (7)$$

where $p(F_i)$, $p(F_j)$ and $p(F_i, F_j)$ are the probability density functions. Finally, the feature vector $\mathbf{F_s}$ consisting only the features $\{f_{si}\}$ $(i \in 1, 2, \cdots, L_s)$ with high values of mRMR ranking are constructed to predict the emotional states.

IV. Prediction of Emotional Rating

In the proposed method, a regression technique is required to map the features to continuous-level emotional dimension. We employ the support vector regression (SVR) technique to predict the emotional states from the proposed audiovisual features by acknowledging that it is a well-established statistical learning theory applied successfully in many prediction tasks in computer vision. The kernel SVR implicitly maps the dynamic features into a higher dimensional feature space to find a linear hyperplane, wherein the emotional state can be predicted with a predefined soft error margin. Given a training set of known emotional rating $\mathbf{\Theta}(t) \in \{\mathbf{F_s}(t), r(t)\}$, where $\mathbf{F_s}(t) \in \mathbb{R}^{L_s}$ and $-1 \leq r(t) \leq 1$, the emotional state is predicted using the test feature $\hat{\mathbf{F}}_{\boldsymbol{s}}(t)$ as a regression function given by

$$\hat{r}(t) = \sum_{i=1}^{L_s} \beta_i \Phi \left(f_{si}(t), \hat{f}_{si}(t) \right) + b \quad (8)$$

where β_i are the Lagrange multipliers of a dual optimization problem, $\Phi(\cdot)$ is a kernel function, f_{si} are the support vectors, and b is the weight of bias. In order to map the audiovisual features into the higher dimensional feature space for prediction, the most frequently used kernel functions such as the linear, polynomial, and radial basis function (RBF) can be used. With a view to select the parameters of the SVR, a grid-search on the hyper-parameters is used by adopting a cross-validation scheme. The parameter settings that produce the best cross-validation accuracy are used for predicting the emotional state from the proposed dynamic features under test.

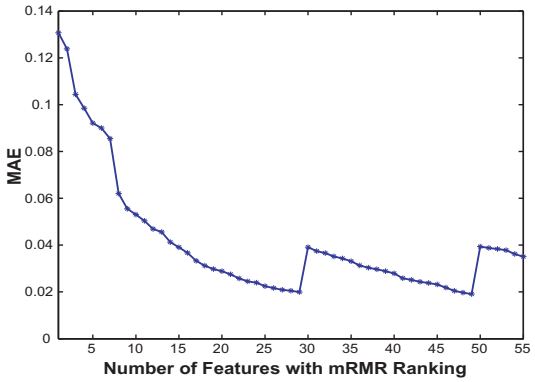

Fig. 4. Prediction performance of dimension valence in terms of MAE with increasing length of proposed dynamic features ranked according mRMR.

TABLE I
Comparison of Prediction Performance of Emotional Dimensions in terms of MAE Using Different Features

Features	Valence	Arousal
Audio (MFCC)	0.1202	0.1614
Visual (LBP-TOP)	0.1049	0.1163
Audio (MFCC with δ and $\delta\delta$ + Zero-Crossing Rate + Log Frame Energy + Centroid with δ) [22]	0.1390	0.1624
Visual (Gabor Energy) [11]	0.1185	0.1586
Proposed Audiovisual (MFCC with δ and $\delta\delta$ + LBP-TOP)	**0.0891**	**0.1114**

V. Experiments

Experimentations are carried out to evaluate the performance of the prediction method using the proposed audiovisual features. The experiments are presented in separate subsections by detailing the nature of the database, the setup that was employed, and the results that were obtained.

A. Database

In the experiments, we consider spontaneous and naturalistic interactions between pairs of humans provided in the SEMAINE (Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression) database [23]. The database was developed by capturing both the audio and video signals of the interactions with a target to analyze the issues of sensitive artificial listener (SAL) agents. The experiments consider the solid-SAL scenario, wherein nonverbal engagements such as the back-channelling, and eye contact between the participants and operators are recorded. The database contains 24 such recording sessions comprising a total of 190 multimedia clips. The audio signal of the databse is recorded at 48 kHz with 24 bits per sample. The grayscale video signal is recorded at 49.979 frames per second and 8 bits per pixel. Continuous ratings of the dimensions valence and arousal in

the range $(-1, 1)$ are provided for each of the frames of the clips in the database.

B. Setup

According to the frame-rate of video signal and sampling frequency of the audio signal of the database, single frame corresponds to the 20ms short-time of audio signal. Thus, in order to calculate the MFCC, the audio signal is segmented into non-overlapping short-time signals with a duration of 20ms. Mel-scale filter bank consisting of 15 uniformly-spaced triangular filters and a bandwidth of 4kHz is applied to the power spectrum of the short-time signal. We have considered all the DCT coefficients from the order 2 to 13 for the calculation of MFCC. The audio features are extracted from the actual MFCCs as well as from their first and second derivatives, also known as δ and $\delta\delta$, respectively. Thus, the initial length of features contributed from each of the short-time frame of audio signals becomes 36.

To extract the visual features, the face region is cropped using the well-known Viola-Jones face detection algorithm [24]. The features in terms of LBP-TOP of the cropped face region are estimated using 5 numbers of neighboring frames by considering that emotional ratings remain nearly unchanged for such a group of frames. The number of neighboring points in a frame is set to 8 to estimate the LBPs. In such a case, the initial number of visual features represented by histogram of all possible uniform patterns becomes 58. Due to normalization and fusion of audio and visual features, the total number of dynamic features for a set of non-overlapping 5-frames prior to the selection process becomes $L = 94$.

In order to select the effective features for prediction of emotional rating the mRMR ranking of each of the features are obtained from a training set of data. Approximately 25% frames of video clips are considered for choosing the training data and the rest of the frames as the test data for predicting the emotional dimensions. In order to train the features with the emotional rating, we have discretized the range of rating in 10-levels uniformly. The sets of 50-neighboring frames that have nearly constant rating (variation is less that 20%) for each of the levels are chosen for the purpose of training. Five-fold cross-validation technique is applied on the training data set to optimize the model of SVR. The parameters such as the length of RBF kernel, weights, and bias of the SVR are optimized in terms of the mean absolute error (MAE) of prediction. The optimized SVR is employed first to select effective length of features those are ranked by the criterion mRMR, and finally for prediction of emotional dimension on the testing set of data.

C. Results

Fig. 4 shows the MAE values obtained for the trained data to predict the dimension valence using the increasing length of features ranked according to mRMR. It is seen from this figure that the MAE of prediction decreases with the inclusion of first few features that have high values of mRMR ranking. However, if the feature length exceeds $L_s = 30$, the MAE values are not ensured to be minimum. The prediction performance of the dimension arousal also shows similar performance characteristics, and hence, we have selected first 30 mRMR ranked audio-visual features to predict emotional dimension of the testing set in the experiments. According to the proposed approach of selection approximately 20% features are counted from the audio signals and rest from the video frames.

The overall performance of prediction is compared with individual features that we have used, i.e., MFCC and LBP-TOP, as well as bundles of audio features including the MFCC with derivatives, zero-crossing rates, log-energy, centroid with derivatives [22] and visual features including the Gabor energy [11]. Table I shows the overall prediction performance of the testing clips in terms of MAE for the comparing methods. It is seen from this table that in general the visual features in terms of texture patterns of facial images performs better than the audio features that represent frequency characteristics of speech. It can be also observed that the prediction of emotional dimension arousal is more challenging than that of dimension valence. Due to the proposed normalization, fusion, and selection processes of the audiovisual features, the MAE of prediction performance for both valence and arousal dimensions are found to the lowest among the comparing methods.

Fig. 5 shows instantaneous prediction of the emotional dimensions valence and arousal for sampled frames of video clips having recording and session IDs (1, 2) and (10, 41), respectively, wherein the sampling period is 5. The training data of 250 frames in these clips or equivalently 50 sampled frames are encircled. It is seen from Fig. 5 that the proposed method can predict the ratings of both dimensions closely. The closeness of prediction is significant, when the changes are small in the testing frames as revealed in the facial appearances of the participants. If there is a sudden change of a dimension, then the proposed prediction can follow the trend within few seconds as per the frame-rate. Thus, the proposed instantaneous emotion prediction technique can be effective in developing real-time artificial listener agents.

VI. Conclusion

Automatic prediction of emotional states is crucial for developing SAL that has many potential applications requiring interaction between machines and humans. The generalized approach of prediction of emotional state follows the steps of extraction of affective features, selection of features, and mapping of selected features using a regressor. This paper has investigated the performance of instantaneous prediction of two commonly-referred emotional dimensions, namely, the valence and arousal, using the low-level audiovisual features including the MFCC and LBP-TOP under certain settings. Extracted features are combined with a feature-level fusion process and then selected by using the mutual information-based mRMR ranking. These low-level features are mapped on the emotional dimensions using the SVR technique. Performance of the proposed audiovisual features is compared

978-1-5090-4090-2/16/$31.00 ©2016 IEEE

ISBN 978-1-5090-4089-6

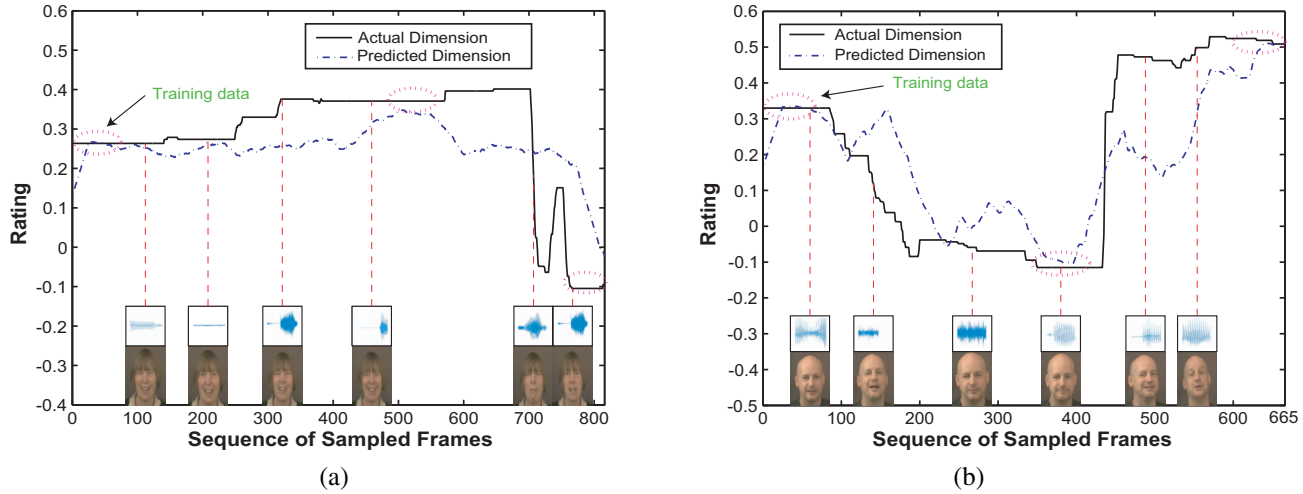

Fig. 5. Instantaneous prediction of emotional dimensions using the proposed audiovisual features. The dimensions are (a) valence and (b) arousal.

with existing audio and visual features for prediction of instantaneous rating of valence and arousal dimensions. The MAEs calculated using different length of feature vector show that the prediction performance improves significantly, when approximately 30% of top ranked features are considered for the regression. Experiments on instantaneous prediction reveal that a moderate length audiovisual features as proposed in this paper can provide a few seconds of settling time even when an emotional dimension changes sharply.

REFERENCES

[1] 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.

[2] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410–430, 2015.

[3] P. Ekman, *Handbook of Cognition and Emotion*. John Wiley & Sons, Ltd, 2005, ch. 3, pp. 45–60.

[4] R. B. Dietz and A. Lang, "Affective agents: Effects of agent affect on arousal, attention, liking and learning," in *Proc. Int. Conf. Cognitive Technology*, San Francisco, CA, 1999.

[5] M. K. Greenwald, E. W. Cook, and P. J. Lang, "Affective judgement and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *Journal of Psychophysiology*, vol. 3, pp. 51–64, 1989.

[6] L. S. Chen and T. S. Huang, "Emotional expressions in audiovisual human computer interaction," in *Proc. IEEE Int. Conf. Multimedia and Expo*, New York, 2000, pp. 423–426.

[7] L. C. D. Silva and P. C. Ng, "Bimodal emotion recognition," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, Grenoble, 2000, pp. 332–335.

[8] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting Psychologists Press, 1978.

[9] S. M. Imran, S. M. M. Rahman, and D. Hatzinakos, "Differential components of discriminative 2D Gaussian-Hermite moments for recognition of facial expressions," *Pattern Recognition*, vol. 56, pp. 100–115, 2016.

[10] M. Wollmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, pp. 153–163, 2013.

[11] M. Dahmane and J. Meunier, "Continuous emotion recognition using Gabor energy filters," in *Lecture Notes in Computer Science: Affective Computing and Intelligent Interaction*, 2011, vol. 6975, pp. 351–358.

[12] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recognition and Workshops*, Santa Barbara, CA, 2011, pp. 314–321.

[13] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proc. ACM Int. Conf. Multimodal Interaction*, Santa Monica, CA, 2012, pp. 501–508.

[14] G. A. Ramirez, T. Baltrušaitis, and L.-P. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *Proc. Int. Conf. Affective Computing and Intelligent Interaction*, Memphis, TN, 2011, pp. 396–406.

[15] Y. Cui, S. Luo, Q. Tian, S. Zhang, Y. Peng, L. Jiang, and J. S. Jin, *Mutual Information-Based Emotion Recognition*. Springer, New York, 2013, pp. 471–479.

[16] H. Meng and N. Bianchi-Berthouze, "Naturalistic affective expression classification by a multi-stage approach based on hidden Markov models," in *Proc. Int. Conf. Affective Computing and Intelligent Interaction*, Memphis, TN, 2011, pp. 378–387.

[17] D. Wu, T. D. Parsons, and S. Narayanan, "Acoustic feature analysis in speech emotion primitives estimation," in *Proc. Int. Conf. InterSpeech*, Makuhari, Japan, 2010, pp. 785–788.

[18] S. Panchapagesan, "Frequency warping by linear transformation of standard MFCC," in *Proc. Int. Conf. InterSpeech*, Pittsburgh, PA, 2006, pp. 397–400.

[19] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, pp. 951–928, 2007.

[20] B. M. S. B. Talukder, B. Chowdhury, T. Howlader, and S. M. M. Rahman, "Intelligent recognition of spontaneous expression using motion magnification of spatio-temporal data," in *Lecture Notes in Computer Science: Intelligence and Security Informatics*, 2016, vol. 9650, pp. 114–128.

[21] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[22] M. Wollmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-cowie, and R. Cowie, "Abandoning emotion classes - Towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. Int. Conf. InterSpeech*, Brisbane, Australia, 2008, pp. 597–600.

[23] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[24] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Int. J. Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.