Continuous Multi-modal Emotion Prediction in Video based on Recurrent Neural Network Variants with Attention

Joyal Raju[1], Yona Falinie A. Gaus[1], Toby P. Breckon[1,2]
Department of {Computer Science[1] | Engineering[2]}, Durham University, UK

***Abstract*—Automatic perception and understanding of human emotion is becoming an increasingly attractive research field in artificial intelligence and human-computer interaction. Emotion portrayal within conversation plays a significant role in the semantics of a sentence. However, emotion is not only biologically determined but is also influenced by the environment. Therefore, cultural differences exist in some aspects of emotions, and it is important for the next generation of computer systems to adapt the cross-cultural difference in order to enable more naturalistic interactions between humans and machines. In this paper, we investigate the suitability of state-of-the-art deep learning architectures based on recurrent neural network (RNN) variants with explicit attention modelling to bridge the gap across different cultures (German and Hungarian) for emotion prediction in video. Three different attention based network architectures are proposed in this work:- early attention fusion, extended multi-attention fusion and attention-based encoder-decoder. Our RNN variants with explicit attention modelling approach achieves very promising Concordance Correlation Coefficient results, which outperform the baseline on Arousal of 0.637 *vs.* 0.614 (baseline), for Valence of 0.689 *vs.* 0.615 and for Liking of 0.625 *vs.* 0.222.**

***Index Terms*—affective computing, emotion, multi-modal, cross-cultural, attention network, emotion recognition, motion detection**

I. INTRODUCTION

Emotions are a complex state of feeling that results in physical and psychological changes, variously associated with thoughts, feelings, behaviour, and a degree of pleasure or displeasure [1]. They form a very important semantic component in human conversations - without the context of the speakers emotion, the true intention of the utterance may be ambiguous. Therefore, the affective computing field aims to recognise human emotion to enable more naturalistic human-computer interaction [2]. There are two type of emotion representation models in affective computing:- categorical and dimensional [3]. As for the categorical model, a persons emotional state is described via a discrete set of affective attributes [4], such as *happiness*, *sadness*, *fear*, *disgust*, *anger*, *surprise*. In contrast, within the dimensional model the emotional state is mapped to a continuous coordinate points in a 3D or 2D Cartesian space, such as valence-arousal-dominance (VAD) [5]. VAD is the emotion state representation in continuous space where valence represents the pleasantness ranging from positive to negative, arousal represents the intensity of emotion ranging from excited to calm, and dominance represents the degree of control ranging from uncontrolled to in control [5].

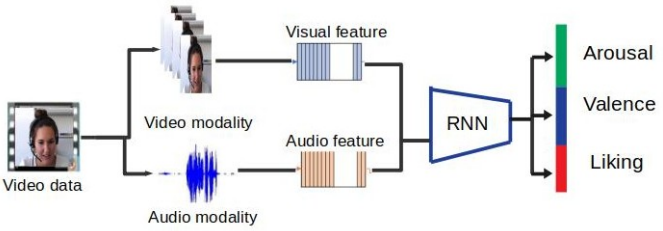

Fig. 1. Overview of the proposed multi-modal emotion prediction approaches.

The relationship between categorical and dimensional models are studied extensively by Sun *et al* [6]. For example, positive valence relates to a happy state, while negative valence relates to a sad or angry state. Whilst categorical models are easier to understand, the limited discrete set of categories may not suitably reflect the subtlety of emotions. In contrast, dimensional models can express subtler and more complicated emotional states when compared to their categorical counterparts.

Constructing a good dimensional emotion prediction approach relies upon three aspects: various multi-modal feature extraction, an effective feature fusion strategy, and a versatile numerical regression model [7] [8] [9]. Traditionally, researchers rely on handcrafted audio-visual features. For video, appearance features such as spatial-temporal facial textures extracted via the Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [10] and geometric features such as facial landmarks [11] [12] have been often used. As for audio, low level descriptor features provided in the Geneva Minimalistic Acoustic Parameter Set (eGEMAPS) have also been used. Recently, the rise of deep learning based methods, particularly in convolutional neural networks (CNN) and recurrent neural networks (RNN), motivates affective computing research to use CNN to extract deep audio-visual features as viable deep representations for effective emotion prediction [7] [13] [9].

The performance of emotion prediction approaches are often influenced by cultural diversity [14]. In previous literature [15] [16] [17], some aspects of emotion have been shown to be culturally different. For example, traditionally-defined Western culture perceives high arousal states as happiness, meanwhile traditionally-defined Eastern culture perceives happiness from experiencing low arousal states [17]. As such, in this work our goal is to evaluate the effectiveness of CNN architectures in order to improve the generalization

of emotional prediction in such cross-cultural scenarios. Our main aim is to predict emotion automatically in the *valence-arousal* emotional dimension, and a third dimension describing *liking* (or sentiment) continuously over time. In summary, our contributions are:

- a broad analysis spanning both hand-crafted features and deep learning features within the context of continuous emotion prediction from video. For the hand-crafted audio features, we use lower-level descriptor features such as eGeMAPS and Mel Frequency Cepstral Coefficient (MFCC). For video, we utilize Facial Action Units (FAU) to extract appearance and geometric information from different facial attributes. In terms of deep learning features, we use Deep Spectrum [18] features from pre-trained CNN [19] [20] [21] on audio. As for video modality, we employ a VGG-16 [19] as well as a ResNet-50 [22] network that are pre-trained with the Affwild [23] dataset as our deep features representation.
- an evaluation of RNN variants with explicit attention modelling across two different cultures, German and Hungarian in the context of continuous emotion prediction from video. We further exploit three different RNN architectures: early attention fusion, extended multi-attention fusion and attention-based encoder-decoder. Our proposed RNN variants demonstrates promising results, outperforming the prior work of [24].

II. RELATED WORK

Applications of emotion prediction face many challenges, mostly due to variations across individuals such as culture, gender, demographic and so on. Notable work such as Chiou *et al.* [25] attempt to solve such variation issues by combining emotional datasets from Mandarin and German to improve the cross-cultural performance. Neumann *et al.* [26] address the inconsistencies between two cultures (French and English) by utilizing a small amount of data from the target culture in order to finetune an emotion model originally trained on a larger dataset from the source culture. Gideon *et al.* [27] uses a domain adaptation approach on cross-culture speech emotion prediction. In this section, we present a focused summary of the current state-of-the-art with respect to topics related to the methodology of automatic emotion prediction proposed in this work.

A. Multi-modality Features in Emotion Prediction

In continuous emotion prediction, various multi-modal features have been utilized, ranging from hand-crafted features to deep learning features. Notable works such as Sánchez-Lozano *et al.* [28] make use of Local Binary Patterns (LBP) and Gabor features from the visual modality and low-level descriptors such as MFCC from audio. On the other hand, Wollmer *et al.* [29] and Brady *et al.* [7] focus on low-level descriptor features as an input towards Long Short Term Memory (LSTM) networks and Support Vector Regression (SVR), respectively. As deep learning over hand-crafted features, emotion prediction research has similarly tended to prefer deep features representations from varying deep neural network architectures. Notable works such as Chen *et al.* [13], Huang *et al.* [30] and Zhao *et al.* [9] demonstrate that deep learning features achieve comparable or even better performance than handcrafted features, across video and audio modalities.

Whilst multi-modal features are widely regarded as the most encompassing representation of emotion, multi-modal fusion is also essential to achieve better performance for emotion prediction approaches. There are mainly three strategies to achieve multi-modal fusion, namely early-fusion, late fusion and model-level fusion. Previous work by Huang *et al.* [30] adopts late fusion to combine the predictions of different features. Meanwhile, Chen *et al.* [13] explore early fusion from all available modalities. Subsequent work by Huang *et al.* [31] compare the performance between these two methods. This study [31] shows that that late fusion is good at predicting arousal and valence, while early fusion is more suitable for liking prediction. The latest work by Huang *et al.* [32] employed model-level fusion by fusing audio-visual modalities via a multi-headed attention module and hence achieving superior performance than early fusion and late fusion alone. Therefore in this work we follow Huang *et al.* [32] approach by adopting a model-level fusion strategy in RNN architecture in order to accommodate representative features both the video and audio modalities.

B. Model Architecture in Emotion Prediction

With respect to the continuous emotion prediction, various regression approaches have been employed. Most commonly two regression models are frequently utilised: Kernel based Support Vector Regression and Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) [33] [13] [34] [24]. Previous work by Wollmer *et al.* [33] and Chen *et al.* [13] utilize both LSTM-RNN and SVR to perform regression analysis on the arousal and valence dimension. These studies reveal that LSTM manage to capture the temporal information of the emotional dimension significantly, outperforming SVR. Later work by Huang *et al.* [34] utilize three kinds of temporal architectures, including LSTM, Time-Delay Neural Networks (TDNN) and a multi-headed attention network, to learn different temporal modeling in the sequence hence showing that a combination of these approaches obtains the best result. Meanwhile, Chen *et al.* [13] fully utilize Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks (DBLSTM) for each unimodal feature representation. Similarly in this work we adopt LSTM-RNN architecture that incorporate an attention network in continuous emotion prediction.

C. Attention in Emotion Prediction

These aforementioned studies [13] [34] [31] make the assumption that emotion can be reliably predicted at every single frame, which maybe an unreasonable assumption. As for example, in the audio modality, the characteristics of each uttered word may not offer an equitable insight into the sentiment of the entire sentence. Similarly, any single still image frame from a video sequence may not accurately

convey the emotional state of the human subject. Mirsamadi *et al.* [35] tackle this situation by assigning a weight to each frame depending on how emotionally salient the features are, by using an attention mechanism. This allows the network to focus on the video frames in the sequence that containing strong emotional characteristics. Lee *et al.* [36] extend attention weights towards both textural (visual) features and audio features such that the attention weights indicate the modality whose features are most useful for every frame. Wang *et al.* [37] expand the concept of a multi-modal network architecture for this task by utilising multiple attention layers. Each input for the audio/video modality is weighted using an attention mechanism similar to the method used by Mirsamadi *et al.* [35] to emphasize emotionally meaningful features from each modality. These approaches [35] [36] [37] readily demonstrate the added value of an attention mechanism as they are shown to outperform the same architectural approach with the attention component removed. Inspired by the effectiveness of dynamically ignoring unreliable modalities, we explore the use of attention mechanisms allowing our model to learn to dynamically assign larger weights to more useful features.

III. METHODOLOGY

In this work, we investigate three different RNN architecture variants, that each incorporate an attention network subcomponent in a variety of architecture. All of them take both audio and visual features as input (see Section IV) and output a continuous time series of three values, representing a subject emotional dimension for the *arousal*, *valence* and *liking* dimensions, as shown in Figure 1.

A. Early Attention Fusion Model (EAF)

The first variant RNN architecture is illustrated in Figure 2. In this approach, we fully utilize CNN Deep Spectrum and facial features for video whilst using MFCC features for audio. Our approach first applies a fully connected layer and a softmax function respectively. Subsequently, we calculate the attention weight obtain from a softmax function for each modality. Subsequently, these features will become an input to two layer bi-directional LSTM. In this stage, the LSTM will produce one fixed-length vector from final hidden state which encompasses all the necessary information of the of the three-dimensional continuous emotion representation in use.

B. Attention based Encoder and Decoder Model (AED)

A key limitation of RNN techniques is that they are limited in their ability to track long-term dependencies on the emotion. Since emotions are dynamic [34], the ability of the architecture to capture emotional long-term dynamic should be well modeled. In order to address this issue, we implement the architecture from Bahdanau *et al.* [38], where an encoder-decoder model can learns to align and translate jointly, as shown in Figure 3. It consists of a bidirectional LSTM as an encoder, and attention weights connecting each input location to each output location. This architecture allows the LSTM decoder to focus on all hidden states instead of just the final hidden state. The additional use of attention within this architecture enables the decoder a flexibility to identify the parts of the features that may relevant for three-dimensional continuous emotion representation in use.

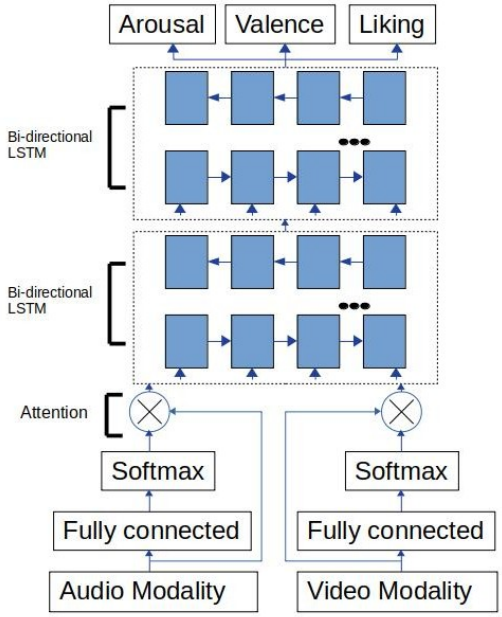

Fig. 2. Early Attention Fusion architecture (EAF).

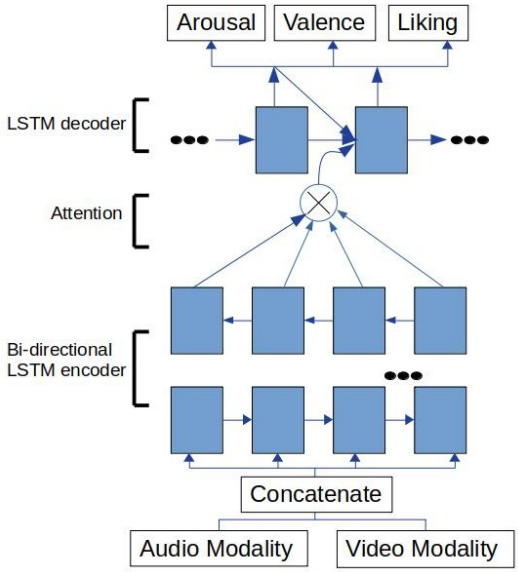

Fig. 3. Attention based Encoder and Decoder architecture (AED).

C. Extended Multi-Attention Fusion Network (EMAFN)

As discussed in Section III-A and III-B, both of the aforementioned architectures are good candidates for continuous emotion prediction, since they manage to capture dynamic relationship existing between consecutive features taken from both audio and video. Therefore, we further utilise these two temporal architecture by replacing bi-directional LSTM (Figure 2) with bi-directional LSTM with attention mechanisms (Figure 3) as shown in Figure 4. In this approach, we extend

conventional attention by allowing multiple attention in both feature and model level, respectively. These multiple attention enables the features to learn emotional dynamic and at the same time concentrates only the parts of the features which relevant for continuous emotion prediction.

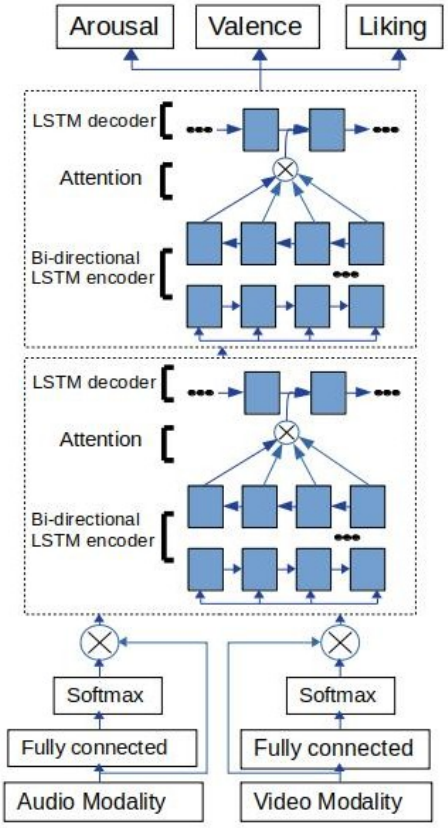

Fig. 4. Extended Multi-Attention Fusion Network architecture (EMAFN).

IV. EXPERIMENTAL SETUP

In this section, we discuss two experimental aspects of our RNN variant architectures with attention, as explained in Section III. Namely, our choice of datasets and our choice of underlying feature extraction approach.

A. Datasets

Our proposed architectures are trained on features extracted from the Audio/Visual Emotion Challenge (AVEC) 2019 dataset [24]. This dataset consists of audio and video recordings of dyadic interactions between friends of both German and Hungarian human subjects, under uncontrolled settings using webcams and microphones. The videos have been annotated at the individual frame level across the emotional dimensions of *arousal*, *valence* and *liking* by human annotators that were native speakers of the corresponding language that a given recording was in. The dataset is provided as part of the AVEC 2019 challenge [24] and is partitioned into training and development partitions. The training partition contains thirty-four videos each for participants of German and Hungarian descent (68 videos in total comprising of approximately 123k annotated frames). The development partition contains fourteen videos each for participants of German and Hungarian descent (28 videos in total comprising of approximately 38k annotated frames). As we do not have access to the original testing partition (set) from the AVEC 2019 challenge [24], we instead solely use the training partition during training of the models, and the development partition is used as the testing partition.

B. Feature Extraction

From the audiovisual recordings, we fully utilises two types of features: low level descriptor features and deep learning based features, which are explained in detail in the remainder of this section.

Low-Level Descriptor Features: For audio we use the features defined in the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) and Mel Frequency Cepstral Coefficient (MFCC) using the openSMILE toolkit [39] that describe spectral, cepstral, prosodic and voice quality information [24] to generate an audio descriptor. Similarly, we utilize appearance, geometric information from the video dataset in the form of Facial Action Units (FAU). FAU are descriptors of the positions of different facial features and the intensity of the expression it contributes to. These FAU together with gaze orientation for the participant within each video frame are extracted using the openFACE toolkit [40] to generate a visual facial descriptor. These low-level features are summarised by computing their mean and standard deviation using a sliding window of 4s in length and a hop size 100ms, from which a bag-of-words feature representation is used to capture the distribution of these features. This is achieved with the use of the openXBOW toolkit [41] using codewords to form a bag of words dictionary representation of size 100. These extracted feature representations are thus concatenated to form a joint 386 dimension audio and 153 dimension video feature descriptor which forms the dense tensor input to each of our proposal RNN variant architectures outlined in Section III.

Deep Features: Mel-Spectrogram images of the audio are produced using Deep Spectrum [42] for a window size of 4s in length and a hop size 100ms. These images are then passed through a set of pre-trained deep convolutional neural networks, from which we extract activations from the late-stage layers within the architecture to use. In this manner we obtain a 4096 dimension feature vector from the activations each of VGG-16 [19] and AlexNet [20] in each of the locations of second fully connected layer. Similarly we obtain 1024 and 1920 dimensional feature vectors from the activations each of DenseNet-121 and DenseNet-201 [21] in each of the locations of last average pooling layer. For visual features we extract features from a VGG-16 in each of the locations of first fully-connected layer and a ResNet-50 n each of the locations of global average pooling layer, both of which have been pre-trained on the Aff wild dataset [24]. As a result, we obtain 4096 dimensional deep feature vector from VGG-16 and a 2048 dimensional deep feature vector from ResNet-50 for each frame. These extracted feature representations are thus concatenated to form a joint 11136 dimension audio and 6144 dimension video feature descriptor which forms the dense tensor input to each of our proposal RNN variant architectures outlined in Section III.

V. EVALUATION

In this section, we evaluate the performance of the RNN variant architectures with attention (Section III) under the respective evaluation metrics and review the performance on emotional dimensions of *arousal*, *valence* and *liking* respectively in a cross-cultural context.

A. Performance Evaluation

We treat the emotion dimensional prediction of *arousal*, *valence* and *liking* as a regression task. Therefore, to evaluate the quality of time series prediction which represents each of the aforementioned emotion dimensions, we use the concordance correlation coefficient (CCC), ρ_c, for a time series variable x compared to a ground truth time series variable y is defined as follows:

$$\rho_c = \frac{2\rho_{x,y}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{1}$$

where symbol $\rho_{x,y}$ is the pearson correlation coefficient (PCC) defined as follows

$$\rho_{x,y} = \frac{covariance(x,y)}{\sigma_x, \sigma_y} \tag{2}$$

and σ_x^2 and σ_y^2 are the variance of each time series and μ_x and μ_y are the mean values of each series. During training, we transformed this equation into a loss function so that the weight updates serve to maximise this statistic. A CCC of 1 indicates a perfect correlation between the two time series variables, whilst a CCC of 0 indicates little to no correlation. Our choice of this statistic is based on the fact that it is amplitude (scale) and phase (temporal location) invariant [43] which helps to mitigate for labelling inaccuracies within the ground truth that is attributable to the reaction time of the annotators [44].

B. Analysis

The results are reported in Table I for AVEC 2019 baseline [24] along with our proposed architectures as detailed in Section III. In emotional dimension of *arousal*, the AED architecture improves emotion detection accuracy with an increase performance over the baseline results. The AED architecture achieves CCC value of 0.635, 0.638, and 0.637 (Table I, upper) when tested on German, Hungarian and German plus Hungarian cultures together, respectively. The AED architecture also improves upon the performance in the emotional dimension of *valence*, achieving CCC value of 0.708 and 0.689 (Table I, middle) for both Hungarian and German plus Hungarian cultures together, respectively. At the same time, the AED architecture provides somewhat average performance on German culture, with a CCC value of 0.676, as opposed to 0.684 in the baseline results. We further observe that EMAFN provides great improvement upon the baseline resuts in emotional dimension of *liking*. The EMAFN architecture manages to achieve a CCC value of 0.386, 0.803 and 0.625 (Table I, bottom) when tested on German, Hungarian and German plus Hungarian cultures together, respectively. Overall, this indicates that encoder-decode model with attention weight in AED architecture makes a positive effect on performance. We also have shown that attention can be an effective tool in the detection of emotion when the audiovisual dataset is diverse and includes multiple cultures.

TABLE I
CONCORDONCE CORRELACTION COEFFICIENT RESULTS FOR EMOTIONAL DIMENSIONS OF GERMAN AND HUNGARIAN CULTURES FROM BASELINE APPROACH AND OUR PROPOSED ARCHITECTURE.

Culture	Baseline [24]	[45]	[46]	EAF	EMAFN	AED
			arousal			
German	0.629	0.789	**0.791**	0.486	0.132	0.635
Hungarian	0.583	0.583	0.585	0.222	0.181	**0.638**
German + Hungarian	0.614	0.724	**0.737**	0.365	0.155	0.637
			valence			
German	0.684	**0.794**	0.778	0.510	0.659	0.676
Hungarian	0.508	0.572	0.463	0.152	0.586	**0.708**
German + Hungarian	0.615	**0.708**	0.653	0.345	0.631	0.689
			liking			
German	0.048	0.352	0.441	0.160	**0.386**	-0.129
Hungarian	0.260	0.311	0.208	0.261	**0.803**	0.224
German + Hungarian	0.222	0.320	0.425	0.215	**0.625**	0.062

We further compare our use of RNN architecture variance with attention against the two best entrants of the AVEC 2019 Cross-Cultural Emotion Sub-challenge [45] [46] in Table I. In comparison, our results offer the best performance on emotional dimension of *liking* via EMAFN architecture with CCC value of 0.803 and 0.625 (Table I, bottom) on both Hungarian and German plus Hungarian respectively. Our proposed AED architecture also offers competitive performance, with the best CCC value of 0.638 (Table I, upper) and 0.708 (Table I, middle) on emotional dimension of *arousal* and *valence* for Hungarian culture respectively.

VI. CONCLUSION

In this paper, we explore the combination of efficient deep learning and hand-crafted features across audio and video, and novel variant RNN architectures with attention in the cross-cultural context. We explore three different RNN variants with attention, the early attention fusion model (EAF), the extended multi attention fusion network (EMAFN) and the attention-based encoder-decoder model (AED). Our proposed variant RNN architectures with attention manage to capture the emotional dynamic within continuous emotional dimension, namely *arousal*, *valence* and *liking*. The cross-cultural experimental results demonstrate that our proposed method can improve the emotional prediction performance over baseline results and additionally provide competitive performance against leading contemporary techniques on the AVEC 2019 benchmark dataset. Future work will explore both the use of transformer architectures and the additional use of text as an additional input modality.

REFERENCES

[1] M. Cabanac, "What is emotion?" *Behavioural Processes*, vol. 60, no. 2, pp. 69 – 83, 12 2002.

[2] R. W. Picard, *Affective computing*. MIT press, 2000.

[3] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, "Affective computing for large-scale heterogeneous multimedia data: A survey," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 3s, pp. 1–32, 2019.

[4] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[5] H. Schlosberg, "Three dimensions of emotion." *Psychological review*, vol. 61, no. 2, p. 81, 1954.

[6] K. Sun, J. Yu, Y. Huang, and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *Int. Conf. on Multimedia and Expo*. IEEE, 2009, pp. 566–569.

[7] 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 *Int. Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 97–104.

[8] S. Chen and Q. Jin, "Multi-modal conditional attention fusion for dimensional emotion prediction," in *Int. Conf. on Multimedia*, 2016, pp. 571–575.

[9] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 65–72.

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

[11] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Multi-scale temporal modeling for dimensional emotion recognition in video," in *Int. Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 11–18.

[12] ——, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Int. Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 65–72.

[13] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 19–26.

[14] N. Lim, "Cultural differences in emotion: differences in emotional arousal level between the east and the west," *Integrative medicine research*, vol. 5, no. 2, pp. 105–109, 2016.

[15] J. A. Russell, "Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies." *Psychological bulletin*, vol. 115, no. 1, p. 102, 1994.

[16] H. R. Markus and S. Kitayama, "Culture and the self: Implications for cognition, emotion, and motivation." *Psychological review*, vol. 98, no. 2, p. 224, 1991.

[17] Y. Uchida and S. Kitayama, "Happiness and unhappiness in east and west: themes and variations." *Emotion*, vol. 9, no. 4, p. 441, 2009.

[18] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 3512–3516.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. on Learning Representations*, 2015.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[23] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *Int. Journal of Computer Vision*, vol. 127, no. 6, pp. 907–929, 2019.

[24] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Int. on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

[25] B.-C. Chiou and C.-P. Chen, "Speech emotion recognition with cross-lingual databases," in *Fifteenth Annual Conf. of the Int. Speech Communication Association*, 2014.

[26] M. Neumann *et al.*, "Cross-lingual and multilingual speech emotion recognition on english and french," in *Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 5769–5773.

[27] J. Gideon, M. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)," *IEEE Transactions on Affective Computing*, 2019.

[28] E. Sánchez-Lozano, P. Lopez-Otero, L. Docio-Fernandez, E. Argones-Rúa, and J. L. Alba-Castro, "Audiovisual three-level fusion for continuous estimation of russell's emotion circumplex," in *ACM Int. Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 31–40.

[29] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2362–2365.

[30] J. Huang, Y. Li, J. Tao, Z. Lian, Z. Wen, M. Yang, and J. Yi, "Continuous multimodal emotion prediction based on long short term memory recurrent neural network," in *Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 11–18.

[31] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and M. Yang, "Multimodal continuous emotion recognition with data augmentation using recurrent neural networks," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 57–64.

[32] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 3507–3511.

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

[34] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Efficient modeling of long temporal contexts for continuous emotion recognition," in *Int. Conf. on Affective Computing and Intelligent Interaction*. IEEE, 2019, pp. 185–191.

[35] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 2227–2231.

[36] W. Y. Choi, K. Y. Song, and C. W. Lee, "Convolutional attention networks for multimodal emotion recognition from speech and text data," in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language*, 2018, pp. 28–34.

[37] Y. Wang, J. Wu, and K. Hoashi, "Multi-attention fusion network for video-based emotion recognition," in *Int. Conf. on Multimodal Interaction*, 2019, pp. 595–601.

[38] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Int. Conf. on Learning Representations*, 2015.

[39] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Int. Conf. on Multimedia*, 2013, p. 835–838.

[40] T. Baltrusaitis *et al.*, "Openface 2.0: Facial behavior analysis toolkit," in *Int. Conf. on Automatic Face Gesture Recognition*, 2018, pp. 59–66.

[41] M. Schmitt and B. Schuller, "openxbow – introducing the passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, vol. 18, no. 96, pp. 1–5, 2017.

[42] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. J. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *INTERSPEECH*, 2017.

[43] F. Weninger, F. Ringeval, E. Marchi, and B. W. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio." in *IJCAI*, vol. 2016, 2016, pp. 2196–2202.

[44] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 3–9.

[45] H. Chen, Y. Deng, S. Cheng, Y. Wang, D. Jiang, and H. Sahli, "Efficient spatial temporal convolutional features for audiovisual continuous affect recognition," in *Int. on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 19–26.

[46] J. Zhao, R. Li, J. Liang, S. Chen, and Q. Jin, "Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions," in *Int. on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 37–45.