IMPLICIT FUSION BY JOINT AUDIOVISUAL TRAINING FOR EMOTION RECOGNITION IN MONO MODALITY

Jing Han[1], *Zixing Zhang*[2], *Zhao Ren*[1], *Björn Schuller*[1,2]

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]GLAM – Group on Language, Audio & Music, Imperial College London, UK

jing.han@informatik.uni-augsburg.de

ABSTRACT

Despite significant advances in emotion recognition from one individual modality, previous studies fail to take advantage of other modalities to train models in mono-modal scenarios. In this work, we propose a novel joint training model which implicitly fuses audio and visual information in the training procedure for either speech or facial emotion recognition. Specifically, the model consists of one modality-specific network per individual modality and one shared network to map both audio and visual cues into final predictions. In the training process, we additionally take the loss from one auxiliary modality into account besides the main modality. To evaluate the effectiveness of the implicit fusion model, we conduct extensive experiments for mono-modal emotion classification and regression, and find that the implicit fusion models outperform the standard mono-modal training process.

Index Terms— Joint training, audiovisual learning, emotion regression, emotion classification

1. INTRODUCTION

Automatic emotion recognition plays a vital role in the development of emotional artificial intelligence [1, 2]. Over the past few decades, great advances have been made in a variety of modalities, such as facial expression, hand gesture, head position, body posture, speech, text, and physiological signals [3–8]. Further, increasing research intended to make use of multiple modalities by means of developing multi-modal emotion recognition systems. These multi-modal systems are innovated to use certain fusion technologies to improve the model performance when compared with mono-modal systems. To date, there are a plethora of fusion strategies available. For example, in [9, 10], the features extracted from the audio and video modalities are combined together to train models; in [11, 12], the decisions from audio-based and video-based models are fused for a final prediction.

However, in the evaluation phase, these systems often require the synchronous presence of the modalities that are employed in the training phase. The absence of any involved modality often leads to the corruption or the performance degradation of pre-trained multi-modal models [3]. In contrast, in real-life scenarios, it is a common case that signals from some particular modality are missing. For example, the camera could be not always fixed in front of a user, or not always available under light, which results in invalid or missing visual signals. Similarly, a user could be silent although she/he is emotional. A straightforward way to address this issue is to integrate an additional component, such as voice activity detection and face detection, in the front-end of the multi-modal recognition systems [3]. Once the absence of particular signals is detected, the prediction process could be automatically re-directed to a mono-modal system. The mono-modal system, nevertheless, is trained via mono-modal signals, and normally inferior to the multi-modal system.

In this contribution, we propose a novel fusion approach, namely *implicit fusion*. This approach is particularly innovated to enhance the performance of a mono-modal system, by exploiting the information from other auxiliary modalities in the training phase. That is, we use the data from multiple modalities to jointly train the system, with an assumption that the knowledge from different modalities could be transferred/fused to the system; whereas in the evaluation phase, those auxiliary modalities are not required anymore. Therefore, the proposed approach differs from the training process for conventional mono-modal systems that are trained merely with one modality. It also differs from previous fusion strategies (explicit fusion henceforth) that are specifically designed for multi-modal emotion recognition systems and normally need the same modalities in both training and evaluation phases as aforementioned.

Furthermore, our work is partially inspired by the multi-task learning paradigm, where multiple tasks are jointly trained with a shared network and several task-specific networks. It has been repeatedly demonstrated that such a learning process can lead to a better generalisation of the representation learnt from the shared networks [10, 12, 13]. Similarly, our motivation is that, an auxiliary modality could be beneficial for training a mono-modal framework.

In this paper, we focus on the modalities of audio and video for emotion recognition, because humans mainly rely on facial expressions and vocal intonations when perceiving others' emotional states [14]. The major contributions of this work include: (1) proposing a novel implicit fusion method to explore knowledge from auxiliary modalities; (2) jointly training a model with audio and visual data for mono-modal emotion recognition; and (3) investigating the effectiveness of the model for both the categorical emotion classification and the dimensional emotion regression.

2. RELATED WORK

In the literature, a number of fusion paradigms have been investigated for multi-modal emotion recognition [3, 15–20]. In general, these paradigms can be categorised into three groups, i. e., *feature-level* fusion, *decision-level* fusion, and *model-level* fusion. Feature-level fusion (*aka* early fusion), is implemented by simply concatenating features from multiple modalites into one combined vector as

This work was supported by the TransAtlantic Platform "Digging into Data" collaboration grant (ACLEW: Analysing Child Language Experiences Around The World), with the support of the UK's Economic & Social Research Council through the research Grant No. HJ-253479, and by the European Union's Horizon H2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 766287 (TAPAS).

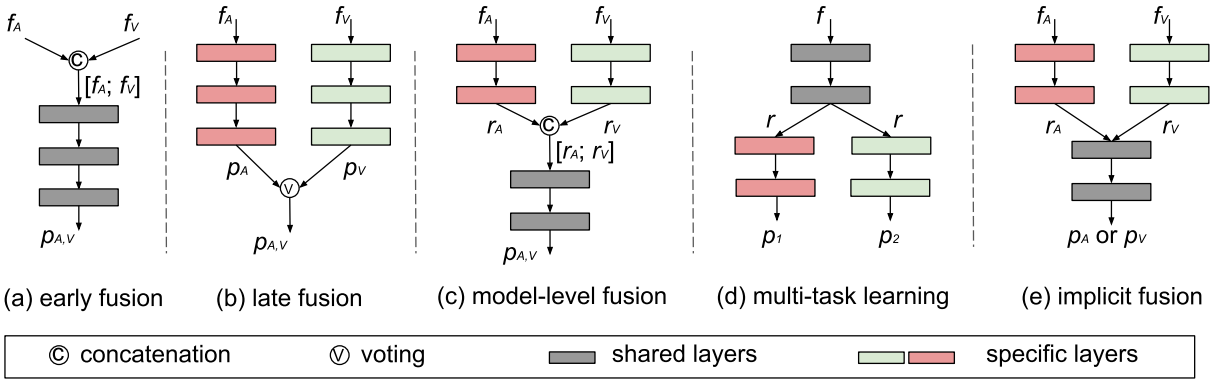

Fig. 1. Overview of the implicit fusion framework (e) and other explicit fusion approaches (a,b,c), as well as multi-task learning (d).

the input of a prediction model. By implementing early fusion, better results have been achieved and reported in [9, 10]. However, it often suffers from the high dimensionality of the feature space and the synchronisation of different modalites [10]. Decision-level fusion (*aka* late fusion), on the contrary, combines predictions reaped from different modalities to come up with a final prediction via a voting strategy. It has been studied and applied successfully in affective computing [11, 12, 21]. In this method, however, the mutual correlation between the modalities is overlooked. As a compromise of early fusion and late fusion, model-level fusion has been proposed which fuses the intermediate representations of different modalities [17–19].

Different from all of the previous fusion strategies, our proposed fusion strategy integrates the information from different modalities in an implicit manner, rather than the explicit concatenating or voting. With this manner, only one modality is demanded during the evaluation phase. This offers the framework a vital important advantage as the single modality setting is often met in practise.

3. JOINT AUDIOVISUAL TRAINING

To leverage the complementary information from audio and video modalities, joint training can be applied in either an *explicit* or an *implicit* manner, as demonstrated in Fig. 1. Note that, in this paper we employ Recurrent Neural Networks (RNNs) as the recognition model because of its efficient learning capability and great success in emotion recognition [20, 22].

3.1. Explicit fusion

As illustrated in Fig. 1(a)-(c), audio features f_A and visual features f_V are explicitly fused in conventional multi-modal emotion recognition systems. More specifically, the concatenation of the two modalities takes place in different levels, i. e., $[f_A; f_V]$ in early fusion, voting based on p_A and p_V in late fusion, and $[r_A; r_V]$ in model-level fusion.

As a consequence, the obtained model can be applied to predict emotions for a given instance, if and only if both f_A and f_V are available as inputs of the model.

3.2. Implicit fusion

The proposed implicit fusion approach for joint audiovisual training is illustrated in Fig. 1(e).

Let us denote an audio feature vector as $f_A \in \mathbb{R}^M$ and its corresponding visual feature vector as $f_V \in \mathbb{R}^N$, where M and N are the dimensions of the audio and visual vectors, respectively. As depicted in Fig. 1(e), f_A and f_V are fed into two specific layers blocks, the process of which can be formulated as follows:

$$r_A = SP_A(f_A),\ r_V = SP_V(f_V), \tag{1}$$

where the function $SP_A(\cdot) : \mathbb{R}^M \rightarrow \mathbb{R}^K$ and the function $SP_V(\cdot) : \mathbb{R}^N \rightarrow \mathbb{R}^K$ map each input of different modalities into the same subspace, resulting in corresponding K-dimension representations r_A and r_V. After that, the following shared layers are applied to estimate the final predictions, and this process can be formulated as follows:

$$p_A = SH(r_A),\ p_V = SH(r_V), \tag{2}$$

where the function $SH(\cdot) : \mathbb{R}^K \rightarrow \mathbb{R}$ estimates final predictions p_A and p_V, separately.

To efficiently aggregate the advantages of different modalities for mono-modal emotion recognition (i. e., speech emotion recognition or facial emotion recognition), the model is trained with a set of audiovisual features $\{(f_A, f_V)\}$. When the model is applied for speech emotion recognition, the joint loss function $\mathcal{J}(\theta)$ is calculated by:

$$\mathcal{J}(\theta) = \mathcal{L}_A + \alpha \cdot \mathcal{L}_V, \tag{3}$$

where θ denotes the network parameters to be optimised, $\mathcal{L}_A$ and $\mathcal{L}_V$ stand for the loss of audio and visual data, respectively, and α denotes the weight of visual prediction loss to regulate its contribution to $\mathcal{J}(\theta)$. The term $\alpha \cdot \mathcal{L}_V$ enforces the optimisation to take the auxiliary modality information into account. Similarly, for facial emotion recognition, the joint loss function in Eq. (3) is altered into

$$\mathcal{J}(\theta) = \mathcal{L}_V + \alpha \cdot \mathcal{L}_A. \tag{4}$$

Moreover, the value of α is optimised on the development set, by achieving a best performance for the selected modality.

One may note that, the structure of implicit fusion is similar to model-level fusion, as both possess specific layers to learn r_A and r_V, followed by shared layers to provide a final prediction. However, different from model-level fusion, r_A and r_V are fed into shared layers alternatively when training the implicit fusion model, rather than the concatenation of the two. As a result, the final prediction is still for each single modality, i. e., p_A for audio or p_V for video.

In addition, the structure of implicit fusion is about the same as multi-task learning, which is given in Fig. 1(d), where the representation r can be learnt from the original feature f via shared layers. In multi-task learning, outputs from an auxiliary task p_2 are utilised to update the shared parameters, in order to better estimate p_1. Correspondingly, in implicit fusion, inputs from an auxiliary modality are

exploited implicitly, through optimising the parameters of the shared layers.

4. EXPERIMENTS AND RESULTS

This section is devoted to empirically investigating the proposed implicit fusion approach for *categorical emotion classification* and *dimensional emotion regression*.

4.1. Databases and Features

4.1.1. OMG-Emotion

As to *categorical emotion classification*, the One-Minute Gradual-Emotional (OMG-Emotion) Behavior dataset [23] was employed. The OMG-Emotion dataset is composed of 567 emotional monologue videos collected from Youtube, with an average length of one minute. These videos were then divided into utterance-level clips, and annotated by at least five annotators [23]. Seven categorical emotions were considered, i. e., *neutral, happiness, sadness, anger, surprise, fear,* and *disgust*. Majority voting was then applied to compute the gold standard based on all annotations of the same segment. Moreover, the dataset is split into the training, development, and test sets, resulting in 2440, 617, and 2229 segments for each partition, respectively. Note that, in this work, we performed experiments and reported performances only on the development set, as labels of the test set are not yet accessible.

To extract acoustic features on the OMG-Emotion dataset, we used the eGeMAPs feature set [24], resulting in 88 features for each utterance, same as done with RECOLA (cf. Section 4.1.2). For visual representations, firstly MTCNN [25] was applied for face detection and alignment on each frame. After that, frame-level intermediate deep features of size 4096 were extracted from the "fc-7" layer of the VGG-Face model [26], which was pre-trained on a large number of facial images. Finally, an average pooling was conducted on all frames of the same utterance to deliver an utterance-level representation.

4.1.2. RECOLA

For *dimensional emotion regression*, we utilised RECOLA, a standard database previously applied in the Audio/Visual Emotion Challenge (AVEC) in 2015, 2016, and 2018 [11, 21, 27]. The RECOLA dataset consists of audiovisual recordings of spontaneous and natural interactions from 27 French-speaking participants in order to investigate affective and social behaviours expressed by humans in real-life conditions from multimodal cues. Moreover, time- and value-continuous dimensional emotion annotations in terms of *arousal* and *valence* are given with a constant frame rate of 40 ms for the first five minutes of each recording, by averaging all six annotators and meanwhile taking the inter-evaluator agreement into consideration [28]. The dataset is further equally divided into three disjoint parts, by balancing the gender, age, and mother tongue of the participants. Therefore, each part contains nine unique recordings, resulting in 67.5 k segments in total for each part (training, development, or test).

For a fair comparison with other methods on RECOLA, we employed the same acoustic and visual features as the features provided in the AVEC challenges. In particular, 88 acoustic features and 632 geometric visual features were obtained for each segment. For full details on the RECOLA database and feature sets, please refer to [11, 21].

4.2. Implementation and Evaluation

The extracted features were first standardised in an online manner. That is, the means and variances were computed on the training partition, which were then applied over the corresponding development and test partitions for standardisation.

Table 1. Performance of the proposed and other models for classifying seven emotional categories in terms of F1 on the development set with the OMG-Emotion dataset. Results that obtain the best performance are highlighted.

F1 [%]	audio	video
SVM [23]	33.0	—
CNN [23]	—	37.0
RNN (baseline)	36.5	37.9
RNN (implicit fusion)	**40.2**	**42.1**

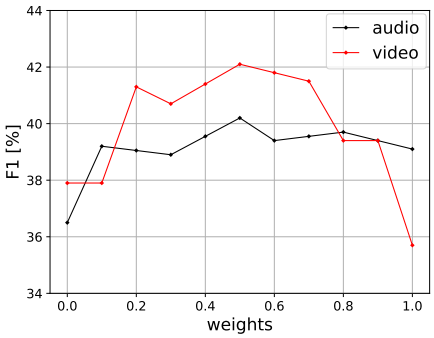

Fig. 2. Impact of the weights of the auxiliary modality on the performance (F1) when training jointly with both audio and visual signals and then evaluating with only audio or only video.

In these preliminary experiments, we constructed the frameworks with RNN equipped with Gated Recurrent Unit (GRU), each hidden layer consisting of 100 nodes. In particular for OMG-Emotion, we employed one hidden layer for each modality-specific branch, followed by one shared hidden layer. For RECOLA, more data were available in training, thus two hidden layers for each modality and another two shared hidden layers were applied. These settings were selected based on our previous empirical experiences [22]. In the joint training process, the network was trained with an Adam optimiser with an initial learning rate of 0.001.

Besides, the weight α was optimised in the range of [0.0, 1.0] by a grid search with a step size of 0.1. To accelerate the training process, the network parameters were updated for every minibatch of 128 audiovisual instances. It should be noticed that, only features from a single modality were utilised on the development and test partitions, to simulate the scenarios where only mono-modal data is accessible.

Furthermore, for the baseline, we performed the training on each single modality independently. That was done by setting α to be 0.0 in Eq. (3) and Eq. (4), respectively.

Additionally, we conducted the same operations of annotation delay compensation and post-processing procedure of predictions on the RECOLA database, following previous works in [19–21].

Finally, to measure the performance of the frameworks, we utilised the metrics of F1 for OMG-Emotion and Concordance Correlation Coefficient (CCC) for RECOLA, as suggested by previous studies in [21, 23]. For a more in-depth explanation of CCC, the reader is referred to [21]. In general, a higher F1 or CCC indicates a better prediction performance.

4.3. Results and Discussion

4.3.1. Results on OMG-Emotion

For our experiments on OMG-Emotion, we conducted seven-class categorical emotion classification tasks on audio and visual signals.

Table 2. Concordance Correlation Coefficient (CCC) on RECOLA when evaluating via individual *audio* and *video* modalities on the development (*dev.*) and *test* sets in the dimensions of *arousal* and *valence*, respectively. Results were reported for the proposed implicit fusion approach, corresponding baselines and other state-of-the-art approaches. The Multi-Task Learning (MTL) frameworks, and the Dynamic Difficulty Awareness Training (DDAT) frameworks both have two variants by exploring reconstruction error (RE) and perception uncertainty (PU), respectively. Results that obtain the best performance on the test set are highlighted.

CCC	arousal				valence			
	audio		video		audio		video	
	dev.	test	dev.	test	dev.	test	dev.	test
SVR [11]	.796	.648	.379	.272	.455	.375	.612	.507
DNN+Curriculum learning [29]	.687	.591	.394	.267	.159	.174	.300	.269
MTL (RE based) [30]	.788	.629	.502	.324	.519	.331	.632	.488
MTL (PU based) [12]	.803	.654	.508	.327	.506	.416	.643	.452
DDAT (RE based) [22]	.807	**.694**	.544	.400	.508	**.422**	.639	.471
DDAT (PU based) [22]	.811	.664	.513	.397	.498	.407	.632	.501
RNN (baseline)	.766	.605	.499	.399	.504	.381	.619	**.529**
RNN (implicit fusion)	.769	.611	.515	**.413**	.513	.395	.622	.527

Table 1 presents the performance of the models in terms of F1 on the development set. From the table, we can observe that on this database, our RNN-based baseline models outperform the other methods reported in the literature [23], i.e., Support Vector Machine (SVM) and Convolutional Neural Network (CNN). More specifically, for speech emotion classification our baseline yields higher F1 than SVM (36.5% vs 33.0%), and our baseline performs better than CNN (37.9% vs 37.0%) in the video modality.

Additionally, comparing the performance achieved by our corresponding baselines and the proposed implicit fusion model, it is noticed that, the latter approach performs better than the former one by a large margin, i. e., 40.2% vs 36.5% for audio and 42.1% vs 37.9% for video. These experimental results may indicate that, the proposed implicit fusion approach is plausible to promote performances of mono-modal emotion classification further when information from auxiliary modalites are integrated during training.

Furthermore, to demonstrate the effect of the weights of the auxiliary modality α for the performance of emotion prediction, we computed the performance in terms of F1 for each predefined value of α on the audio and video modalities, as shown in Fig. 2. When $\alpha = 0.0$, i. e., no contribution from the auxiliarly modality, the model is learnt based on only the loss of each single modality, separately. When α increases, i. e., the contribution of the auxiliary modality during training increases, the performance of mono-modal emotion recognition (audio or video) is improved first, until a point where the contribution of the auxiliary modality might actually penalise the learning objective too much and even harm the learning of the main modality, and thus performances starts to decrease. To this end, a proper value of the weight α needs to be identified for the tasks at hand. We can observe from the figure that, the best performance for both audio and video emotion classification on the OMG-Emotion database is reaped when $\alpha = 0.5$.

4.3.2. Results on RECOLA

Table 2 presents the implicit fusion results in terms of CCC for the prediction of the arousal and valence dimensions on RECOLA. One may observe that, when implementing implicit fusion, the obtained results are consistently higher than the related baselines on the development partition. This confirms the importance of integrating auxiliary modality for mono-modal emotion recognition. Also, similar observations can be seen on the test partition in most cases (i. e., 3 out of 4 cases). This exception case is highly attributed to the mismatch between the two partitions.

Meanwhile, as illustrated in Table. 2, our approach achieves comparable or superior performance to other state-of-the-art methods applied on the RECOLA database, such as Support Vector Regressor (SVR) [11], a curriculum learning model [29], multi-task learning [12], and a Dynamic Difficulty Awareness Training (DDAT) framework [22]. Particularly, perception uncertainty-based DDAT is a more recent and promising approach, which exploits the perception uncertainty to estimate the difficulty of learning specific information, as emotion prediction is a subjective task without a ground truth [22]. Despite the fact that our proposed models are worse than the DDAT models on the audio modality, the implicit fusion can be incorporated with DDAT in the future, to further boost the performance of mono-modal emotion recognition.

Moreover, one may notice that when predicting arousal based on only video data, the best ever CCC (.413) is achieved by an implicit fusion model. This indicates that our approach can largely supply additional knowledge from audio to alleviate the weakness of video signals for arousal prediction.

5. CONCLUSION

In contrast to previous studies that train audiovisual data jointly for multi-modal emotion recognition, we, for the first time, exploited the audiovisual information to build models targeted at mono-modal scenarios. By adding an addition loss from an auxiliary modality to penalise the learning process, complementary information from other modalities is combined implicitly. The proposed methods were evaluated on two datasets for emotion classification and regression, respectively. Experimental results have demonstrated that the proposed methods clearly improve the prediction performance of a mono-modal model by involving a complementary modality into its learning process.

In the future, we will investigate the efficiency of the proposed implicit fusion in other applications, such as speech recognition and scene detection. Additionally, more auxiliary modalities will be taken into consideration in future to further facilitate speech/facial emotion prediction, e. g., the head position and physiological signals. Furthermore, we also plan to combine the implicit fusion with the dynamic difficulty awareness training [22] to advance mono-modal emotion recognition.

6. REFERENCES

[1] R. W. Picard, *Affective Computing*, MIT press, Cambridge, MA, 1997.

[2] J. Han, Z. Zhang, N. Cummins, and B. Schuller, "Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives," *IEEE Computational Intelligence Magazine*, 2018, 13 pages.

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[4] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *Proc. IEEE International Colloquium on Signal Processing and its Applications*, Penang, Malaysia, 2011, pp. 410–415.

[5] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *Proc. ICASSP*, Brisbane, Australia, 2015, pp. 4749–4753.

[6] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[7] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *Proc. ICASSP*, New Orleans, LA, 2017, pp. 5115–5119.

[8] J. Huang, Y. Li, J. Tao, Z. Lian, and J. Yi, "End-to-end continuous emotion recognition from video using 3D ConvLSTM networks," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 6837–6841.

[9] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. ICMI*, State College, PA, 2004, pp. 205–211.

[10] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multicultural dimensional continues emotion recognition in dyadic interactions," in *Proc. AVEC*, Seoul, South Korea, 2018, pp. 65–72.

[11] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. AVEC*, Amsterdam, Netherlands, 2016, pp. 3–10.

[12] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proc. ACM MM*, Mountain View, CA, 2017, pp. 890–897.

[13] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, Jan. 2017.

[14] A. Mehrabian, "Communication without words," *Communication Theory*, pp. 193–200, 2008.

[15] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[16] Y. Wang and L. Guan, "Recognizing human emotion from audiovisual information," in *Proc. ICASSP*, Philadelphia, PA, 2005, pp. 1125–1128.

[17] 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, Apr. 2011.

[18] 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 *Proc. AVEC*, Brisbane, Australia, 2015, pp. 73–80.

[19] 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 *Proc. AVEC*, Brisbane, Australia, 2015, pp. 41–48.

[20] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller, "Strength modelling for real-world automatic continuous affect recognition from audiovisual signals," *Image and Vision Computing*, vol. 65, pp. 76–86, Sep. 2017.

[21] 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 *Proc. AVEC*, Brisbane, Australia, 2015, pp. 3–8.

[22] Z. Zhang, J. Han, and B. Schuller, "Dynamic difficulty awareness training for continuous emotion prediction," *IEEE Transactions on Multimedia*, vol. PP, Sep. 2018, 13 pages.

[23] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A.r Sutherland, and S. Wermter, "The OMG-emotion behavior dataset," in *Proc. IJCNN*, Rio, Brazil, 2018, pp. 1408–1412.

[24] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "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, Apr. 2016.

[25] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[26] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. British Machine Vision Conference*, Swansea, UK, 2015, pp. 1–12.

[27] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, E. Ciftçi, H. Güleç, A. Salah, and M. Pantic, "AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proc. AVEC*, Seoul, South Korea, 2018, pp. 3–13.

[28] F. Ringeval, A. Sonderegger, J. S. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. FG*, Shanghai, China, 2013, pp. 1–8.

[29] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *arXiv preprint arXiv:1805.10339*, May 2018.

[30] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech," in *Proc. ICASSP*, New Orleans, LA, 2017, pp. 2367–2371.