An Audiovisual Child Emotion Recognition System for Child-Robot Interaction Applications

Panagiotis P. Filntisis[1], Niki Efthymiou[1], Gerasimos Potamianos[2], and Petros Maragos[1]
[1]*School of ECE, National Technical University of Athens, 15773 Athens, Greece*
[2]*Department of ECE, University of Thessaly, 38221 Volos, Greece*
{filby,nefthymiou}@central.ntua.gr, gpotam@ieee.org, maragos@cs.ntua.gr

***Abstract*—We present an audiovisual emotion recognition system tailored to child-robot interaction scenarios. Our proposed system is based on deep learning and the Temporal Segment Networks framework, receives input from both the child's speech and video modalities (the latter represented as RGB and optical flow streams), and tackles several challenges that arise in emotion recognition and child-robot interaction. The system is evaluated on the EmoReact child emotion recognition dataset, significantly outperforming the state-of-the-art on this corpus. In addition, extensive ablation studies are conducted.**

***Index Terms*—Audiovisual Emotion Recognition, Multimodality, Child-Robot Interaction, Audiovisual Fusion**

I. Introduction

Emotion is one of the fundamentals of human communication, and its significance in our social lives has kickstarted numerous studies on its automatic recognition, focusing on various applications that range from support lines and call centers, to e-Health, education, and human-robot interaction.

Concerning the latter, during the interaction between a human and a social robot, the analysis of the human emotional state should have an important role in order to develop empathic robotic agents [1]. Emotional awareness of robotic agents allows them to adapt their behavior towards the human based on the perceived affect, creating a more human-like and natural communication. Especially, as far as Child-Robot Interaction (CRI) is concerned, it has been shown that robotic actions in accordance with children's emotion create a positive and stronger bond between the two, increasing trust and establishing long-term interactions [2], [3].

Building an emotion recognition system for children is challenging and presents many obstacles. Children not only differ from adults in their natural characteristics (e.g., voice pitch, body height) but also exhibit different behavioral patterns, which for example can result in abrupt movements and occlusions [4], [5]. To counter these, a robust system for child emotion recognition should leverage information from multiple modalities, exploiting the fact that different emotions can be expressed through different information channels. Further, recognition systems should be computationally efficient, especially in the context of real-life CRI scenarios. Finally, an additional challenge concerns general lack of high-quality, large children emotion datasets [5] that are crucial in developing state-of-the-art deep learning supervised techniques. Children corpora tend to be of small size due to the fact that they are hard to obtain, one of the more important reasons being data sensitivity.

In this paper, we present an audiovisual emotion recognition system aiming to address the aforementioned challenges. The system takes as input both the child's speech, as well as the visual channel, in the form of the raw RGB data stream, which can be used to effectively identify static facial expressions, and the optical Flow stream, which is effective in modeling the dynamics of emotions. This selection of different modalities is verified by ablation studies that analyze the contribution of each modality for the prediction of different emotions and identify the most effective fusion scheme that can be used to combine information from all channels. In addition, the deep learning based methods that we have employed allow for computationally efficient training and inference, and they can be developed on small datasets, avoiding overfitting. We perform extensive ablation studies on the EmoReact dataset, which, to the best of our knowledge, is the only dataset of children expressing emotions both verbally and visually, and establish a good trade-off between computational load and system performance. Finally, our approach is verified by comparing our system to the previous best published results on the EmoReact dataset, significantly outperforming them.

The rest of the paper is organized as follows: Section II discusses related work on emotion recognition for children and CRI. Section III describes in detail our proposed audiovisual emotion recognition architecture, and Section IV presents our thorough experimental results on the EmoReact dataset. Finally, Section V provides our conclusions and directions for future work.

II. Related Work

Recognition of children affect, as we mentioned above, is crucial in creating empathic robots deployed during CRI. Thus, interesting research works have been presented for designing advanced emotion recognition modules to equip intelligent robots. Goulart et al. proposed in [6] a computational system for estimating children emotion during CRI, deploying visual information from both RGB and infrared thermal cameras.

This research is carried out/funded in the context of the project "Intelligent Child-Robot Interaction System for designing and implementing edutainment scenarios with emphasis on visual information" (MIS 5049533) under the call for proposals "Researchers' support with an emphasis on young researchers-2nd Cycle". The project is co-financed by Greece and the European Union (European Social Fund- ESF) by the Operational Programme Human Resources Development, Education and Lifelong Learning 2014-2020.

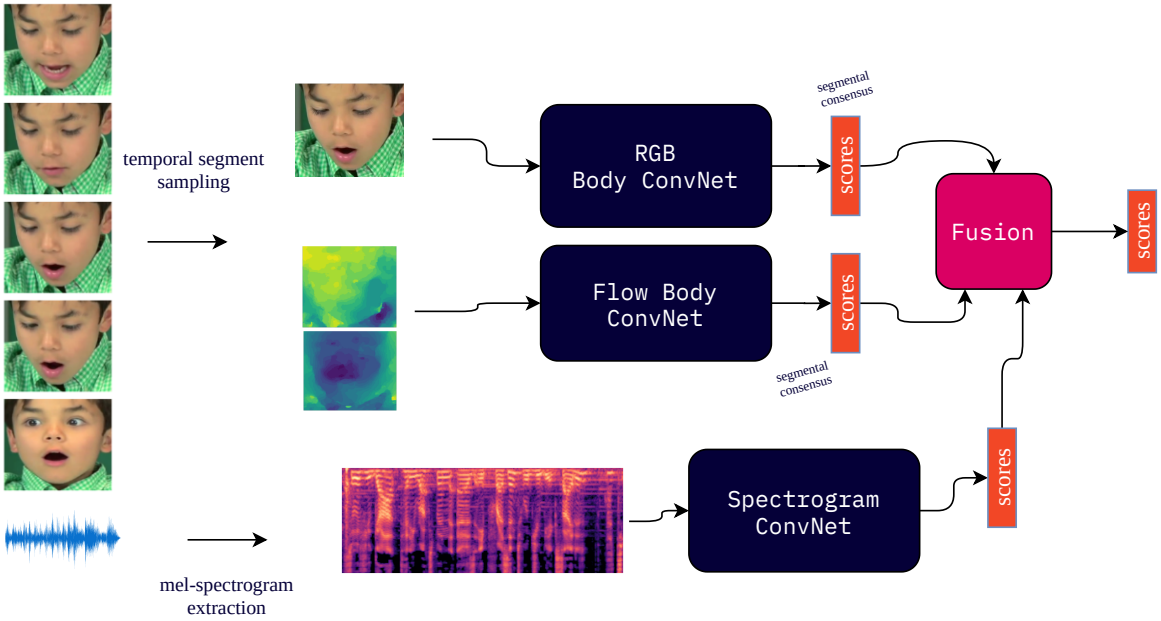

Fig. 1. The proposed multimodal emotion recognition architecture for child-robot interaction.

The proposed system detects the facial regions of interest that are relevant to five basic emotions. Lopez-Rincon in [7] proposed a Convolutional Neural Network (CNN) combined with a Viola-Jones face detector, trained using the AffectNet database [8], and tuned it with children data in order to recognize children facial emotional expressions. Marinou et al. [9] proposed an automated approach using 3d skeleton data and a CNN architecture for action and continuous emotion recognition during robot-assisted therapy sessions of children with Autism Spectrum Disorders (ASD). In [10], a system perceived children affective expressions while playing chess with an iCat robot and modified the behavior of the robot to be more friendly and increase children engagement. Similarly, Filippini et al. [11] classified children emotional states to understand their engagement level using thermal signal analysis during interaction with the Mio Amico Robot. An adaptive robot behavior based on the perceived emotional responses was also developed for a NAO robot in [12].

A number of studies have emphasized the importance of leveraging multiple modalities for emotion recognition in adults [13]–[15]. In [16], 3D CNN were used to extract deep spatiotemporal features from both the video and audio (represented as short-time Fourier transform) in order to determine emotion scores. Kim et al. [17] used deep belief networks for audiovisual feature generation, while [18] combined a two-branch feature extraction scheme with a long short-term memory network for continuous dimensional emotion recognition.

On the contrary, there is a lack of works studying multiple modalities for emotion recognition in children. Apart from the face, which is the most commonly used channel for identifying emotion [19], there are other modalities equally powerful to reveal children affect such as speech and body movements. In [20], an ensemble of AlexNet networks was applied on multiple spectrograms in order to extract deep features, which were then used by an SVM to identify emotions in the EmoReact dataset. For the same dataset, [5] combined traditional audio features and features extracted from the OpenFace framework [21] (action units, shape parameters and head orientation) with an SVM for audiovisual emotion recognition. In [4], we proposed a two-branch architecture modeling body movements along with the facial expressions to identify emotions in children during CRI scenarios. In comparison to that work, here we investigate a different modality (audio) instead of the body skeleton, in order to tackle occlusions and increase robustness, and also use a different dataset (EmoReact), which includes both visual and aural expression of emotion. Furthermore, while the CNN architecture in that work considered each frame in the video separately, here we use CNN architectures that take into account multiple frames in the video using temporal sampling, as well as leverage the video dynamics using the optical flow representation.

III. Method

The architecture of the proposed emotion recognition system is shown in Figure 1. The system is composed of different branches, each one focusing on a different input channel/modality.

A. Visual Branch

The visual branch is based on the Temporal Segments Network (TSN) framework [22]. During training, the input video is split into K different segments of equal duration M, and in the next step, a snippet of length $N < M$ consecutive frames is randomly sampled from each segment, resulting in K snippets T_k. Subsequently, each snippet is fed to a CNN, yielding class scores S_k for each snippet. In the last step, the

scores of the different snippets are fused using the segmental consensus function H that is applied on the representations of all different snippets to obtain the final scores:

$$S = H(S_k) = H(F_v(T_k; \boldsymbol{W_v})|_{k \in K}) \quad (1)$$

where $F_v(T_k; \boldsymbol{W_v})$ denotes the application of a CNN with parameters $\boldsymbol{W_v}$ on the snippet T_k. The most common consensus function that can be used is averaging, while others include maximum or weighted averaging (we use simple averaging). The CNN is then trained using standard cross-entropy loss in the case of multiclass classification, or binary cross-entropy in the case of multilabel classification (which is the case of emotion recognition we consider).

Traditionally, TSNs take input from both the RGB of the input video, as well as the optical flow, with each one trained separately and then fused using average or weighted average fusion. As with TSNs for action recognition, we also use both modalities, since the optical flow can be used to model the dynamics that arise during expressions of emotion, while the RGB modality can best identify static expressions such as smiles. We also need to mention that we crop the input video (both RGB and Flow) around the child's face, by using the facial landmarks obtained by OpenFace [21].

The paradigm of TSNs offers several benefits to emotion recognition and CRI in particular. Considering an input video with a child expressing emotion, the archetype facial expressions and action units that correspond to each emotion are not present throughout the video, but usually only during a short period of it. As a result, temporal sampling allows the network to access several parts of the video and model its long-range temporal structure, thus being more likely to observe the corresponding facial expression. In addition, compared to processing the entire video, the sampling process ignores redundant information in consecutive video frames, helping avoid overfitting and offering a type of data augmentation, valuable for children emotion databases of small size.

Finally, since the ultimate goal of the system is its deployment in real-life CRI scenarios, it is important to consider computational costs of training, as well as the ability to run in real time. Due to the fact that the system does not consider the entire input video chunk, the computational load is reduced significantly, both at training, as well as during inference.

B. Audio Branch

In the audio branch, considering the input waveform of the video, we first extract its mel-spectrogram representation and then apply a CNN $F_a(\boldsymbol{W}_a)$ on it in order to extract the audio representation. Here, we bypass the cumbersome feature extraction methods by considering the mel-spectrogram of the waveform as an image, and applying standard computer vision techniques. Next, as with the visual modality, a fully connected layer is used in order to obtain the final emotion scores. The audio branch is susceptible to overfitting because the full spectrogram is fed to the network, contrary to the visual branch where temporal sampling is used. To counter this, we apply a more aggressive regularization scheme with high penalty for L2 regularization during training.

C. Training and Audiovisual Fusion

In order to fuse information from the visual and audio modalities, we consider two different types of fusion between both RGB-audio, as well as Flow-audio modalities: feature fusion and score fusion, and two training schemes: independent training and joint training.

During joint training, the RGB (or Flow) and audio CNN are trained concurrently, and depending of the fusion scheme, we either concatenate their feature vectors (feature fusion) before the last fully connected layer, or average the scores (score fusion) obtained after the last fully connected layer. In order to achieve feature fusion under joint training, we repeat the audio feature vector K times (where K is the number of segments/snippets), and associate each visual snippet with the audio feature vector for the whole video, through concatenation of the feature vectors. In contrast, in independent training the RGB (or Flow) and audio networks are trained separately, and we then average their emotion scores.

IV. Experimental Framework and Results

A. Database

The dataset we use is the EmoReact dataset. The EmoReact dataset [5] contains videos of 63 children (32F, 31M, aged 4 to 14) reactions to different topics, and has been collected from the YouTube channel React. The number of all videos across the training (432 videos), validation (303 videos), and test set (367) is 1102. Each video is annotated with one or more emotions, from a total of 8 emotion labels: Curiosity, Uncertainty, Excitement, Happiness, Surprise, Disgust, Fear, and Frustration. To the best of our knowledge, the EmoReact dataset is the only dataset of children expressing emotion, both verbally and visually.

B. Implementation Details

The CNN backbone of the visual and audio branches is a residual CNN with 50 layers (ResNet50) [23]. Specifically for the CNN of the visual RGB branch, we have pretrained it on the largest facial expression dataset, AffectNet [8], achieving 59.47% accuracy on the validation set (test set is not available). Because the label distribution of AffectNet is highly skewed, we employ balanced sampling so that the network sees the underrepresented classes more often. The residual networks of the audio branch and Flow modality are pretrained on ImageNet (we obtain the weights of the network as provided by the PyTorch framework).

We train all networks and modalities with stochastic gradient descent for 60 epochs, starting with a learning rate of 1e-2, momentum 0.9, and regularization with weight decay (L2 regularization) 5e-4. The learning rate is reduced by a factor of 10 at 20 and 40 epoch milestones[1]. Training is done using binary cross-entropy loss. For evaluation, we select the

[1]We have made the code for the experiments publicly available at https://github.com/filby89/multimodal-emotion-recognition

TABLE I
ROC AUC AND AVERAGE TIME ELAPSED PER EPOCH WITH VARYING NUMBER OF SAMPLED SNIPPETS.

Segments	ROC AUC		sec/train epoch	sec/val epoch
	Balanced	Unbalanced		
RGB				
1	0.685	0.773	11	7
3	0.713	0.786	27	20
5	0.709	0.787	40	26
10	0.715	0.788	73	51
Flow				
1	0.585	0.741	37	23
3	0.596	0.744	101	70
5	0.623	0.757	166	115
10	0.627	0.759	294	210

TABLE II
RESULTS ON THE EMOREACT DATASET FOR DIFFERENT FUSION AND TRAINING SCHEMES BETWEEN THE RGB-AUDIO AND FLOW-AUDIO MODALITIES.

Fusion	Training	ROC AUC	
		Balanced	Unbalanced
Single Modality	Audio	0.715	0.750
	Visual (RGB)	0.713	0.786
	Visual (Flow)	0.623	0.757
Score Fusion RGB-audio	Joint Training	0.720	0.756
	Independent Training	**0.747**	**0.799**
Score Fusion Flow-audio	Joint Training	0.719	0.746
	Independent Training	0.725	0.787
Feature Fusion RGB-audio	Joint Training	0.719	0.769
Feature Fusion Flow-audio	Joint Training	0.707	0.744

epoch with the best validation area under receiver operating characteristic (ROC AUC), and apply the corresponding network on the test set, reporting class-balanced and unbalanced ROC AUC. Especially in the case of audio, we found out that a more aggressive regularization scheme is needed to avoid overfitting, and thus we increased the weight decay tenfold to 5e-3.

C. Results

a) Number of segments: As a first ablation study, we consider the number of segments (and as a consequence the number of snippets), which are used during training of the visual branch with the RGB and Flow modalities. We consider 4 different values: 1, 3, 5, and 10, and report in Table I the results on the ROC AUC (balanced per class and unbalanced), as well as average time taken per epoch for training and inference, on a computer with an RTX 2080 GPU.

We can see that in the case of RGB, increasing the number of segments above 3 does not result in significant performance difference, showing that even a small number of segments can achieve satisfactory performance. However, increasing the number of segments increases significantly both the training and inference times. For the Flow modality, we see that selecting 5 as a number of segments results in a balanced trade-off between performance and execution time, since the performance increase using 10 segments is minuscule. For the following experiments, we use 3 segments for RGB and 5 segments for the Flow modality.

TABLE III
FINAL ROC AUC RESULTS ON THE EMOREACT DATASET.

	ROC AUC	
	Balanced	Unbalanced
Audio		
audio features + SVM [5]	0.610	-
dnn ensemble + SVM [20]	0.718	-
Ours (End-to-End)	0.715	0.750
Visual		
openface + SVM [5]	0.620	-
Ours (Flow)	0.623	0.757
Ours (RGB)	0.713	0.786
AudioVisual		
[5]	0.640	-
Ours (RGB+Audio+Flow)	**0.754**	**0.809**

b) Audiovisual fusion and training schemes: Next, we experiment with the different kinds of fusion schemes that can be used to merge the RGB and audio, as well as the Flow and audio modalities: feature vs. score fusion, as well as the pretraining scheme: joint training of both networks vs. independent training. The results of this study are shown in Table II. Training the networks independently and then averaging their scores achieves the best result in both cases of audiovisual fusion (RGB-audio and Flow-audio), when compared to both the single modalities, as well as their fusion using joint training. This could be attributed to the fact that while the TSN framework inherently avoids overfitting using the temporal sampling, in the case of audio this is not the case, since the full spectrogram is used, and more elaborate schemes of regularization are needed.

c) Emotion by modality: Next, we explore the strengths and weaknesses of each different modality, by showing the different ROC AUC scores for each emotion, in Figure 2. We observe that especially for Happiness, RGB is the most appropriate modality, while Fear and Disgust, are best identified through the children's speech. Flow, in almost all cases underperforms when compared to the other modalities, however in the case of Excitement and Surprise it achieves a high score, which can be explained by the more intense movements a person does when expressing these emotions. The figure also shows the result of average score fusion using independent training for all three modalities, RGB, Flow, and audio. We can see that overall, fusion increases the total balanced and unbalanced scores, however in the case of Uncertainty, Excitement, and Happiness, it results in slightly lower score when compared to RGB only.

d) Final Results: We present the final results of the emotion recognition system on EmoReact in Table III, where we have also added the result of average score fusion between the three different modalities (using independent training), as well as the previous reported best results in the literature. For the audio modality, our architecture achieves significantly better ROC AUC than [5], which used a carefully selected speech features set with an SVM, as well as similar results with Nagarajan et al. [20]. However, our approach is end-to-end and simple to implement, while Nagarajan et al. employed

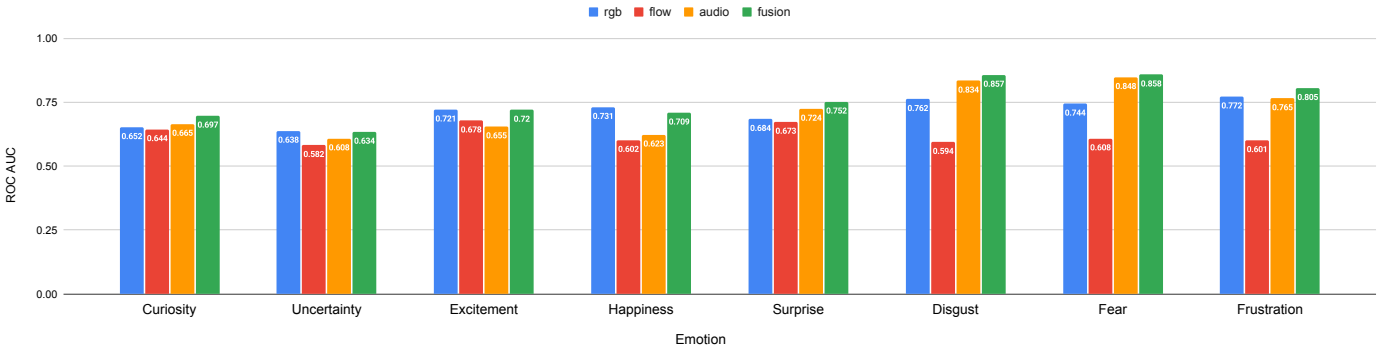

Fig. 2. ROC AUC per emotion, for each different modality and their average score fusion.

an elaborate scheme involving multiple AlexNet architectures for feature extraction and an SVM on top of them to achieve the final result.

In the visual modality, our RGB TSN architecture improves significantly upon the best previous published result, which used features extracted from the OpenFace framework with an SVM [5].

Finally, our audiovisual fusion scheme using all three modalities with independent training further increases the ROC-AUC up to 0.754, resulting in significant score improvement upon all previous studies.

V. Conclusion

In this paper we proposed a novel multimodal emotion recognition system that can be used for deducing the emotion of children, with the ultimate goal being child-robot interaction scenarios. To that end, we have used deep learning methods that tackle challenges met in CRI: small datasets, real-time inference, and computationally low-cost training. We have also thoroughly explored several aspects of our architecture and identified the contribution of different parts of our network to the final outcome. We have evaluated the emotion recognition system on the EmoReact dataset of children expressing their emotions multimodally, and showed that it achieves high performance and state-of-the art results. In the future, we aim to conduct extensive emotion recognition evaluations in real-life CRI scenarios with numerous children and custom use-cases.

References

[1] K. Hone, "Empathic agents to reduce user frustration: The effects of varying agent characteristics," *Interacting with computers*, vol. 18, pp. 227–245, 2006.

[2] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Empathic robots for long-term interaction," *International Journal of Social Robotics*, vol. 6, pp. 329–341, 2014.

[3] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Transactions on Computer-Human Interaction*, vol. 12, pp. 293–327, 2005.

[4] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction," *IEEE Robotics and Automation Letters*, vol. 4, pp. 4011–4018, 2019.

[5] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "EmoReact: a multimodal approach and dataset for recognizing emotional responses in children," in *Proc. ICMI*, 2016.

[6] C. Goulart, C. Valadão, D. Delisle-Rodriguez, D. Funayama, A. Favarato, G. Baldo, V. Binotte, E. Caldeira, and T. Bastos-Filho, "Visual and thermal image processing for facial specific landmark detection to infer emotions in a child-robot interaction," *Sensors*, vol. 19, pp. 2844, 2019.

[7] A. Lopez-Rincon, "Emotion recognition using facial expressions in children using the NAO robot," in *Proc. CONIELECOMP*, 2019, pp. 146–153.

[8] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, pp. 18–31, 2017.

[9] E. Marinoiu, M. Zanfir, V. Olaru, and C. Sminchisescu, "3D human sensing, action and emotion recognition in robot assisted therapy of children with autism," in *Proc. CVPR*, 2018.

[10] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. Mcowan, "Multimodal affect modeling and recognition for empathic robot companions," *Int. Journal of Humanoid Robotics*, vol. 10, pp. 1350010, 2013.

[11] C. Filippini, E. Spadolini, D. Cardone, D. Bianchi, M. Preziuso, C. Sciarretta, V. del Cimmuto, D. Lisciani, and A. Merla, "Facilitating the child–robot interaction by endowing the robot with the capability of understanding the child engagement: The case of Mio Amico robot," *International Journal of Social Robotics*, pp. 1–13, 2020.

[12] M. Tielman, M. Neerincx, J.J. Meyer, and R. Looije, "Adaptive emotional expression in robot-child interaction," in *Proc. HRI*, 2014.

[13] H. Bänziger, T.and Pirker and K. Scherer, "Gemep-geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions," in *Proc. LREC*, 2006, vol. 6, pp. 15–19.

[14] L. C. De Silva, "Audiovisual emotion recognition," in *Proc. Int. Conf. on Systems, Man and Cybernetics*, 2004.

[15] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang, "Affective multimodal human-computer interaction," in *Proc. Int. Conf. on Multimedia*, 2005.

[16] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes, "Deep spatio-temporal features for multimodal emotion recognition," in *Proc. WACV*, 2017.

[17] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. ICASSP*, 2013.

[18] P. Tzirakis, G. Trigeorgis, M. A Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *Journal of Selected Topics in Signal Processing*, vol. 11, pp. 1301–1309, 2017.

[19] B. De Gelder, "Why bodies? twelve reasons for including bodily expressions in affective neuroscience," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 364, pp. 3475–3484, 2009.

[20] B. Nagarajan and V. R.M. Oruganti, "Cross-domain transfer learning for complex emotion recognition," in *Proc. TENSYMP*, 2019.

[21] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proc. FG*, 2018, pp. 59–66.

[22] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. ECCV*. Springer, 2016.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.