Audiovisual Classification of Group Emotion Valence Using Activity Recognition Networks

1st João Ribeiro Pinto
INESC TEC and Universidade do Porto
Porto, Portugal
joao.t.pinto@inesctec.pt

2nd Tiago Gonçalves
INESC TEC and Universidade do Porto
Porto, Portugal
tiago.f.goncalves@inesctec.pt

3rd Carolina Pinto
Universidade do Porto
Porto, Portugal
up201506006@fe.up.pt

4th Luís Sanhudo
INESC TEC and Universidade do Porto
Porto, Portugal
luis.sanhudo@inesctec.pt

5th Joaquim Fonseca
Bosch Car Multimedia
Braga, Portugal
joaquim.fonseca2@pt.bosch.com

6th Filipe Gonçalves
Bosch Car Multimedia
Braga, Portugal
filipe.goncalves@pt.bosch.com

7th Pedro Carvalho
INESC TEC and Polytechnic of Porto
Porto, Portugal
pedro.m.carvalho@inesctec.pt

8th Jaime S. Cardoso
INESC TEC and Universidade do Porto
Porto, Portugal
jaime.cardoso@inesctec.pt

2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS) | 978-1-7281-7574-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/IPAS50080.2020.9334943

***Abstract*—Despite recent efforts, accuracy in group emotion recognition is still generally low. One of the reasons for these underwhelming performance levels is the scarcity of available labeled data which, like the literature approaches, is mainly focused on still images. In this work, we address this problem by adapting an inflated ResNet-50 pretrained for a similar task, activity recognition, where large labeled video datasets are available. Audio information is processed using a Bidirectional Long Short-Term Memory (Bi-LSTM) network receiving extracted features. A multimodal approach fuses audio and video information at the score level using a support vector machine classifier. Evaluation with data from the EmotiW 2020 AV Group-Level Emotion sub-challenge shows a final test accuracy of 65.74% for the multimodal approach, approximately 18% higher than the official baseline. The results show that using activity recognition pretraining offers performance advantages for group-emotion recognition and that audio is essential to improve the accuracy and robustness of video-based recognition.**

***Index Terms*—activity, audio, deep learning, group emotion, recognition, valence, video**

I. Introduction

Emotion recognition is a fast-growing research topic, due to its potential for enhanced human-computer interfaces and automatic services that immediately respond to the emotions of the user or client [1]. Horror videogames that adapt the gameplay and soundtracks based on the player's fear, as well as autonomous vehicles that adapt the travel experience based on the occupants' emotions, are only two of the endless innovations attainable through emotion recognition [2].

State-of-the-art methods for emotion recognition are mainly based on facial expressions, and important hurdles have been overcome in this field [1], [2]. Group-level emotion is a fairly uncharted research topic that extends the analysis to the emotional state displayed by a group of people as a whole [3]. While there are several challenges in individual emotion recognition, approaches for group emotion recognition also need to deal with the variety of emotions, their valence, and arousal levels, that can differ among members of the same group. This topic was the focus of the EmotiW 2020 [4] sub-challenge that motivated this work. The scarce data and the difficulty in obtaining annotations is the reason why few have addressed this topic [5]–[7], and why current approaches still offer low accuracy levels.

The task of group emotion recognition shares some similarities with the recognition of human activity based on the video. Unlike the former, the latter boasts several large and thoroughly labeled datasets, such as the Kinetics [8] or the ActivityNet [9], even when restricting to data focused on groups rather than on individuals. These larger sets of available data have allowed for the development of very robust and high-performing algorithms, such as the I3D [8], the SlowFast networks [10], or the stagNet [11].

While methods based on visual information compose most of the literature, some works discuss the advantages of including additional sources of information, especially audio [12]–[15]. Specifically, it has been shown that using audio complements some of the flaws of video-based recognition [12], despite offering subpar accuracy results when in a unimodal recognition system. These results have confirmed the advan-

This work is supported by: European Structural and Investment Funds in the FEDER component, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) [Project no. 039334; Funding Reference: POCI-01-0247-FEDER-039334] and by the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia within the PhD grants "SFRH/BD/137720/2018", "SFRH/BD/06434/2020", and "SFRH/BD/129652/2017".

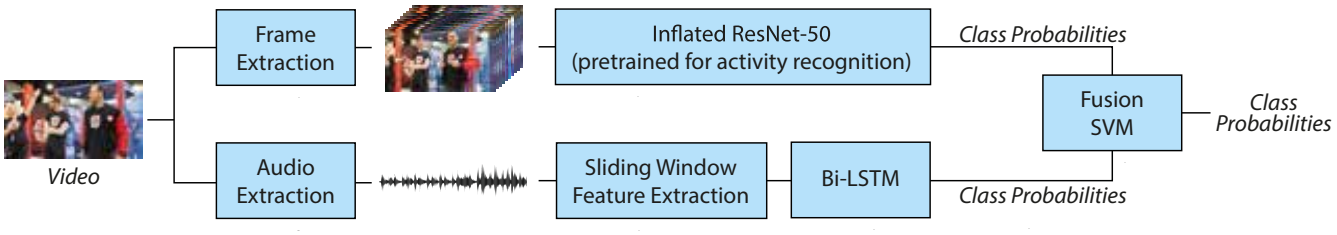

Fig. 1. Illustration of the structure of the proposed method for audio-video group emotion recognition. The proposed methodology for group emotion recognition processes a video in two streams: one processes concatenated video frames using an inflated ResNet-50 pretrained on a large activity recognition dataset, and the other extracts sliding window features from the audio and processes them using a bi-directional long short-term memory (Bi-LSTM) network. A support vector machine (SVM) classifier receives class probabilities from each stream and returns a final class prediction.

tages of combining audio information with a strong method for video-based recognition.

This work explores the novel application of inflated convolutional neural networks (CNN) to classify emotion valence at the group level in videos. The network uses weights pretrained for activity recognition, to take advantage of the greater availability of data to boost performance on our target task. We also study the use of audio for improved performance, through score-level fusion, with a Bi-LSTM network receiving spectral features. Throughout the experiments, we assess the performance of the proposed method for multimodal and unimodal classification, analyze its behavior in different scenarios, and compare it directly with the EmotiW 2020 sub-challenge official baseline.

II. Methodology

A. General overview

The proposed algorithm is composed of three modules: a video-based emotion recognition model, an audio-based emotion recognition model, and a multimodal fusion module (see Fig. 1). Video and audio-based emotion recognition modules are trained independently, while the fusion module, based on a multiclass SVM receives the softmax scores provided by the other two. Thus, the proposed method consists of a pipeline that relies on late audio-video fusion, at the score level, using a multi-class SVM emotion recognition classifier.

B. Video-based emotion recognition

The video-based emotion recognition module is based on an inflated bidimensional (2D) convolutional neural network (CNN), similar to I3D [8], the state-of-the-art in activity recognition. The model is an end-to-end network: it receives frames extracted from a video, ordered and concatenated over a time dimension, and returns class probabilities for that video. The architecture of the network follows the structure of a ResNet-50 (see Fig. 2), proposed by He *et al.* [16], whose name stands for residual networks. The shortcut connections that perform identity mapping on each residual learning block enable the stable training of models with more convolutional layers, resulting in deeper representations of the input data.

The inflated ResNet-50 consists of a bidimensional ResNet-50 model where the convolutional filters and layers have been converted into 3D. This allows them to process several frames simultaneously as a single input. Downsampling operation before the first block of each type enables learning multi-resolution features. This model has been pretrained[1] to discriminate between 339 activity classes on the Multi-Moments In Time database [17]. To offer probability outputs for each of the three group-level emotion valence classes, the last fully-connected layer of this network is replaced by a three-neuron fully-connected layer, followed by softmax activation, trained on the EmotiW 2020 sub-challenge train dataset.

C. Audio-based emotion recognition

The audio-based recognition module (see Fig. 1) is composed of two main processes: feature extraction on sliding windows, and a Bi-LSTM recognition model. Audio features were extracted using pyAudioAnalysis[2] which contains an off-the-shelf feature set with 34 available features, including signal zero-crossing rate, signal energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, mel-frequency cepstral coefficients (MFCC), chroma vector, and chroma deviation. All these features are extracted over sliding windows of 25 milliseconds with a time step of 10 milliseconds.

The features are received by a Bi-LSTM model with local attention that returns the class probabilities for the respective audio (see Fig. 3), adapted from [18]. Its weighted-pooling strategy enables the focus on the specific sound parts which contain strong emotional characteristics, controlled by an attention function trained simultaneously with the Bi-LSTM model.

D. Score-level ensemble

The softmax scores obtained by both video and audio based emotion recognition models are then concatenated in a feature vector, composed of six class probability values, and given to a multi-class SVM that combines the separate audio and video predictions into a single decision.

[1] Available at: https://github.com/zhoubolei/moments_models
[2] Available at: https://github.com/tyiannak/pyAudioAnalysis

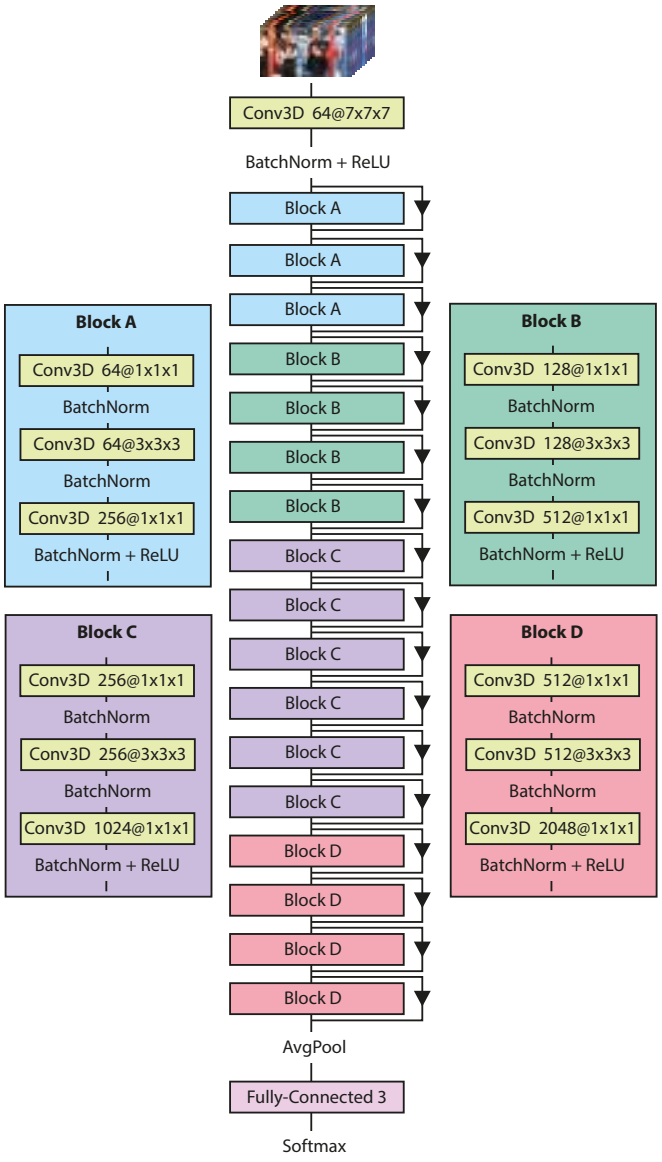

Fig. 2. Structure of the video-based group emotion recognition module, based on an inflated ResNet-50.

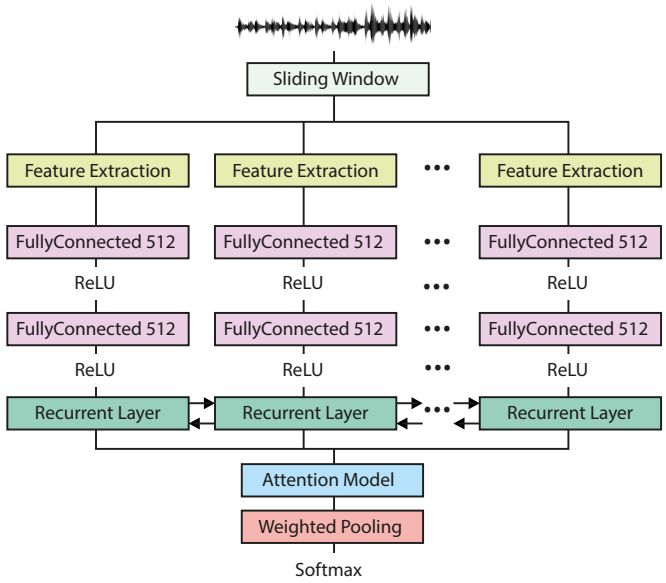

Fig. 3. Structure of the audio-based group emotion recognition module, based on a Bi-LSTM network [18].

III. Experimental settings

A. Data

All the experiments were conducted on an adapted version of the "Video-level Group AFfect" (VGAF) dataset [19] for the EmotiW 2020 AV Group-level sub-challenge [4]. The VGAF is a video-based database that contains labels for emotion and cohesion. The data was collected from the YouTube platform and consists of videos under the creative commons license (CC0) and present keywords that correspond to the range of emotions and cohesion.

Since the number of individuals per video is variable, and the groups on each video can also present a varying number of persons over time, the videos have been divided so that each video-clip has always the same number of persons per frame. Each VGAF clip was manually labeled by different annotators for emotion and cohesion and every annotator was informed of the basic concepts of emotion and cohesion. Only videos with mutual consensus were kept in the final database. The labels for group emotion are related to emotion valence (*i.e.*, positive, neutral, and negative) whereas the group cohesion labels are in the range $[0-3]$, being 0 the state of very low cohesion (dominance over the group members) and 3 the state of very high cohesion.

For the EmotiW 2020 AV Group-Level Emotion sub-challenge, the task was the classification of group emotion. The VGAF dataset videos were divided into five-second videos: 2661 for the train, 766 for the validation, and 756 for the test. Each video (except those in the test set) is accompanied by a discrete group emotion valence ground-truth label. In the training dataset, there are 802 positive videos, 923 neutral videos, and 936 negative videos. In the validation set, there are 302 positive videos, 280 neutral videos, and 184 negative videos.

B. Baseline algorithm

The baseline algorithm is the audiovisual group-level emotion recognition sub-challenge baseline [19] of the EmotiW 2020 Grand Challenge. This method is composed of two streams, for audio and video data processing, fused at the feature-level. The video stream is a pretrained Inception V3 network that separately processes frames extracted from a video. The extracted features are combined using a long short-term memory (LSTM) network. The audio stream is composed of a fully-connected network that receives OpenSMILE [20] features extracted from the audio. The outputs from the video and audio streams are concatenated and used by a fully-connected layer to offer two outputs: the probabilities for the

three emotion valence classes and an emotion cohesion value on the $[0-3]$ range.

C. Preprocessing

The videos on the EmotiW 2020 AV Group-level Emotion sub-challenge were subject to preprocessing before being used by the proposed method. For each five-second video, ten frames were extracted, thus resulting in 2 frames per second. This is an adaptation from the original model pretrained on the Multi-Moments In Time database (MMIT), which worked at 5 frames per second. We found that reducing the frame rate did not harm performance and sped up the recognition process. Before being used on the audio-based recognition module, the audio was extracted from each file by converting them (originally in the MP4 format) to audio files (in the WAV format). Regarding audio, we noticed that after feature extraction some of the generated features could assume non-number values. For training purposes, we removed these samples and trained only with valid ones. For inference during validation/test, we replaced non-number values by zero.

D. Training

The audio-based recognition module was trained from scratch[3]. The weights were randomly initialized, and the model was trained over a maximum of 200 epochs, with a batch size of 128, categorical-cross-entropy as the loss function, and using the Adam optimizer with an initial learning rate of 10^{-2}. To prevent overfitting, we used dropout and early-stopping with a patience value of 15 epochs.

The video-based recognition module, pretrained on the MMIT database, was adapted to output probabilities for each of the three valence classes in group emotion recognition. This was achieved by replacing the last fully-connected layer with a new one, with three neurons.

Since this layer needs to be trained, all weights of the network have been frozen (except those of this layer). The network was briefly fine-tuned until convergence over a maximum of 250 epochs, with batch size 32, using the Adam optimizer with an initial learning rate of 10^{-5}.

When training the audio-based module and the video-based module the hyperparameters have been selected empirically, to maximize performance in the validation set. The hyperparameters for the fusion module (*e.g.*, the regularization parameter "C", the kernel, or the polynomial degree of the kernel) were found through a grid-search. Once the optimal hyperparameters were found, the train and validation set were combined to make a "full train" set, and thus take full advantage of all available labeled data for better performance in the test set.

E. Experiments

In this work, the aim is not only to assess the proposed method's performance for group-level emotion recognition but also to examine its behavior in several conditions.

[3]Code adapted from: https://github.com/RayanWang/Speech_emotion_recognition_BLSTM

TABLE I
ACCURACY (%) OF THE PROPOSED METHOD ON THE VALIDATION SET (A - ONLY AUDIO; V - ONLY VIDEO; A+V - MULTIMODAL).

Method	Accuracy			
	Overall	*Positive*	*Neutral*	*Negative*
Proposed (A)	47.19	19.93	**69.64**	57.61
Proposed (V)	**62.40**	**69.20**	52.50	66.30
Proposed (A+V)	61.83	58.13	61.78	**67.93**
Baseline (A)	50.23	-	-	-
Baseline (V)	52.09	-	-	-
Baseline (A+V)	50.23	-	-	-

We use the accuracy metric and confusion matrices to examine the overall performance of the method and also analyze its class-wise accuracy. The performances of the multimodal method and its audio and video-based modules, separately, are evaluated in both the validation set (with available ground-truth labels) and the test set (accuracy values delivered by the EmotiW 2020 sub-challenge organization upon request). The performance is compared with the official sub-challenge baseline, following the results reported in the paper [19].

The performance is also evaluated according to the number of people in the video. Since the number of people in each video is not included, we use the MTCNN method [21] on each frame of each video, and infer the group size based on the average number of detected faces: a group with less than five detected faces are considered small (total of 502 videos on the validation set), otherwise, it is considered a large group (264 videos on the validation set). With this, we aim to evaluate the difficulties associated with recognizing emotion in large groups, where cohesion is likely to be generally lower.

IV. RESULTS AND DISCUSSION

The performance results of the proposed method, and the comparison with the official sub-challenge baseline, on the validation and test sets, is presented, respectively, in Table I and Table II.

On the validation set, the performance offered by the proposed multimodal method is superior to the baseline. The accuracy attained by the video-only approach is close to that offered by the multimodal method, over 62%. This is evidence of the advantages of using pretrained networks (in this case, transferred from the task of human activity recognition). The audio-only approach offers considerably lower performance (47%) than the audio-only baseline (50%), which indicates the use of OpenSMILE features and fully-connected networks may be better fitted for group emotion recognition based on audio.

From the validation to the test set (Table II), the official video-only baseline suffers a sharp performance decay (from 52% to 42% accuracy), which is also felt with the proposed video-only approach (albeit not as dramatic, from 62% to 59% accuracy). Fusing with audio on a multimodal approach reduced that decrease in the case of the official baseline (from 50% to 48% accuracy), and even reversed it in the case of the proposed method (from 62% to 66% accuracy). This confirms

TABLE II
ACCURACY (%) OF THE PROPOSED METHOD ON THE TEST SET (A - ONLY AUDIO; V - ONLY VIDEO; A+V - MULTIMODAL).

Method	Accuracy			
	Overall	*Positive*	*Neutral*	*Negative*
Proposed (V)	58.86	**55.76**	57.93	63.04
Proposed (A+V)	**65.74**	54.38	**77.99**	60.00
Baseline (V)	42.00	-	-	-
Baseline (A+V)	47.88	45.00	10.00	**70.00**

TABLE III
CONFUSION MATRIX OF AUDIO-BASED RECOGNITION (ON THE VALIDATION SET).

		Predicted Class		
		Positive	*Neutral*	*Negative*
True Class	***Positive***	60	139	102
	Neutral	35	195	50
	Negative	24	54	106

TABLE IV
CONFUSION MATRIX OF VIDEO-BASED RECOGNITION (ON THE VALIDATION SET).

		Predicted Class		
		Positive	*Neutral*	*Negative*
True Class	***Positive***	209	59	34
	Neutral	98	147	35
	Negative	44	18	122

TABLE V
CONFUSION MATRIX OF MULTIMODAL RECOGNITION (ON THE VALIDATION SET).

		Predicted Class		
		Positive	*Neutral*	*Negative*
True Class	***Positive***	175	99	27
	Neutral	78	173	29
	Negative	27	32	125

the idea present in the literature that, while audio alone is not suitable for recognition, it offers additional information that is essential for the robustness and accuracy of the method.

Analysing the class-wise accuracies and the confusion matrices (Table III, Table IV, and Table V), one can notice that video is, overall, the best modality to recognise emotions. The advantages of using video rely mainly on the "extreme" classes, positive and negative, which denote visual information is more advantageous to recognize strong group emotions. The proposed audio-only approach attains very poor accuracy in the positive class.

Since the positive class is the minority class in the training dataset, the results of the audio-only approach may partially be explained by this slight class imbalance. However, as mentioned before, the video-only approach does not verify this, which is fortunate when combining both approaches into the multimodal proposed method. Using both modalities slightly decreases the accuracy of positive and negative videos, when compared with the video-only approach, but takes advantage of the audio information to considerably improve accuracy on neutral videos and achieve overall better performance.

TABLE VI
ACCURACY (%) ON THE VALIDATION SET FOR VIDEOS OF SMALL GROUPS *vs.* LARGE GROUPS (A - ONLY AUDIO; V - ONLY VIDEO; A+V - MULTIMODAL).

Method	Group Size	
	$N < 5$	$N \geq 5$
Proposed (A)	48.61	44.49
Proposed (V)	66.33	54.92
Proposed (A+V)	65.74	54.37

At last, the results of the group size study are presented in Table VI. In both audio and video-only approaches, as well as the multimodal method, the recognition performance is higher in smaller groups. The performance values should serve as a rough reference, since the process of face detection may present errors, and the number of faces may not accurately describe the number of people in the video's group (which may include occluded faces or people facing the opposite direction of the camera).

Nevertheless, the performance differences are considerable and show expected behavior: it should be harder for larger groups to consistently show the same emotion than smaller groups. Hence, emotion cohesion should be higher, on average, for smaller groups, and thus the certainty of the algorithms when recognizing the emotion valence. This could perhaps be addressed using hierarchical methodologies (from individual-level to group-level) as used in current group activity recognition approaches (discussed in the related work section).

Through an analysis of some videos where the proposed model failed, a pattern emerged. While there are certain videos where the error was evident, there are several examples where it is very difficult to notice that an apparently neutral scene displays, in fact, a positive or negative group emotion.

Some examples are shown in Fig. 4. On the top left, a short video of a calm conversation on a TV show that is labeled as positive. On the top right, a negative emotion video that the proposed method classified as positive, since the video only covers the moment before the boy being bullied started crying.

Fig. 4. Some examples of validation set videos where the model offered unsuccessful predictions (top left: label positive, predicted neutral; top right: label negative, predicted positive; bottom left: label positive, predicted neutral; bottom right: label negative, predicted neutral).

On the bottom left, a conference presentation that the method classified as neutral since the positive ground-truth emotion could only be verified by the facial expressions. On the bottom right, a conversation deemed neutral by the proposed method, where only a closer inspection of the audio shows that it is, in fact, part of a protest or a similar confrontation.

Although the information about the underlying ground-truth labels is indeed present on these example videos, it is hidden in contextual clues, expressions in small faces, or the content of conversations. Exploring ways to integrate these aspects into the recognition of group-level emotion could be the way to avoid the most common mistakes of the proposed method, and ultimately achieve better overall performance and robustness.

V. Conclusion

In this work we proposed a novel method for the automatic recognition of group emotion that uses a late-fusion multimodal approach, combining scores from both video and audio based emotion recognition models that are used to feed a multiclass SVM that returns a final class probability. This method showed significant improvement against the baseline, confirming that the use of acquired knowledge from activity recognition is useful for group-emotion recognition and that the joint utilization of audio and video benefits the learning of the model.

On the other hand, taking into account the maximum accuracy value, we believe that there is still room for further improvements. Further efforts should be devoted to the study of the links between the tasks of activity recognition and emotion recognition, especially on the group-level. Approaches for abnormal behavior recognition through video anomaly detection, such as [22], could offer meaningful improvements on the automatic distinction between positive and negative emotions. Also, this method could benefit from OpenSMILE features in the audio-based emotion recognition module, on the development of multimodal approaches that are based on early-fusion (*e.g.*, input or intermediate-layer levels), and on the design of a "fully" end-to-end network that receives both video and audio as input and learns the relevant features for the classification task (*e.g.*, through regularization methods such as loss functions with different terms and weights).

Acknowledgments

The authors wish to acknowledge the EmotiW 2020 Grand Challenge [4] organizers and the authors of the MMIT dataset and pretrained models used in this work.

References

[1] P. M. Ferreira, F. Marques, J. S. Cardoso, and A. Rebelo, "Physiological inspired deep neural networks for emotion recognition," *IEEE Access*, vol. 6, pp. 53 930–53 943, 2018.

[2] D. Mehta, M. F. H. Siddiqui, and A. Y. Javaid, "Facial Emotion Recognition: A Survey and Real-World User Experiences in Mixed Reality," *Sensors*, vol. 18, no. 2, 2018.

[3] L. Tan, K. Zhang, K. Wang, X. Zeng, X. Peng, and Y. Qiao, "Group Emotion Recognition with Individual Facial Emotion CNNs and Global Image Based CNNs," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI '17, 2017, p. 549–552.

[4] A. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "EmotiW 2020: Driver Gaze, Group Emotion, Student Engagement and Physiological Signal based Challenges," in *ICMI '20: 22nd ACM International Conference on Multimodal Interaction*. New York, NY, USA: ACM, 2020.

[5] B. Sun, Q. Wei, L. Li, Q. Xu, J. He, and L. Yu, "Lstm for dynamic emotion and group emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI '16, 2016, p. 451–457.

[6] Q. Wei, Y. Zhao, Q. Xu, L. Li, J. He, L. Yu, and B. Sun, "A new deep-learning framework for group emotion recognition," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI '17, 2017, p. 587–592.

[7] A. Gupta, D. Agrawal, H. Chauhan, J. Dolz, and M. Pedersoli, "An attention model for group-level emotion recognition," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ser. ICMI '18, 2018, p. 611–615.

[8] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4724–4733.

[9] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 961–970.

[10] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6201–6210.

[11] M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, and L. Van Gool, "stagNet: An Attentive Semantic RNN for Group Activity and Individual Action Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 549–565, 2020.

[12] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5491–5500.

[13] R. Cosbey, A. Wusterbarth, and B. Hutchinson, "Deep Learning for Classroom Activity Detection from Audio," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3727–3731.

[14] D. Liang and E. Thomaz, "Audio-based activities of daily living (ADL) recognition with large-scale acoustic embeddings from online videos," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 1, Mar. 2019.

[15] W. Wang, F. Seraj, N. Meratnia, and P. J. M. Havinga, "Privacy-Aware Environmental Sound Classification for Indoor Human Activity Recognition," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA '19, 2019, p. 36–44.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[17] M. Monfort, K. Ramakrishnan, A. Andonian, B. A. McNamara, A. Lascelles, B. Pan, Q. Fan, D. Gutfreund, R. Feris, and A. Oliva, "Multi-moments in time: Learning and interpreting models for multi-action video understanding," 2019.

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

[19] G. Sharma, S. Ghosh, and A. Dhall, "Automatic group level affect and cohesion prediction in videos," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2019, pp. 161–167.

[20] F. Eyben, "Real-time speech and music classification by large audio feature space extraction," *Springer Theses*, vol. 10, 2016. [Online]. Available: https://www.springer.com/de/book/9783319272986

[21] 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, 2016.

[22] P. Augusto, J. S. Cardoso, and J. Fonseca, "Automotive Interior Sensing - Towards a Synergetic Approach between Anomaly Detection and Action Recognition Strategies," in *Proceedings of the Fourth IEEE International Conference on Image Processing Systems and Applications (IPAS)*, Dec. 2020.