EmoP3D: A Brain like Pyramidal Deep Neural Network for Emotion Recognition

Emanuel Di Nardo[1][0000−0002−6589−9323], Alfredo Petrosino[1][0000−0002−8736−1997], and Ihsan Ullah[2][0000−0002−7964−5199]

[1] CVPR Lab, University of Naples Parthenope, Napoli, Italy
{emanuel.dinardo, alfredo.petrosino}@uniparthenope.it
[2] Data Mining & Machine Learning Group, Discipline of Information Technology, National University of Ireland Galway, Galway, Ireland
ihsan.ullah@nuigalway.ie

Abstract. The paper reports a new model based on the understanding and encompassing intelligence from brain i.e. biological pyramidal neurons, tailored for emotion recognition. Our objective is to introduce and utilize usage of non-Convolutional layers in models and show comparable or state-of-the-art performance for multi-class emotion recognition problem. We open-sourced the optimized code for researchers. Our model shows state-of-the-art performance on two emotion recognition datasets (eNTERFACE and Youtube) enhancing previous best result by 9.47% and 20.8%, respectively.

Keywords: Emotion Recognition · Pyramidal Neural Network · 3DPyraNet · Convolutional Neural Network

1 Introduction

Despite wide applicability and success of deep neural networks, understanding the human visual system for emotion recognition (ER) needs more study about the representation and processing of visual information in the brain. Indeed, human emotion recognition from sensors such as speech, heartbeat, breathing, muscle tension, etc. are not as successful in real-time as of visual sensors.Emotion recognition shows the mood, personality, motivation, and intentions of a person. For example, it can help in maintaining law and order situation in a crowded scenario by identifying the force behind the positive or negative intentions of a person in the scene which can be controlled before getting out of control. Humans can easily read and understand the facial expression of a person. However, it is not the same for machines. Recognition of human emotions (e.g. happy, sad, tensed) from speech can be wrongly classified due to the noise in the scene. Whereas, one can not have implanted sensors over the human body all the time. Therefore, human facial expression recognition in the videos is a highly researched area of computer vision *CV* and machine learning *ML*. Psychology divide human emotions in six i.e. anger, disgust, fear, happy, sad, and surprise.

Our model utilized the visual features of humans to classify among these six categories.

The applications of emotion recognition are not limited to surveillance. It is widely used for video summarization, e-commerce, normalizing facial images by removing emotions, and helping people with neurological disorder [26]. For example, research has shown that people affected by multiple sclerosis disease can face difficulty in understanding the features or expression that help in understanding the emotions of others [3].

The traditional *CV* approach is to use spatio–temporal features extracted with handcrafted descriptors and then classifying them with a state-of-the-art classifier. *ML* approaches try to automatically learn features from the training samples and give accurate predictions based on the trained model. In this paper, we will enhance a recently introduced deep neural network called 3D pyramidal neural network (3DPyraNet) [23, 25], already proposed by the authors. The model is based on the motivation from pyramidal neurons in the brain. It uses a pyramidal structure and introduces a new layer called correlation layer as well as a new weighting scheme that helps in better learning and reduction of the number of parameters as compared to other state-of-the-art deep models. In this paper, we enhance the model introducing per frame normalization in the normalization layer as well as increased size of the model. The model is evaluated on two challenging datasets for emotion recognition, i.e eNTERFACE, created in a laboratory mimicking real-world emotions, and Youtube, collected from real-world YouTube videos.

The paper is organized as follows: Section 2 provides an overview of the work being done until now. Section 3 provides a motivational background behind the proposed model. Further, its sub-sections explain the details about existing techniques that are modified, combined and enhanced for our proposed models. Datasets and data preparation are discussed in section 4. Section 5 provides details about the used benchmark datasets and achieved results. Finally, Section 6 concludes this paper.

2 Background

ER from videos is a challenging task. Prior approaches [17, 12] use to define face landmarks in order to extract important features that are specific to face elements and their position. Later, other CV approaches such as [1] integrate time motion image and the quantized image matrix to extract visual features. In [28], a kernel-based technique identifies optimal transformations that represent the coupled patterns between features from multiple modalities. One example is [16], that was developed to work on audio-visual features. It performs a multi-class classification. However, it expects that only one emotion at a time can be recognized. Visual input is not fully raw information hence relevant key-frames that are the most representative frames in an image sequence with emotions are extracted using a clustering-based strategy. Out of these, geometric features, metric distances, angles, and others are learned by an Inception network.

MARN model [30] tries to discover relations between multiple modalities using a Recurrent Neural Network with a hybrid long-short-term memory architecture (LSTHM). Each input is processed by the proposed LSTHM. The output of each network is concatenated and used as input for a multi-attention block. All features are concatenated and a deep neural network is used with K softmax layers to output coefficients for each modality. Chen et al. [4] proposed a model based on reinforcement learning that utilizes word-level fusion and uses LSTM with Temporal Attention to predict human emotion. One of the most efficient technique is [26] that also works on multi-modality. Visual features are extracted using 2D and 3D convolutional neural networks. 2D model uses VGG-16 architecture [18] and 3D features are extracted with C3D [5]. In both cases, a feature vector of 4096 dimensions is extracted. Each model performs a temporal fusion using scores produced by softmax layer and low dimensional vectors of 297 and 394 features are extracted. All modalities are combined to get the resulting feature vector. Definitely, temporal fusion is done using a length variable LSTM in order to take all the descriptor as inputs obtaining a VGG-LSTM and a C3D-LSTM network.

The most recent approach is proposed in [8] where many architectures are adopted at one time. It uses HoloNet [29], pre-trained DenseNet [9] and ResNet [7]. At each epoch, the models are fine-tuned by many supervised support networks that help to generalize the emotional states. This deep and mixed model takes advantage of multi-modal handcrafted features and shows promising results.

Pyramid structure has a deep role in improving the performance of neural network models. Both Neural Network & Image Pyramids have similar structure i.e. exponential reduction and coarse to fine refinement. Pyramidal ImageNet (gradual decreasing) Vs. the non-Pyramidal ImageNet model showed equal or better performance despite fewer parameters [23]. PyramidNet enhanced ResNet by similar concept as [23] but in reverse order (gradually increasing) [6]. Literature review shows that different models uses different datasets which make it hard to benchmark one model with others. Our model will be evaluated on a dataset called eNTERFACE, that is used by many researchers, and on ICT Youtube. In next section, the enhanced 3DPyraNet model is explained.

3 Proposed Model

The proposed model (EmoP3D) is an enhancement of 3D pyramidal neural network (3DPyraNet) proposed in [23]. The proposed 3D pyramidal architecture was based on the concept of coarse to fine refinement or the decision making in a pyramidal structure inside the brain. More specifically, similar to the pyramidal neurons in a biological brain [22]. Pyramidal neurons exist in cortical layers. They communicate with each other as well as with sub-cortical regions of the brain through their long axonal projections forming a pyramid structure before taking a decision and transferring to higher layers. The cortical layers in the mammalian cerebral cortex perform a major role in important cognitive func-

tions, perception and motor control. In [22], the role of pyramidal neurons at layer 2, 3 and 5 is compared and it is shown how it affects the motor skill learning of mice. Further, it is in human nature to work and decide based on the course to fine refinement (pyramid structure) e.g. shopping in a market. Similarly, traditional feature subset selection technique has the same concept of selecting the most relevant features out of a large set of ambiguous and redundant features.

3DPyraNet was further inspired by pyramidal structures in [2], weighting scheme in [19], and 3D spatio-temporal structure from 3D Convolutional Neural Network (3DCNN) [10]. It works on utilizing spatio-temporal features in a video as input. 3DCNN and CNN do reduce spatial resolution, however, increase the number of maps in each higher layer which avoids the strict pyramid structure of the network.

On the contrary, [24] adopted a strict pyramidal architecture. This topology tends to reduce the dimensions of the input layer by layer, in order to obtain fine non-ambiguous high-quality features. It uses *Weighted Sum* (WS), or more commonly the cross-correlation product for each receptive field. It is the basic network operation in this model instead of convolution. The most fundamental property is that the number of weights in a layer is equal to the input image or feature map size. The sliding kernel and then WS results in a partial weight sharing scheme. This means that each neuron has a unique local weight that is updated at each iteration inside the receptive field of specific output neurons. This reduces the burden on each weight parameter and enhances the performance of the network. The pyramidal structure and weighting scheme has a "sparse-to-influence" effect. It helps to remove irrelevant features and preserves the most discriminative and relevant features during the training process. Therefore, it can learn complex structure even with fewer feature maps and hidden layers.

The strict structure preserves the number of feature maps for each layer. The weight matrix and input mask have equal size. After each weighted sum operation, it is reduced as it goes to a higher layer. This helps in reduction of features as well as in maintaining and reducing the number of learn-able parameters as compared to recent deep networks with a large number of parameters. In order to maintain pyramid structure and extract more features, three 3d weight matrices are incorporated to extract more temporal information at the first hidden layer, as shown in Figure. 1. This approach fuses the local weights connectivity with weight sharing because 3DWS layers preserve the local connectivity on 2D space, but working on small temporal region weights are shared on depth, repeating the same weights volume on each time-depth step. Temporal extension extracts correlation among objects and actions in adjacent frames. The output in a layer l of a neuron $N^l_{u,v,z}$ for a receptive field $R^l_{u,v,z}$ using 3D weighted sum operation become:

$$y^l_{u,v,z} = f_l\Big(\sum_{i \in R^l_{u,v,z}} \sum_{j \in R^l_{u,v,z}} \sum_{k \in R^l_{u,v,z}} [W^l_{i,j,k} \circ x^{l-1}_{i,j,k}] + b^l_{u,v,z}\Big) \tag{1}$$

As it is possible to see the bias term b^l has the same shape as the output neuron map, it means that each neuron has an own bias term to be learned. It is different

Dataset	Input-size	WS1	L3P	WS5	FC
Both	16x100x100x1	14x97x97x1x3	12x48x48x1x3	10x45x45x1x3	60750

Table 1. Network structure per layer for both datasets (eNTERFACE and YouTube)

from the standard bias concept of other network models where it has the size of the output feature maps. The adopted 3D temporal pooling layer is different from any other kind of pooling layer because it not only reduces spatial and temporal dimensions but also performs a data transformation on previous weighted sum layer to avoid scale and translation problems. Mathematically it is defined as:

$$y^l_{u,v,z} = f_l(W^l_{u,v,z} \times \max_{i,j,k \in R^l_{u,v,z}} y^{l-1}_{i,j,k} + b^l_{u,v,z}) \tag{2}$$

3.1 Model structure

A general description about the key layers i.e. WS layer and a max-pooling layer of the proposed model is given in previous section 3. Further, it consists of the activation layer, a normalization layer, and fully connected layer. The model is composed of two 3DWS layers linked together by a 3D max pooling layer. In normalization layer, each frame is normalized individually with zero mean unit variance. Each 3DWS layer is followed by an activation layer which is followed by the normalization layer. Similarly, each 3D-Max-Pooling layer is followed by activation and normalization layers. To preserve the strict property, each correlation layer uses three feature maps. To ensure a fast convergence and avoid gradients problems a *Leaky ReLU* activation function [11] is used. Complete structure can be seen in Fig. 1. Finally an output layer, a softmax classifier is used for class probability estimation. Each 3DWS layer reduces the number of feature maps by two as well as reduces the spatial size of feature maps based on the size of receptive field and stride used at that layer. Similarly, in pooling layer, not only pooling reduces the spatial resolution but the 3D structure reduces the feature maps by two. Hence it maintains a continuous constant reduction in both spatial and temporal dimensions forming a pyramid structure that helps in providing most discriminative features at the final layer. The model takes an input clip of size $16 \times 100 \times 100 \times 1$ (depth, height, width, channels). We have taken this to make it similar to the work done in [31]. However, the network is not restrictive to the input size and this can be changed according to the application. This change of input size changes the structure of the network resulting in a different number of parameters to train.

A total of three weight matrices are used with a receptive field of 3x4x4 for depth, height and width dimensions in weighted sum layers. It is also adopted a valid max pooling structure with a receptive field of 2 with a step of 2 in order to halves the spatial dimensions and reduce the temporal domain in a progressive, soft manner. Back-propagation algorithm is used to train the network. Mini-batch gradient descends with a batch size of 100 samples for each dataset is

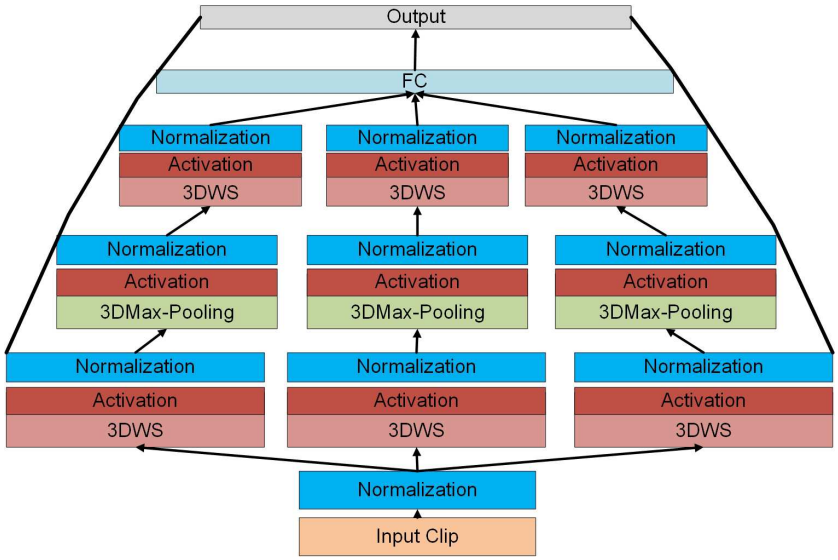

Fig. 1. 3DPyraNet Architecture

used with Momentum optimization algorithm [21]. Table. 1 shows the size of each layer of the network. It shows that the network reduces its size gradually and resulting in an FC layer of size 60750.

4 Dataset and Data Preparation

There are many datasets for ER, but they are for the most compatible only with 2D methods. Since we propose a method that is able to capture temporal features across multiple frames, it is mandatory to use a dataset that allow to model this behavior. Two datasets in the literature that are mostly used are eNTERFACE '05 [13] and Youtube [14]. eNTERFACE '05 dataset adopts 42 subjects from 14 different countries. Each subject reacted to a story told to catch real emotions from them. They express 6 different emotions and, for each of them, there are five reactions. There is a total of 1166 video sequence. Youtube dataset starts with the fact that people express emotions in different ways. It contains videos of people from different age and gender in various topics. Furthermore, these are not recorded in a controlled environment because an emotion recognition system should be able to work in many contexts. There are 20 female subjects and 27 male subjects with age in the range 14-60. It provides three polarity classes i.e anger, worried, and happiness. Each video sequence is pre-processed to detect and extract the subject's face. It is done using a "Multitask CNN" proposed in [31] and available in [20]. Some sample facial images are shown in Figure. 2. In this paper, each image is gray-scaled and re-sized to have 100×100 height and width dimensions. A sequence of 16 consecutive frames with an overlap of

Fig. 2. Sample Images from eNTERFACE (First row) and YouTube (Second Row)

8 frames is used as input clip. In the Youtube dataset, each clip is a mix of emotions so the most prevalent one is chosen if there is an emotion overlap in a sequence.

5 Results and Discussion

Using the proposed architecture, we adopted a 10-Fold cross-validation approach to evaluate the performance on each dataset. Comparing our model with state-of-the-art methods, it is possible to see in table 2 that our model outperforms previous techniques on both datasets. The most important comparison is done with [16, 4, 30] that is the most recent end-to-end approach proposed for emotion recognition on the same dataset. Our model outperforms the previous best result provided by a CNN based model (AVER-CNN) by more than 9% enhancement in the accuracy over eNTERFACE emotion recognition dataset. Whereas, EmoP3D outperformed previous models i.e. LSTM(A) and MARN by more than 20% increase in accuracy. A confusion matrix over a single run is also shown to inspect per class influence. We can observe from the left matrix of Fig. 3 that overall performance is good over the whole dataset, only two classes that are *Sadness* and *Surprise* shows to be more difficult to understand by the network. Both of them are even hard for a normal human being to recognize. From Fig. 3 right matrix, the most confusing class is the neutral sentiment, it is clear that "neutral" isn't a true defined and distinguished sentiment and it is also difficult for human understanding to identify it. We also optimized the code and open-sourced the cuda based code of EmoP3D implementation for researchers. The

Approach	eNTERFACE	Youtube
AVER-Geometric[15]	41.59%	-
KCMFA[27]	58%	-
AVER-CNN[15]	62%	-
LSTM(A) (Binary)[4]	-	52.3%
MARN [30]	-	54.2%
EmoP3D (Ours)	**71.47%**	**75%**

Table 2. Comparison in-terms of accuracy with other state-of-the-art models

	Sample size	Time (s) for a batch of 100 samples	Time (s) for a single input	Speedup
3DPyraNet	13x80x100	105	1.05	-
Ours EmoP3D	16x100x100	1.32	0.0132	**∼79%**

Table 3. Processing time comparison of 3DPyraNet implementation vs. EmoP3D

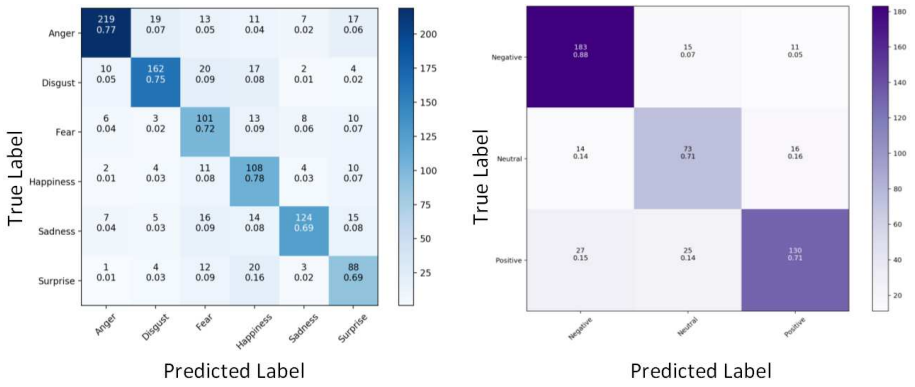

Fig. 3. Confusion matrices for eNTERFACE (left) and YouTube (right) result

code is available on Github[3]. Table. 3 shows comparison of the time taken by new code vs old code. It gave us more than 79% speedup.

6 Conclusion

Our biological pyramidal neurons inspired architecture is able to outperform other models used for ER. It demonstrates that the temporal domain has an important role in the expression of emotion. It is a process that involves many temporal stages that need to be well defined. This work opens the way to many scenarios that can help to identify emotions deeply. For example, it is possible

[3] https://github.com/CVPRLab-UniParthenope/EmoP3D/projects

to extend this work to multiple domains using also information provided by voice or where possible from what a person is telling. This information can be used together in a multimodal approach to enhance the ability of a machine to identify, understand and maybe reproduce human emotions. The proposed architecture shows its generalization power by showing good results in both control (eNTERFACE) and uncontrolled (YouTube) environment datasets. In the future, we intend to further increase the network depth and use ResNet structure with 3DWS layers.

References

1. Bejani, M., Gharavian, D., Charkari, N.M.: Audiovisual emotion recognition using anova feature selection method and multi-classifier neural networks. Neural Computing and Applications **24**(2), 399–412 (2014)
2. Cantoni, V., Petrosino, A.: Neural recognition in a pyramidal structure. IEEE Transactions on Neural Networks **13**(2), 472–480 (Mar 2002). https://doi.org/10.1109/72.991433
3. Cecchetto, C., Aiello, M., DAmico, D., Cutuli, D., Cargnelutti, D., Eleopra, R., Rumiati, R.I.: Facial and bodily emotion recognition in multiple sclerosis: The role of alexithymia and other characteristics of the disease. Journal of the International Neuropsychological Society **20**(10), 10041014 (2014). https://doi.org/10.1017/S1355617714000939
4. Chen, M., Wang, S., Liang, P.P., Baltrušaitis, T., Zadeh, A., Morency, L.P.: Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. pp. 163–171. ACM (2017)
5. Fan, Y., Lu, X., Li, D., Liu, Y.: Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 445–450. ACM (2016)
6. Han, D., Kim, J., Kim, J.: Deep pyramidal residual networks. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. vol. 2017-Janua, pp. 6307–6315 (2017). https://doi.org/10.1109/CVPR.2017.668
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
8. Hu, P., Cai, D., Wang, S., Yao, A., Chen, Y.: Learning supervised scoring ensemble for emotion recognition in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. pp. 553–560. ACM (2017)
9. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. CoRR **abs/1608.06993** (2016), http://arxiv.org/abs/1608.06993
10. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 221–231 (Jan 2013). https://doi.org/10.1109/TPAMI.2012.59
11. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml. vol. 30, p. 3 (2013)
12. Mansoorizadeh, M., Charkari, N.M.: Multimodal information fusion application to human emotion recognition from face and speech. Multimedia Tools and Applications **49**(2), 277–297 (2010)

13. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The enterface05 audio-visual emotion database. In: Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on. pp. 8–8. IEEE (2006)
14. Morency, L.P., Mihalcea, R., Doshi, P.: Towards multimodal sentiment analysis: Harvesting opinions from the web. In: Proceedings of the 13th international conference on multimodal interfaces. pp. 169–176. ACM (2011)
15. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing pp. 1–1 (2017). https://doi.org/10.1109/TAFFC.2017.2713783
16. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing (2017)
17. Paleari, M., Huet, B.: Toward emotion indexing of multimedia excerpts. In: Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop on. pp. 425–432. IEEE (2008)
18. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC. vol. 1, p. 6 (2015)
19. Phung, S.L., Bouzerdoum, A.: A pyramidal neural network for visual pattern recognition. IEEE Transactions on Neural Networks **18**(2), 329–343 (March 2007). https://doi.org/10.1109/TNN.2006.884677
20. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
21. Sutton, R.S.: Two problems with backpropagation and other steepest-descent learning procedures for networks. In: Proceedings of Eightth Annual Conference of the Cognitive Science Society, 1986 (1986)
22. Tjia, M., Yu, X., Jammu, L.S., Lu, J., Zuo, Y.: Pyramidal neurons in different cortical layers exhibit distinct dynamics and plasticity of apical dendritic spines. Frontiers in neural circuits **11**, 43 (2017)
23. Ullah, I., Petrosino, A.: About pyramid structure in convolutional neural networks. Proceedings of the International Joint Conference on Neural Networks **2016-Octob**, 1318–1324 (2016). https://doi.org/10.1109/IJCNN.2016.7727350
24. Ullah, I., Petrosino, A.: Spatiotemporal features learning with 3dpyranet. In: Blanc-Talon, J., Distante, C., Philips, W., Popescu, D., Scheunders, P. (eds.) Advanced Concepts for Intelligent Vision Systems. pp. 638–647. Springer International Publishing, Cham (2016)
25. Ullah, I., Petrosino, A.: A spatio-temporal feature learning approach for dynamic scene recognition. In: Shankar, B.U., Ghosh, K., Mandal, D.P., Ray, S.S., Zhang, D., Pal, S.K. (eds.) Pattern Recognition and Machine Intelligence. pp. 591–598. Springer International Publishing, Cham (2017)
26. Vielzeuf, V., Pateux, S., Jurie, F.: Temporal multimodal fusion for video emotion classification in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. pp. 569–576. ACM (2017)
27. Wang, Y., Guan, L., Venetsanopoulos, A.N.: Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. IEEE Transactions on Multimedia **14**(3), 597–607 (June 2012). https://doi.org/10.1109/TMM.2012.2189550
28. Wang, Y., Guan, L., Venetsanopoulos, A.N.: Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. IEEE Transactions on Multimedia **14**(3), 597–607 (2012)

29. Yao, A., Cai, D., Hu, P., Wang, S., Sha, L., Chen, Y.: Holonet: Towards robust emotion recognition in the wild. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 472–478. ICMI 2016, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2993148.2997639, http://doi.acm.org/10.1145/2993148.2997639
30. Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E., Morency, L.P.: Multi-attention recurrent network for human communication comprehension. arXiv preprint arXiv:1802.00923 (2018)
31. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)