An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild

Panagiotis Antoniadis*, Ioannis Pikoulis*, Panagiotis P. Filntisis, Petros Maragos

School of ECE, National Technical University of Athens, 15773 Athens, Greece

{pantoniadis97, pikoulis.giannis}@gmail.com, filby@central.ntua.gr, maragos@cs.ntua.gr

Abstract

In this work we tackle the task of video-based audio-visual emotion recognition, within the premises of the 2nd Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW2). Poor illumination conditions, head/body orientation and low image resolution constitute factors that can potentially hinder performance in case of methodologies that solely rely on the extraction and analysis of facial features. In order to alleviate this problem, we leverage both bodily and contextual features, as part of a broader emotion recognition framework. We choose to use a standard CNN-RNN cascade as the backbone of our proposed model for sequence-to-sequence (seq2seq) learning. Apart from learning through the RGB input modality, we construct an aural stream which operates on sequences of extracted mel-spectrograms. Our extensive experiments on the challenging and newly assembled Aff-Wild2 dataset verify the validity of our intuitive multi-stream and multi-modal approach towards emotion recognition "in-the-wild". Emphasis is being laid on the the beneficial influence of the human body and scene context, as aspects of the emotion recognition process that have been left relatively unexplored up to this point. All the code was implemented using PyTorch[1] and is publicly available[2].

1. Introduction

Automatic affect recognition constitutes a subject of rigorous studies across several scientific disciplines and bears immense practical importance as it has extensive applications in environments that involve human-robot cooperation, sociable robotics, medical treatment, psychiatric patient surveillance and many other human-computer interaction scenarios.

---

*Equal contribution

[1] https://pytorch.org/

[2] https://github.com/PanosAntoniadis/NTUA-ABAW2021

Representing human emotions has been a basic topic of research in psychology. While the cultural and ethnic background of a person can affect their expressive style, Ekman indicated that humans perceive certain basic emotions in the same way regardless of their culture [10, 9]. These six universal facial expressions (happiness, sadness, surprise, fear, disgust and anger) constitute the categorical model. Contempt was subsequently added as one of the basic emotions [29]. Due to its direct and intuitive definition of facial expressions, the categorical model is used in the majority of emotion recognition algorithms [16, 12, 42, 1] and large-scale databases (MMI [34], AFEW [8], FER-Wild [32], etc). However, the subjectivity and ambiguity of restricting human emotion to discrete categories result in large intra-class variations and small inter-class differences.

Recently, the dimensional model proposed by Russell [37] has gained a lot of attention where emotion is described using a set of two latent dimensions that are valence (how pleasant or unpleasant a feeling is) and arousal (how likely is the person to take action under the emotional state). Another dimension called dominance is used sometimes to know whether the person is controlling the situation or not. Since a continuous representation can distinguish between subtly different displays of affect and encode small changes in the intensity, some recent algorithms [33, 4, 22] and databases (Aff-Wild [39, 21], Aff-Wild2 [22], OMG-Emotion [2], AFEW-VA [24], etc) have utilized the dimensional model for uncontrolled emotion recognition. Even so, predicting a 2-dimensional continuous value instead of a category increases the task complexity by a lot and lacks intuitiveness.

The remainder of the paper is structured as follows: Firstly, we provide an overview of the latest and most notable related work in the domain of video-based emotion recognition in-the-wild. Subsequently, we analyze our proposed model architecture. Next, we present our experimental results on the Aff-Wild2 dataset, followed by conclusive remarks.

2. Related Work

Emotion recognition has been extensively studied for many years using different representations of human emotion, like basic facial expressions, action units and valence-arousal.

Recently, many studies have tried to leverage all emotion representations and jointly learn these three facial behavior tasks. Kollias et al. [19] proposed the FaceBehaviorNet, the first study that considered joint-learning of all facial behaviour tasks, in a single holistic framework. They utilized many publicly available emotion databases and proposed two strategies for coupling the tasks during training. Later, Kollias et al. released the Aff-Wild2 dataset [22], the first large scale in-the-wild database containing annotations for all three main behavior tasks. They also proposed multitask learning models that employ both visual and audio modalities and suggested using the ArcFace loss [7] for expression recognition. In an additional work, Kollias et al. [23, 20] studied the problem of non-overlapping annotations in multitask learning datasets. They explored task-relatedness and proposed a novel distribution matching approach, in which knowledge exchange is enabled between tasks, via matching of their predictions' distributions. Last year, the First Affective Behavior Analysis in-the-wild (ABAW) Competition [18] was held in conjunction with the IEEE Conference on Face and Gesture Recognition 2020. The competition contributed in advancing the state-of-the-art methods on the dimensional, categorical and facial action unit analysis and recognition on the basis of the Aff-Wild2 dataset.

Although, the aforementioned methodologies boast relatively high recognition scores on the premise of facial expression analysis, they often neglect the usage of other supplementary sources of affective information, such as the *body* and *context*. Related works [25, 30, 11, 36] from the field of context-based visual emotion recognition, follow a more holistic approach towards solving the "in-the-wild" version of the current problem. Kosti et al. [25] introduced the EMotions In Context (EMOTIC) dataset, the first large-scale image database for context-based emotion recognition, annotated on the basis of an extended emotional corpus with 26 discrete categories and VAD dimensions. A baseline model was also provided, consisting of two ConvNet feature extractors (one for each of the *body* and *context* input streams) and one fusion network. Mittal et al. [30] surpassed baseline performance by fusing multiple input modalities, including the face, pose, inter-agent interactions and socio-dynamic context, effectively forming the EmotiCon framework. Furthermore, researchers have extended the concept of context-based visual emotion recognition in the dynamic setting of video sequences. Filntisis et al. [11] improved upon the baseline performance of [28] relative to the Body Language Dataset (BoLD) by incorporating a contextual feature encoding branch and a visual-semantic embedding loss based on Global Vectors (GloVe) [35] word embeddings. Lastly, Pikoulis et al. [36] achieved state-of-the-art performance on BoLD through an intuitive approach that included the use of scene and attribute characteristics as well as multi-stream optical flow.

3. Method

A complete schematic diagram of our proposed model is shown in Fig. 1. Firstly, we will present the structure of the sub-network regarding the RGB visual modality, along with our proposed extensions for the enhancement of emotion understanding. Next, we will do the same for the aural model, and finally we will present the unified audio-visual architecture.

3.1. Seq2Seq

In order to leverage temporal information and emotion labels throughout each video, our method takes as input sequences of frames that contain either visual or aural information (which will be described in the next subsections). After extracting intermediate feature representations for each sequence of frames, we use a standard LSTM (Long Short-Term Memory) [14] model in order to map the extracted feature sequences features to their respective output labels.

3.2. Visual Model

A single RGB image usually encodes static appearance at a specific point in time but lacks the contextual information about previous and next frames. We aspire to enhance the descriptive capacity of the extracted deep visual embeddings through the feature level combination of multiple feature extractors that focus on different parts of the human instance. In our implementation, we utilize the human face as our primary source of affective information, while we also make use of the body and surrounding depicted environment, in a supplementary manner. For all of the subsequent convolutional branches, we use a standard 50-layer ResNet [13], as our feature extractor backbone. The ResNet-50 variant produces 2048-dim deep feature vector representations for each given input image. All ConvNet backbones are pre-trained using various task-specific datasets. As it will be discussed later on, pre-training constitutes the main differentiating factor among the multiple ConvNet backbones that comprise our visual model.

3.2.1 Face

The face is commonly perceived as the window to the human soul and the most expressive source of visual affective information. We introduce an input stream which explicitly operates on the aligned face crops of the primary depicted human agents. The localization, extraction and alignment of
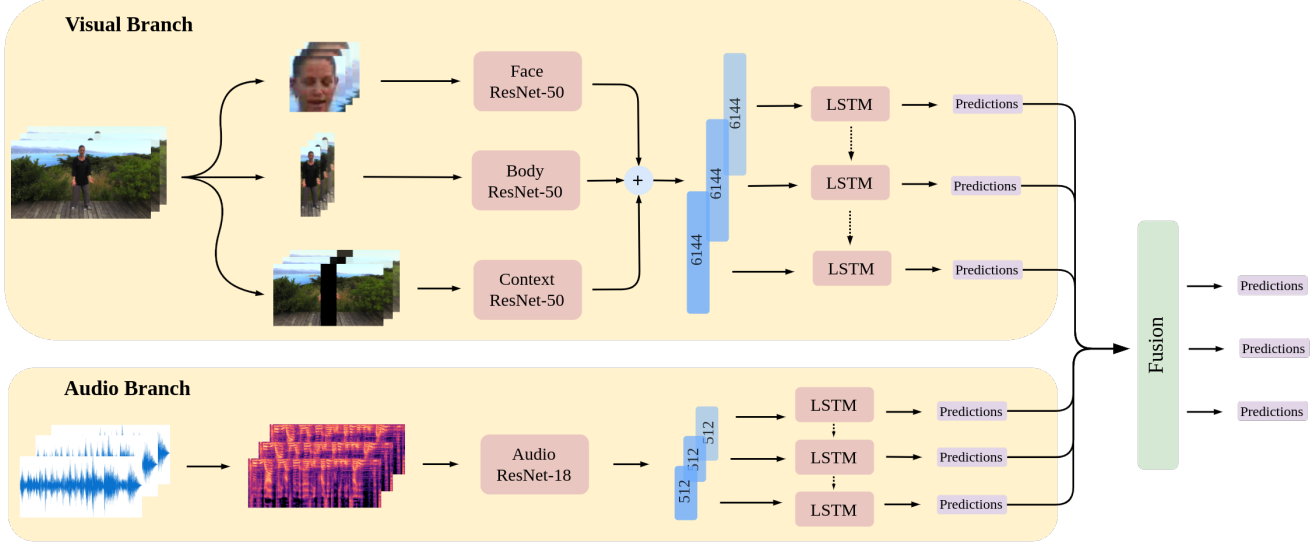
Figure 1. Overview of our proposed method.

face regions per frame has been carried out by the official distributors of the Aff-Wild2 dataset priorly. During frames where face detection and alignment has failed and the corresponding face crops are missing, we feed the ConvNet feature extractor with an input tensor of appropriate size, filled with zeros.

The ConvNet backbone of the face branch receives manual pre-training on AffectNet [31] which constitutes the largest facial expression database, containing over 1M images, annotated on both categorical and dimensional level. We pre-trained the face branch on AffectNet for 5 epochs using a batch size of 64 and a learning rate of 0.001 achieving 64.09% validation accuracy.

3.2.2 Context

We incorporate a context stream in the form of RGB frames whose primary depicted agents have been masked out. For the acquisition of the masks we use body bounding boxes and multiply them element-wise with a constant value of zero. Prior to the acquisition of the body bounding boxes and the corresponding masks, we calculate the 2D coordinates for 25 joints of the body of the primary depicted agent using the BODY25 model of the OpenPose [3] publicly available toolkit. Let $\{(x_n^{(t)}, y_n^{(t)})\}_{n=1}^{25}$ be the detected set of horizontal and vertical joint coordinates of a given agent, at frame t. The bounding box of the agent $\text{bbox}_{\text{agent}}$ within a given image I of height H and width W, is calculated as follows:

$$
\begin{aligned}
e_z &= \lambda_z(\max_n z_n^{(t)} - \min_n z_n^{(t)}), \quad z \equiv x \text{ or } z \equiv y \\
\text{bottom} &= \min\left(\max_n y_n^{(t)} + e_y, H\right), \quad \text{right} = \min\left(\max_n x_n^{(t)} + e_x, W\right) \\
\text{top} &= \max\left(0, \min_n y_n^{(t)} - e_y\right), \quad \text{left} = \max\left(0, \min_n x_n^{(t)} - e_x\right) \\
\text{bbox}_{\text{agent}} &= I(\text{top:bottom}, \text{left:right})
\end{aligned}
\tag{1}
$$

All joints with a detection confidence score that is less than 10%, are discarded. The masked image I_{mask} for a given input image I is calculated as follows:

$$
I_{\text{mask}} = \begin{cases} I(i,j) & \text{if } (i,j) \notin \text{bbox}_{\text{agent}} \\ 0 & \text{otherwise} \end{cases}
\tag{2}
$$

where the tuple (i,j) corresponds to all valid pixel locations. Throughout our experiments we set $\lambda_x = 0.1$ and $\lambda_y = 0.25$.

Contextual feature extraction is a scene-centric task, and therefore we choose to initialize the corresponding ConvNet backbone using the Places365-Standard [43], a large-scale database of photographs, labeled with scene semantic categories. The pre-trained model is publicly available[3].

3.2.3 Body

We incorporate an additional input stream that focuses solely on encoding bodily expressions of emotion, with the aim of alleviating the problem of undetected or misaligned face crops. The contribution of the body branch in the emotion recognition process becomes more evident during frames where the corresponding face crops are not existent. The body stream operates either on the bounding of the depicted agents or the entire image. The ConvNet feature extractor is pre-trained on the object-centric ImageNet [6] database. The pre-trained weights are publicly available[4].

The early feature fusion of all of the aforementioned input streams results in a 6144-dim concatenated feature vec-

---

[3]https://github.com/CSAILVision/places365
[4]https://download.pytorch.org/models/resnet50-0676ba61.pth

tor. Subsequently, the fused features are fed into a bidirectional, single-layer LSTM, with 512 hidden units, for further temporal modeling.

3.3. Aural Model

In the branch that incorporates audio information, starting with a sequence of input waveforms, we extract the mel-spectrogram representation of each. Then, we use a 18-layer ResNet model pretrained on the ImageNet dataset to extract a 512 feature vector for each input waveform. Finally, using an LSTM layer, we map the feature sequence to labels (either expression labels in the case of Track 2 or VA labels in the case of Track 1).

3.4. Loss Functions

For Track 1 of the ABAW competition, we use both a standard *mean-squared error* $\mathcal{L}_{\mathrm{mse}}$ as well as loss term based on the *concordance correlation coefficient* (CCC). The latter is defined as:

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \qquad (3)$$

where s_x and s_y denote the variance of the predicted and ground truth values respectively, $\bar{x}$ and $\bar{y}$ are the corresponding mean values and s_{xy} is the respective covariance value. The range of CCC is from -1 (perfect disagreement) to 1 (perfect agreement). Hence, in our case we define $\mathcal{L}_{\mathrm{ccc}}$ as:

$$\mathcal{L}_{\mathrm{ccc}} = 1 - \frac{\rho_v + \rho_a}{2} \qquad (4)$$

where ρ_v and ρ_a are the respective CCC of valence and arousal.

For Track 2, we use a standard cross-entropy function $\mathcal{L}_{\mathrm{ce}}$. We also enforce semantic congruity between the extracted visual embeddings and the categorical label word embeddings from a 300-dim GloVe [35] model, pre-trained on Wikipedia and Gigaword 5 data, in the same manner as in [11]. More specifically, given an input sample $\mathbf{x}$, we transform the concatenated visual embeddings $f_v(\mathbf{x})$ into the same dimensionality as the word embeddings $f_t(y)$ through a linear transformation $\mathbf{W}_{\mathrm{emb}}$, with y being the ground truth target label. We later apply an MSE loss between the transformed visual embeddings and the word embeddings which correspond to the ground truth emotional label and denote this term as $\mathcal{L}_{\mathrm{emb}}$:

$$\mathcal{L}_{\mathrm{emb}} = \left\| \mathbf{W}_{\mathrm{emb}} f_v(\mathbf{x}) - \sum_c [\![y = c]\!] f_t(y) \right\|_2^2 \qquad (5)$$

where $[\![\cdot]\!]$ is the Iverson bracket and c is the class index. The whole network can be trained in an end-to-end manner by minimizing the combined loss function $\mathcal{L} = \mathcal{L}_{\mathrm{ce}} + \lambda_{\mathrm{emb}} \mathcal{L}_{\mathrm{emb}}$. For simplicity, we set $\lambda_{\mathrm{emb}} = 1.0$ during all of our subsequent experiments.

4. Experimental Results

4.1. Validation Set

Tables 1 and 2 present our results on the Aff-Wild2 validation set, for the Expression and Valence-Arousal sub-challenges respectively, together with a performance comparison with the baseline and top entries from last year's ABAW competition [18].

Table 1. Results on the Aff-Wild2 validation set, for the Expression sub-challenge.

Method	F_1 Score	Accuracy	Total
Baseline [17]	0.30	0.50	0.366
NISL2020 [5]	-	-	0.493
ICT-VIPL [27]	0.33	0.64	0.434
TNT [26]	-	-	0.546
Audio	0.375	0.495	0.415
Visual (F)	0.453	0.584	0.496
Visual (BCF)	0.517	**0.640**	0.558
Visual (BCF) + $\mathcal{L}_{\mathrm{emb}}$	**0.532**	0.639	**0.567**
Visual (F + BCF)	0.543	0.657	0.580
Audio + Visual (F)	0.519	0.645	0.561
Audio + Visual (BCF)	0.536	0.654	0.575
Audio + Visual (F + BCF)	**0.555**	**0.668**	**0.592**

Table 2. Results on the Aff-Wild2 validation set, for the Valence-Arousal sub-challenge.

Method	CCC-V	CCC-A	Total
Baseline [17]	0.23	0.21	0.22
NISL2020 [5]	0.335	0.515	0.425
ICT-VIPL [41]	0.32	0.55	0.435
TNT [26]	0.493	0.613	0.553
Audio	0.243	0.400	0.322
Visual (F)	0.330	0.539	0.435
Visual (BCF)	**0.344**	**0.550**	**0.447**
Visual (F + BCF)	0.358	0.597	0.478
Audio + Visual (F)	0.366	0.582	0.474
Audio + Visual (BCF)	0.382	0.586	0.484
Audio + Visual (F + BCF)	**0.386**	**0.616**	**0.502**

The letter *F* denotes a visual branch trained using only the cropped face in the input video, while *BCF* denotes the visual branch that incorporates both bodily as well as contextual features, as seen in Figure 1. *Audio* denotes the results of the aural branch. In both tables we also include the results of the weighted average late score fusion among all different combinations of *F*, *BCF*, and *Audio*. We readily see that including the body and the context as additional, supplementary information results in a significant performance boost, especially in the case of the Expression sub-challenge. Furthermore, the fusion of any two methods results in increased performance, when compared to the single branches. Finally, in both Tables, the fusion of all three

Table 3. Results on the Aff-Wild2 test set relative to baseline methodologies, the top entries of the ABAW2 competition and our proposed models.

Method	Track 1			Track 2			Track 3		
	CCC-V	CCC-A	Total	F1 Score	Accuracy	Total	F1 Score	Accuracy	Total
MT-VGG [23]	0.43	0.42	0.425	-	-	0.50	-	-	0.60
V-MT-VGG-GRU [23]	0.44	0.51	0.475	-	-	0.51	-	-	0.62
A-MT-VGG-GRU [23]	0.46	0.45	0.455	-	-	0.52	-	-	0.62
A/V-MT-VGG-GRU [23]	0.47	0.52	0.495	-	-	0.53	-	-	0.63
NISL-2021 [5]	**0.533**	**0.454**	**0.494**	0.431	0.654	0.505	0.451	0.847	0.653
Maybe Next Time [38]	-	-	-	0.605	0.729	0.646	0.461	0.877	0.669
CPIC-DIR2021 [15]	-	-	-	0.683	0.771	0.712	0.489	**0.892**	0.690
Netease Fuxi Virtual Human [40]	0.486	0.495	0.490	**0.763**	**0.807**	**0.778**	**0.506**	0.888	**0.697**
Visual (F)	**0.379**	0.341	0.36	0.312	0.593	0.405	-	-	-
Visual (BCF)	0.316	0.352	0.334	0.329	0.611	0.423	-	-	-
Audio + Visual (F + BCF)	0.368	**0.464**	**0.416**	**0.337**	**0.642**	**0.437**	-	-	-

different models results in the best scores, i.e. **0.592** for the Expression sub-challenge and **0.502** for the VA sub-challenge, on the basis of the Aff-Wild2 validation set.

4.2. Test Set

Table 3 presents an extensive performance comparison among the proposed methodologies of [23], the top entries of this year's ABAW2 competition as well as our own models.

It is quite evident that our proposed model, even though it manages to surpass the baseline [17] on the validation set by a large margin, it does not come close to the performance, exhibited by the top submissions [5, 38, 15, 40]. There two main reasons that account for this observed behavior. Firstly, we presume that the distribution of the test set greatly differs from that of the validation set, leading to a significant difference in performance, relative to the two sets. Furthermore, we followed a more holistic and high-level approach to the problem, whereas the aforementioned teams laid their emphasis explicitly on the facial stream and disregarded all other sources of affective information, such as the *body*, *context* or even the *aural* modality. In that way, the production of the best facial expression model basically lies outside the scope of the current analysis. However, through our experiments we verified the beneficial influence of both the *body* and *context*, as supplementary streams that encode valuable information, especially in the case when the applied face extraction and alignment algorithms fail to due to extreme head or body orientations, low illumination and image quality.

More specifically, on the Expression sub-challenge, the sole usage of the facial stream, led to a total score of 0.405, while the inclusion of both bodily and contextual features raised the total score to 0.423. With the introduction of the aural stream, our model reached a maximum of 0.437 on the test set. On the VA sub-challenge we notice a slightly different but similar behavior. The introduction of contextual

and bodily features decreased the performance of our model significantly along the valence dimension but increased it along the arousal dimension. This can be justified by the fact that the latter is highly correlated with body motion while the former is dominated by facial expressions and AUs. The inclusion of the aural stream, once again led to a significant boost in overall performance, resulting in more than 10% increase in CCC-A and a total score of 0.416.

5. Conclusion

We presented our method for the "2nd Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW2)" challenge. Apart from using only face-cropped images, we leverage both contextual (scene), bodily features, as well as the audio of the video, to further enhance our model's perception. Our results show that fusion of the different streams of information results in significant performance increase, when compared to both single streams, as well as previous best published results on the validation set. Even though our model does not come close to the performance of this year's top entries, we have successfully highlighted the beneficial influence of the *body* and *context* input streams in the emotion recognition process, and as a result, propose our multi-stream and multi-modal approach as a potential way to further improve methodologies that solely rely on the analysis of facial expressions.

References

[1] Panagiotis Antoniadis, Panagiotis P Filntisis, and Petros Maragos. Exploiting emotional dependencies with graph convolutional networks for facial expression recognition. *arXiv preprint arXiv:2106.03487*, 2021.

[2] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. The OMG-Emotion behavior dataset. In *IEEE Int. Joint Conf. on Neural Networks (IJCNN)*, 2018.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 43:172–186, 2021.

[4] Wei-Yi Chang, Shih-Huan Hsu, and Jen-Hsien Chien. FATAUVA-Net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[5] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *15th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*. IEEE, 2020.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge 2013. In *Proc. 15th ACM Int. Conf. on Multimodal Interaction*, 2013.

[9] Paul Ekman. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological bulletin*, 1994.

[10] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17:124, 1971.

[11] P. P. Filntisis, N. Efthymiou, G. Potamianos, and P. Maragos. Emotion understanding in videos through body, context, and visual-semantic embedding loss. In *Proc. 16th Eur. Conf. Computer Vision Workshops (ECCVW) - Workshop on Bodily Expressed Emotion Understanding (BEEU)*, 2020.

[12] Behzad Hasani and Mohammad H Mahoor. Facial expression recognition using enhanced deep 3D convolutional neural networks. In *Proc. IEEE Conf on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[15] Yue Jin, Tianqing Zheng, Chao Gao, and Guoqiang Xu. A multi-modal and multi-task learning method for action unit and expression recognition. *arXiv preprint arXiv:2107.04187*, 2021.

[16] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2015.

[17] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. *arXiv preprint arXiv:2106.15318*, 2021.

[18] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first ABAW 2020 competition. In *15th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, 2020.

[19] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.

[20] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: A large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.

[21] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-Wild database and challenge, deep architectures, and beyond. *Int. Journal of Computer Vision (IJCV)*, pages 1–23, 2019.

[22] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and ArcFace. *arXiv preprint arXiv:1910.04855*, 2019.

[23] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021.

[24] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[25] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using EMOTIC dataset. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 42:2755–2766, 2019.

[26] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. In *15th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*. IEEE, 2020.

[27] Hanyu Liu, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Emotion recognition for in-the-wild videos. *arXiv preprint arXiv:2002.05447*, 2020.

[28] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang. ARBEE: Towards automated recognition of bodily expression of emotion in the wild. *Int. Journal of Computer Vision (IJCV)*, 128:1–25, 2019.

[29] David Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16:363–368, 1992.

[30] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. EmotiCon: Context-aware multimodal emotion recognition using Frege's principle. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

[32] Ali Mollahosseini, Behzad Hasani, Michelle J Salvador, Hojjat Abdollahi, David Chan, and Mohammad H Mahoor. Facial expression recognition from world wild web. In *Proc.*

IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2016.

[33] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. on Affective Computing*, 2:92–105, 2011.

[34] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2005.

[35] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proc. 2014 Conf. Empirical Methods Natural Language Processing (EMNLP)*, 2014.

[36] Ioannis Pikoulis, Panagiotis P Filntisis, and Petros Maragos. Leveraging semantic scene characteristics and multi-stream convolutional architectures in a contextual approach for video-based visual emotion recognition in the wild. *arXiv preprint arXiv:2105.07484*, 2021.

[37] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.

[38] Phan Tran Dac Thinh, Hoang Manh Hung, Hyung-Jeong Yang, Soo-Hyung Kim, and Guee-Sang Lee. Emotion recognition with incomplete labels using modified multi-task learning technique. *arXiv preprint arXiv:2107.04192*, 2021.

[39] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-Wild: Valence and arousal 'in-the-wild' challenge. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[40] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. *arXiv preprint arXiv:2107.03708*, 2021.

[41] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, and Shiguang Shan. M^3T: Multi-modal continuous valence-arousal estimation in the wild. In *15th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*. IEEE, 2020.

[42] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *Eur. Conf. on Computer Vision (ECCV)*. Springer, 2016.

[43] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 40:1452–1464, 2018.