Analysing Affective Behavior
in the second ABAW2 Competition

Dimitrios Kollias
University of Greenwich, UK
D.Kollias@greenwich.ac.uk

Irene Kotsia
Middlesex University London, UK

Elnar Hajiyev
Realeyes

Stefanos Zafeiriou
Imperial College London, UK

Abstract

The Affective Behavior Analysis in-the-wild (ABAW2) 2021 Competition is the second -following the first very successful ABAW Competition held in conjunction with IEEE FG 2020- Competition that aims at automatically analyzing affect. ABAW2 is split into three Challenges, each one addressing one of the three main behavior tasks of Valence-Arousal Estimation, seven Basic Expression Classification and twelve Action Unit Detection. All three Challenges are based on a common benchmark database, Aff-Wild2, which is a large scale in-the-wild database and the first one to be annotated for all these three tasks. In this paper, we describe this Competition, to be held in conjunction with ICCV 2021. We present the three Challenges, with the utilized Competition corpora. We outline the evaluation metrics and present the baseline system with its results. More information regarding the Competition is provided in the Competition site: https://ibug.doc.ic.ac.uk/resources/iccv-2021-2nd-abaw/.

1. Introduction

The proposed Workshop tackles the problem of affective behavior analysis in-the-wild, which is a major targeted characteristic of HCI systems used in real life applications. The current 5th societal revolution aims at merging the physical and cyber spaces, providing services that contribute to people's well-being. The target is to create machines and robots that are capable of understanding people's feelings, emotions and behaviors; thus, being able to interact in a 'human-centered' and engaging manner with them, and effectively serving them as their digital assistants.

Affective behavior analysis in diverse environments, such as in people's homes, in their work, operational or industrial environments, will have a positive societal impact. It will provide machines and robots with the ability to interact and assist people in an effective and natural way. Through human affect recognition, the reactions of the machine, or robot, will be consistent with people's expectations and emotions; their verbal and non-verbal interactions will be positively received by humans. Moreover, this interaction should not be dependent on the respective context, nor the human's age, sex, ethnicity, educational level, profession, or social position. As a result, the development of intelligent systems able to analyze human behaviors in-the-wild can contribute to generation of trust, understanding and closeness between humans and machines in real life environments.

Representing human emotions has been a basic topic of research in psychology. The most frequently used emotion representation is the categorical one, including the seven basic categories, i.e., Anger, Disgust, Fear, Happiness, Sadness, Surprise and Neutral [11]. Discrete emotion representation can also be described in terms of the Facial Action Coding System (FACS) model, in which all possible facial actions are described in terms of Action Units (AUs) [10]. Finally, the dimensional model of affect [61, 54] has been proposed as a means to distinguish between subtly different displays of affect and encode small changes in the intensity of each emotion on a continuous scale. The 2-D Valence and Arousal (VA) Space (valence shows how positive or negative an emotional state is, whereas arousal shows how passive or active it is) is the most usual dimensional emotion representation, depicted in Figure 1.

There are a number of related applications spread across a variety of fields, such as medicine, health, driver fatigue, monitoring, e-learning, marketing, entertainment, lie detection, law [55, 1, 30, 17, 63, 56, 32, 48, 21, 20, 4, 59, 60, 46, 52, 57, 2, 42, 12, 18, 19].

The ABAW2 Competition contains three Challenges, which are based on the same database; these target (i) dimensional affect recognition (in terms of valence and arousal) [16, 35, 6, 22, 5, 41, 65, 39, 23, 3], (ii) categor-
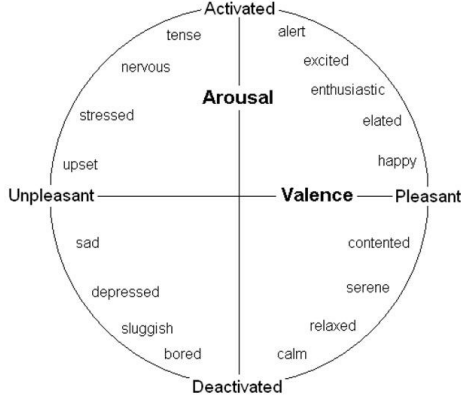
Figure 1. The 2D Valence-Arousal Space

ical affect classification (in terms of the seven basic expressions) [49, 9, 24, 66, 29, 15, 33, 45, 37] and (iii) 12 facial action unit detection [50, 44, 28, 13, 27, 8], in-the-wild. These Challenges produce a significant step forward when compared to previous events. In particular, they use the Aff-Wild2 [26, 38, 40, 34, 36], the first comprehensive benchmark for all three affect recognition tasks in-the-wild: the Aff-Wild2 database extends the Aff-Wild [31, 64, 25], with more videos and annotations for all behavior tasks. Aff-Wild consists of 298 videos, displaying reactions of 200 subjects, with a total video duration of about 30 hours, and 1,250,000 video frames, annotated in terms of valence and arousal. It has been used in the Aff-Wild Challenge in CVPR 2017, with participation of more than 10 research groups. To generate Aff-Wild2, we added 266 more videos, displaying the reactions of 266 more subjects, with a duration of more than 18 hours, and 1,500,000 frames. Aff-Wild2 includes extended spontaneous facial behaviors in arbitrary recording conditions and a significantly increased number of different subjects (466; 280 of which are males and 186 females) and frames (around 2,800,000).

The remainder of this paper is organised as follows. We introduce the Competition corpora in Section 2, the Competition evaluation metrics in Section 3, the developed baseline per Challenge, along with the obtained results in Section 4, before concluding in Section 5.

2. Competition Corpora

The second Affective Behavior Analysis in-the-wild (ABAW2) Competition relies on the Aff-Wild2 database [38, 34, 36]. Aff-Wild2 is the first ever database annotated for all three main behavior tasks: valence-arousal estimation, action unit detection and basic expression classification. These three tasks form the three Challenges of this Competition.

Aff-Wild2 consists of 548 videos with $2, 813, 201$ frames. Sixteen of these videos display two subjects (both

have been annotated). All videos have been collected from YouTube. Aff-Wild2 is an extension of Aff-Wild [31, 64, 25]; 260 more YouTube videos, with $1, 413, 000$ frames, have been added to Aff-Wild. Aff-Wild was the first large scale, captured in-the-wild, dimensionally annotated database, containing 298 YouTube videos that display subjects reacting to a variety of stimuli. Aff-Wild2 shows both subtle and extreme human behaviours in real-world settings. The total number of subjects in Aff-Wild2 is 458; 279 of them are males and 179 females.

The Aff-Wild2 database, in all Challenges, is split into training, validation and test set. At first the training and validation sets, along with their corresponding annotations, are being made public to the participants, so that they can develop their own methodologies and test them. The training and validation data contain the videos and their corresponding annotation. Furthermore, to facilitate training, especially for people that do not have access to face detectors/tracking algorithms, we provide bounding boxes and landmarks for the face(s) in the videos (we also provide the aligned faces). At a later stage, the test set without annotations will be given to the participants. Again, we will provide bounding boxes and landmarks for the face(s) in the videos (we will also provide the aligned faces).

In the following, we provide a short overview of each Challenge's dataset and refer the reader to the original work for a more complete description. Finally, we describe the pre-processing steps that we carried out for cropping and aligning the images of Aff-Wild2. The cropped and aligned images have been utilized in our baseline experiments.

2.1. Aff-Wild2: Valence-Arousal Annotation

545 videos in Aff-Wild2 contain annotations in terms of valence-arousal. Sixteen of these videos display two subjects, both of which have been annotated. In total, $2, 786, 201$ frames, with 455 subjects, 277 of which are male and 178 female, have been annotated by four experts using the method proposed in [7]. The annotators watched each video and provided their (frame-by-frame) annotations through a joystick. A time-continuous annotation was generated for each affect dimension. Valence and arousal values range continuously in $[-1, 1]$. The final label values were the mean of those four annotations. The mean inter-annotation correlation is 0.63 for valence and 0.60 for arousal. Let us note here that all subjects present in each video have been annotated. Figure 2 shows the 2D Valence-Arousal histogram of annotations of Aff-Wild2.

Aff-Wild2 is currently the largest (and audiovisual) in-the-wild database annotated for valence and arousal.

Aff-Wild2 is split into three subsets: training, validation and test. Partitioning is done in a subject independent manner, in the sense that a person can appear only in one of those three subsets. The resulting training, validation and
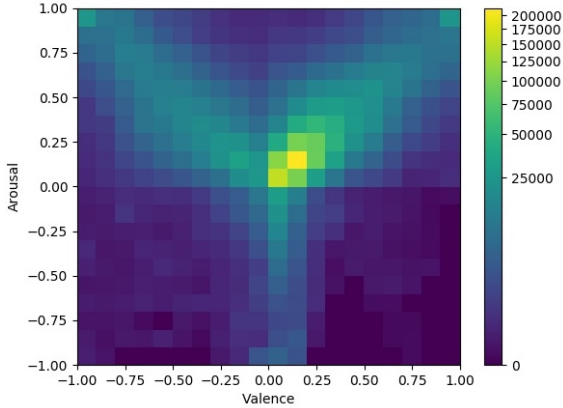
Figure 2. 2D Valence-Arousal Histogram of Aff-Wild2

Table 1. Number of Annotated Images in Each of the Seven Basic Expressions

Basic Expression	No of Images
Neutral	538,411
Anger	52,005
Disgust	31,138
Fear	26,062
Happiness	395,352
Sadness	173,842
Surprise	99,863

tal, $2,565,169$ frames, with 426 subjects, 262 of which are male and 164 female, have been annotated in a semi-automatic procedure (that involves manual and automatic annotations). Aff-Wild2 has been annotated for the occurrence of twelve action units in a frame-by-frame basis. Table 2 shows the name of the twelve action units that have been annotated, the action that they are associated with and the distribution of their annotations in Aff-Wild2.

Aff-Wild2 is currently the largest (and audiovisual) in-the-wild database annotated in terms of action units.

Table 2. Distribution of AU annotations in Aff-Wild2

Action Unit #	Action	Total Number of Activated AUs
AU 1	inner brow raiser	294,591
AU 2	outer brow raiser	136,569
AU 4	brow lowerer	384,969
AU 6	cheek raiser	618,929
AU 7	lid tightener	618,929
AU 10	upper lip raiser	845,793
AU 12	lip corner puller	598,699
AU 15	lip corner depressor	62,954
AU 23	lip tightener	77,793
AU 24	lip pressor	61,460
AU 25	lips part	1,579,262
AU 26	jaw drop	202,447

Aff-Wild2 is split into three subsets: training, validation and test. Partitioning is done in a subject independent manner. The resulting training, validation and test subsets consist of 302, 105 and 127 videos, respectively; the resulting training, validation and test subsets contain 3, 0 and 4, respectively, videos that display two subjects.

2.4. Aff-Wild2 Pre-Processing: Cropped & Cropped-Aligned Images

At first, we split all videos into images (frames). Then, the SSH face detector [47, 58] based on the ResNet [14] and trained on the WiderFace dataset [62] was used to extract face bounding boxes from all the images. The cropped images according to these bounding boxes were provided

test subsets consist of 346, 68 and 131 videos, respectively; the resulting training, validation and test subsets contain 5, 3 and 8, respectively, videos that display two subjects.

2.2. Aff-Wild2: Seven Basic Expression Annotation

539 videos in Aff-Wild2 contain annotations in terms of the seven basic expressions. Seven of these videos display two subjects, both of which have been annotated. In total, $2,595,572$ frames, with 431 subjects, 265 of which are male and 166 female, have been annotated by seven experts in a frame-by-frame basis. A platform-tool was developed in order to split each video into frames and let the experts annotate each videoframe. Let us mention that in this platform-tool, an expert could score a videoframe as having either one of the seven basic expressions or none (since there are affective states other than the seven basic expressions).

Due to subjectivity of annotators and wide ranging levels of images' difficulty, there were some disagreements among annotators. We decided to keep only the annotations on which at least six (out of seven) experts agreed. Table 1 shows the distribution of the seven basic expression annotations of Aff-Wild2.

Aff-Wild2 is currently the largest (and audiovisual) in-the-wild database annotated in terms of the seven basic expressions.

Aff-Wild2 is split into three subsets: training, validation and test. Partitioning is done in a subject independent manner. The resulting training, validation and test subsets consist of 250, 70 and 222 videos, respectively; the resulting training, validation and test subsets contain 3, 0 and 1, respectively, videos that display two subjects.

2.3. Aff-Wild2: Twelve Action Unit Annotation

534 videos in Aff-Wild2 contain annotations in terms of twelve action units. Seven of these videos display two subjects, both of which have been annotated. In to-

to the participating teams. Also, 5 facial landmarks (two eyes, nose and two mouth corners) were extracted and used to perform similarity transformation. The resulting cropped and aligned images were additionally provided to the participating teams. Finally, the cropped and aligned images were utilized in our baseline experiments, described in Section 4.

3. Evaluation Metrics Per Challenge

Next, we present the metrics that will be used for assessing the performance of the developed methodologies of the participating teams in each Challenge.

3.1. Valence-Arousal Estimation Challenge

The Concordance Correlation Coefficient (CCC) is widely used in measuring the performance of dimensional emotion recognition methods, such as in the series of AVEC challenges [53]. CCC evaluates the agreement between two time series (e.g., all video annotations and predictions) by scaling their correlation coefficient with their mean square difference. In this way, predictions that are well correlated with the annotations but shifted in value are penalized in proportion to the deviation. CCC takes values in the range $[-1, 1]$, where $+1$ indicates perfect concordance and -1 denotes perfect discordance. The highest the value of the CCC the better the fit between annotations and predictions, and therefore high values are desired. CCC is defined as follows:

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2},\qquad(1)$$

where s_x and s_y are the variances of all video valence/arousal annotations and predicted values, respectively, $\bar{x}$ and $\bar{y}$ are their corresponding mean values and s_{xy} is the corresponding covariance value.

The mean value of CCC for valence and arousal estimation will be adopted as the main evaluation criterion.

$$\mathcal{E}_{total} = \frac{\rho_a + \rho_v}{2},\qquad(2)$$

3.2. Seven Basic Expression Classification Challenge

The F_1 score is a weighted average of the recall (i.e., the ability of the classifier to find all the positive samples) and precision (i.e., the ability of the classifier not to label as positive a sample that is negative). The F_1 score reaches its best value at 1 and its worst score at 0. The F_1 score is defined as:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}\qquad(3)$$

The F_1 score for emotions is computed based on a per-frame prediction (an emotion category is specified in each frame).

Total accuracy (denoted as $\mathcal{T}Acc$) is defined on all test samples and is the fraction of predictions that the model got right. Total accuracy reaches its best value at 1 and its worst score at 0. It is defined as:

$$\mathcal{T}Acc = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}\qquad(4)$$

A weighted average between the F_1 score and the total accuracy, $\mathcal{T}Acc$, will be the main evaluation criterion:

$$\mathcal{E}_{total} = 0.67 \times F_1 + 0.33 * \mathcal{T}Acc,\qquad(5)$$

3.3. Twelve Action Unit Detection Challenge

To obtain the overall score for the AU detection Challenge, we first obtain the F_1 score for each AU independently, and then compute the (unweighted) average over all 12 AUs (denoted as $\mathcal{A}F_1$) :

$$\mathcal{A}F_1 = \sum_{i=1}^{12} F_1^i\qquad(6)$$

The F_1 score for AUs is computed based on a per-frame detection (whether each AU is present or absent).

The average between the $\mathcal{A}F_1$ score and the total accuracy, $\mathcal{T}Acc$, will be the main evaluation criterion:

$$\mathcal{E}_{total} = 0.5 \times \mathcal{A}F_1 + 0.5 * \mathcal{T}Acc\qquad(7)$$

4. Baseline & Participating Teams' Systems and Results

All baseline systems rely exclusively on existing open-source machine learning toolkits to ensure the reproducibility of the results. In this Section, we first describe the baseline systems developed for each Challenge and then report their obtained results.

At first, let us mention that we utlized the cropped and aligned images from Aff-Wild2, as described in Section 2.4. These images have dimensions $112 \times 112 \times 3$. The pixel intensities are normalized to take values in [-1,1]. No on-the-fly or off-the-fly data augmentation technique [43, 22, 23] was utilized.

4.1. Baseline Systems

The architecture that was used in all 3 Challenges was based on the 13 convolutional and pooling layers of VGG-FACE [51] (its fully connected layers are discarded), followed by 2 fully connected layers, each with 4096 hidden units. In the Valence-Arousal Estimation Challenge baseline, a (linear) output layer follows that gives final estimates
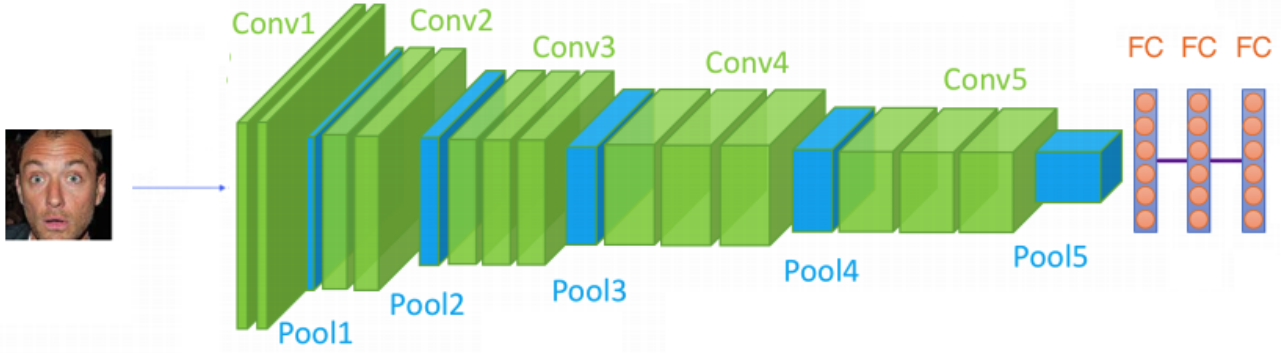
Figure 3. The architecture of the utilized baseline VGG-FACE that has been used in all Challenges; the output is either linear (VA case), or with a softmax unit (in the 7 basic expression case), or with a sigmoid unit (in the 12 AU case)

for valence and arousal. In the Seven Basic Expression Classification Challenge, a final output layer with softmax as activation function follows which gives the 7 basic expression predictions. In the twelve Action Unit Detection Challenge, a final output layer with sigmoid as activation function follows which gives the 12 action unit predictions. Figure 3 shows this basic architecture of the utilized VGG-FACE.

The baseline systems have been pre-trained on the VGG-Face dataset; their convolutional layers were fixed (i.e., non-trainable) and only the three fully connected were trained on Aff-Wild2. These systems have been implemented in TensorFlow; training time was around a day on a Titan X GPU, with a learning rate of 10^{-4} and with a batch size of 256.

4.2. Results

Table 3 presents the CCC evaluation of valence and arousal predictions on the Aff-Wild2 validation set, of the baseline network (VGG-FACE).

Table 3. Baseline results for VA estimation on the validation set of Aff-Wild2; $\mathcal{E}_{total}$ is the mean valence and arousal CCC

Baseline	CCC		$\mathcal{E}_{total}$
	Valence	Arousal	
VGG-FACE	0.23	0.21	0.22

Table 4 presents the performance on the validation set of Aff-Wild2, of the baseline network (VGG-FACE) of the seven basic expression classification Challenge. The performance metric is a weighted average between the F1 score and the total accuracy, as discussed in Section 3.2.

Table 5 presents the performance on the validation set of Aff-Wild2, of the baseline network VGG-FACE of the twelve Action Unit Detection Challenge. The performance metric is the average between the F1 score and the total accuracy, as discussed in Section 3.3.

Table 4. Baseline results for the seven basic expression classification on the validation set of Aff-Wild2; $\mathcal{E}_{total} = 0.67 \times F_1 + 0.33 * \mathcal{T}Acc$

Baseline	F1 Score	Total Accuracy	$\mathcal{E}_{total}$
VGG-FACE	0.30	0.50	0.366

Table 5. Baseline results for 12 action unit detection on the validation set of Aff-Wild2; $\mathcal{E}_{total} = 0.5 \times \mathcal{A}F_1 + 0.5 * \mathcal{T}Acc$

Baseline	Average F1 Score	Total Accuracy	$\mathcal{E}_{total}$
VGG-FACE	0.40	0.22	0.31

5. Conclusion

In this paper we have presented the second Affective Behavior Analysis in-the-wild Competition (ABAW2) 2020. It comprises three Challenges targeting: i) valence-arousal estimation, ii) seven basic expression classification and iii) eight action unit detection. The database utilized for this Competition has been derived from the Aff-Wild2, the large-scale and first database annotated for all these three behavior tasks. We have also presented the baseline networks and their results.

References

[1] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Hojjat Adeli, and D Puthankattil Subha. Automated eeg-based screening of depression using deep convolutional neural network. *Computer methods and programs in biomedicine*, 161:103–113, 2018.

[2] Dimitris Anastassiou and Stefanos Kollias. Digital image halftoning using neural networks. In *Visual Communications and Image Processing'88: Third in a Series*, volume 1001, pages 1062–1069. International Society for Optics and Photonics, 1988.

[3] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter.

The omg-emotion behavior dataset. *arXiv preprint arXiv:1803.05434*, 2018.

[4] G Caridakis, A Raouzaiou, K Karpouzis, and S Kollias. Synthesizing gesture expressivity based on real sequences. In *Workshop Programme*, volume 10, page 19.

[5] Wei-Yi Chang, Shih-Huan Hsu, and Jen-Hsien Chien. Fatauva-net : An integrated deep learning framework for facial attribute recognition, action unit (au) detection, and valence-arousal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017.

[6] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 19–26. ACM, 2017.

[7] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder. 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[8] Didan Deng, Zhaokang Chen, and Bertram E Shi. Fau, facial expressions, valence and arousal: A multi-task solution. *arXiv preprint arXiv:2002.03557*, 2020.

[9] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 118–126. IEEE, 2017.

[10] Paul Ekman. Facial action coding system (facs). *A human face*, 2002.

[11] Paul Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003.

[12] Birte Glimm, Yevgeny Kazakov, Ilianna Kollia, and Giorgos B Stamou. Using the tbox to optimise sparql queries.

[13] Shizhong Han, Zibo Meng, Ahmed-Shehab Khan, and Yan Tong. Incremental boosting convolutional neural network for facial action unit recognition. In *Advances in neural information processing systems*, pages 109–117, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[15] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2983–2991, 2015.

[16] Pooya Khorrami, Tom Le Paine, Kevin Brady, Charlie Dagli, and Thomas S Huang. How deep neural networks can improve emotion recognition on video data. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 619–623. IEEE, 2016.

[17] Junghoe Kim, Vince D Calhoun, Eunsoo Shim, and Jong-Hwan Lee. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *Neuroimage*, 124:127–146, 2016.

[18] Ilianna Kollia, Birte Glimm, and Ian Horrocks. Answering queries over owl ontologies with sparql.

[19] Ilianna Kollia, Yannis Kalantidis, Kostas Rapantzikos, and Andreas Stafylopatis. Improving semantic search in digital libraries using multimedia analysis. *Journal of Multimedia*, 7(2), 2012.

[20] Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian, and Stefanos Kollias. Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. *arXiv preprint arXiv:2106.07524*, 2021.

[21] Dimitrios Kollias, N Bouas, Y Vlaxos, V Brillakis, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and S Kollias. Deep transparent prediction through latent representation analysis. *arXiv preprint arXiv:2009.07044*, 2020.

[22] Dimitrios Kollias, Shiyang Cheng, Maja Pantic, and Stefanos Zafeiriou. Photorealistic facial synthesis in the dimensional affect space. In *European Conference on Computer Vision*, pages 475–491. Springer, 2018.

[23] Dimitrios Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, pages 1–30, 2020.

[24] Dimitris Kollias, George Marandianos, Amaryllis Raouzaiou, and Andreas-Georgios Stafylopatis. Interweaving deep learning and semantic techniques for emotion analysis in human-machine interaction. In *2015 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 1–6. IEEE, 2015.

[25] Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1972–1979. IEEE, 2017.

[26] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. IEEE Computer Society.

[27] 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.

[28] 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.

[29] Dimitrios Kollias, Athanasios Tagaris, and Andreas Stafylopatis. On line emotion detection using retrainable deep neural networks. In *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*, pages 1–8. IEEE, 2016.

[30] Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos Kollias, and Georgios Tagaris. Deep neural architectures for prediction in healthcare. *Complex & Intelligent Systems*, 4(2):119–131, 2018.

[31] 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. *International Journal of Computer Vision*, 127(6-7):907–929, 2019.

[32] Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and Stefanos D Kollias. Transparent adaptation in deep medical image diagnosis. In *TAILOR*, pages 251–267, 2020.

[33] Dimitrios Kollias, Miao Yu, Athanasios Tagaris, Georgios Leontidis, Andreas Stafylopatis, and Stefanos Kollias. Adaptation and contextualization of deep neural network models. In *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*, pages 1–8. IEEE, 2017.

[34] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018.

[35] Dimitrios Kollias and Stefanos Zafeiriou. A multi-component cnn-rnn approach for dimensional emotion recognition in-the-wild. *arXiv preprint arXiv:1805.01452*, 2018.

[36] Dimitrios Kollias and Stefanos Zafeiriou. A multi-task learning & generation framework: Valence-arousal, action units & primary expressions. *arXiv preprint arXiv:1811.07771*, 2018.

[37] Dimitrios Kollias and Stefanos Zafeiriou. Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[38] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[39] Dimitrios Kollias and Stefanos Zafeiriou. Va-stargan: Continuous affect generation. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 227–238. Springer, 2020.

[40] 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.

[41] Dimitrios Kollias and Stefanos P Zafeiriou. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing*, 2020.

[42] S Kollias and Dimitris Anastassiou. Adaptive training of multilayer neural networks using a least squares estimation technique. 1988.

[43] Michael Kuchnik and Virginia Smith. Efficient augmentation via data subsampling. *arXiv preprint arXiv:1810.05222*, 2018.

[44] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. *arXiv preprint arXiv:2002.03399*, 2020.

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

[46] Lori Malatesta, Amaryllis Raouzaiou, Kostas Karpouzis, and S Kollias. Towards modeling embodied conversational agent character profiles using appraisal theory predictions in expression synthesis. *Applied intelligence*, 30(1):58–64, 2009.

[47] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry Davis. SSH: Single stage headless face detector. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[48] Ibrahim M Nasser, Mohammed O Al-Shawwa, and Samy S Abu-Naser. Artificial neural network for diagnose autism spectrum disorder. 2019.

[49] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 443–449. ACM, 2015.

[50] Jaspar Pahl, Ines Rieger, and Dominik Seuss. Multi-label class balancing algorithm for action unit detection. *arXiv preprint arXiv:2002.03238*, 2020.

[51] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[52] Konstantions Rapantzikos, Yannis Avrithis, and Stefanos Kollias. Spatiotemporal saliency for event detection and representation in the 3d wavelet domain: potential in human action recognition. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 294–301, 2007.

[53] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 3–12, 2019.

[54] James A Russell. Evidence of convergent validity on the dimensions of affect. *Journal of personality and social psychology*, 36(10):1152, 1978.

[55] Athanasios Tagaris, Dimitrios Kollias, and Andreas Stafylopatis. Assessment of parkinson's disease based on deep neural networks. In *International Conference on Engineering Applications of Neural Networks*, pages 391–403. Springer, 2017.

[56] Athanasios Tagaris, Dimitrios Kollias, Andreas Stafylopatis, Georgios Tagaris, and Stefanos Kollias. Machine learning for neurodegenerative disorder diagnosis—survey of practices and launch of benchmark dataset. *International Journal on Artificial Intelligence Tools*, 27(03):1850011, 2018.

[57] Nicolas Tsapatsoulis, Kostas Karpouzis, George Stamou, Frederic Piat, and Stefanos Kollias. A fuzzy system for emotion classification based on the mpeg-4 facial definition parameter set. In *2000 10th European Signal Processing Conference*, pages 1–4. IEEE, 2000.

[58] Nicolas Tsapatsoulis and Stefanos Kollias. Face detection in color images and video sequences. In *2000 10th Mediterranean Electrotechnical Conference. Information Technology and Electrotechnology for the Mediterranean Countries. Proceedings. MeleCon 2000 (Cat. No. 00CH37099)*, volume 2, pages 498–502. IEEE, 2000.

[59] Paraskevi Tzouveli, Andreas Schmidt, Michael Schneider, Antonis Symvonis, and Stefanos Kollias. Adaptive reading assistance for the inclusion of students with dyslexia: The agent-dysl approach. In *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, pages 167–171. IEEE, 2008.

[60] Manolis Wallace, Nicolas Tsapatsoulis, and Stefanos Kollias. Intelligent initialization of resource allocating rbf networks. *Neural Networks*, 18(2):117–122, 2005.

[61] CM Whissel. The dictionary of affect in language, emotion: Theory, research and experience: vol. 4, the measurement of emotions, r. *Plutchik and H. Kellerman, Eds., New York: Academic*, 1989.

[62] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[63] Miao Yu, Dimitrios Kollias, James Wingate, Niro Siriwardena, and Stefanos Kollias. Machine learning for predictive modelling of ambulance calls. *Electronics*, 10(4):482, 2021.

[64] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–41, 2017.

[65] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, Shiguang Shan, and Xilin Chen. m^3 t: Multi-modal continuous valence-arousal estimation in the wild. *arXiv preprint arXiv:2002.02957*, 2020.

[66] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *European conference on computer vision*, pages 425–442. Springer, 2016.