ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Multi-Task Learning Challenges

Dimitrios Kollias
Queen Mary University of London, UK
d.kollias@qmul.ac.uk

Abstract

This paper describes the third Affective Behavior Analysis in-the-wild (ABAW) Competition, held in conjunction with IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2022. The 3rd ABAW Competition is a continuation of the Competitions held at ICCV 2021, IEEE FG 2020 and IEEE CVPR 2017 Conferences, and aims at automatically analyzing affect. This year the Competition encompasses four Challenges: i) uni-task Valence-Arousal Estimation, ii) uni-task Expression Classification, iii) uni-task Action Unit Detection, and iv) Multi-Task-Learning. All the 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 in terms of valence-arousal, expressions and action units. In this paper, we present the four Challenges, with the utilized Competition corpora, we outline the evaluation metrics and present both the baseline systems and the top performing teams' per Challenge. Finally we illustrate the obtained results of the baseline systems and of all participating teams. More information regarding the Competition and the leaderboard for each Challenge can be found in the competition's website: http://ibug.doc.ic.ac.uk/resources/cvpr-2022-3rd-abaw.

1. Introduction

Affect recognition based on a subject's facial expressions has been a topic of major research in the attempt to generate machines that can understand the way subjects feel, act and react. The problem of affect analysis and recognition constitutes a key issue in behavioural modelling, human computer/machine interaction and affective computing. There are a number of related applications spread across a variety of fields, such as medicine, health, or driver fatigue, monitoring, e-learning, marketing, entertainment, lie detection and law.

In the past, due to the unavailability of large amounts of data captured in real-life situations, research has mainly focused on controlled environments. However, recently, social media and platforms have been widely used and large amount of data have become available. Moreover, deep learning has emerged as a means to solve visual analysis and recognition problems. Thus, major research has been given during the last few years to the development and use of deep learning techniques and deep neural networks [17, 39] in various applications, including affect recognition in-the-wild, i.e., in unconstrained environments. Moreover, apart from affect analysis and recognition, generation of facial affect is of great significance, in many real life applications, such as for synthesis of affect on avatars that interact with humans, in computer games, in augmented and virtual environments, in educational and learning contexts. The ABAW Workshop exploits these advances and makes significant contributions for affect analysis, recognition and synthesis in-the-wild.

Ekman [13] was the first to systematically study human facial expressions. His study categorizes the prototypical facial expressions, apart from neutral expression, into six classes representing anger, disgust, fear, happiness, sadness and surprise. Furthermore, facial expressions are related to specific movements of facial muscles, called Action Units (AUs). The Facial Action Coding System (FACS) was developed, in which facial changes are described in terms of AUs [5].

Apart from the above categorical definition of facial expressions and related emotions, in the last few years there has been great interest in dimensional emotion representations, which are of great interest in human computer interaction and human behaviour analysis. Dimensional emotion representations are used to tag emotional states in continuous mode, usually in terms of the arousal and valence dimensions, i.e. in terms of how active or passive, positive or negative is the human behaviour under analysis [14, 49, 57].

The third ABAW Competition, held in conjunction with the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2022 is a continuation of the

first [1] [29] and second [2] [37] ABAW Competitions held in conjunction with the IEEE Conference on Face and Gesture Recognition (IEEE FG) 2021 and with the International Conference on Computer Vision (ICCV) 2022, respectively, which targeted dimensional (in terms of valence and arousal) [1–3, 8, 9, 11, 24, 40, 45, 53, 54, 58, 64, 65, 67], categorical (in terms of the basic expressions) [12, 15, 16, 38, 41, 42, 60] and facial action unit analysis and recognition [7, 18, 22, 30, 31, 46, 50, 53]. The third ABAW Competition contains four Challenges, which are based on the same in-the-wild database, (i) the uni-task Valence-Arousal Estimation Challenge; (ii) the uni-task Expression Classification Challenge (for the 6 basic expressions plus the neutral state plus the 'other' category that denotes expressions/affective states other than the 6 basic ones); (iii) the uni-task Action Unit Detection Challenge (for 12 action units); (iv) the Multi-Task Learning Challenge (for joint learning and predicting of valence-arousal, 8 expressions -6 basic plus neutral plus 'other'- and 12 action units). These Challenges produce a significant step forward when compared to previous events. In particular, they use the Aff-Wild2 [28–37, 62], the first comprehensive benchmark for all three affect recognition tasks in-the-wild: the Aff-Wild2 database extends the Aff-Wild [28,32,62], with more videos and annotations for all behavior tasks.

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 and the top performing teams per Challenge, along with the obtained results in Section 4, before concluding in Section 5.

2. Competition Corpora

The third Affective Behavior Analysis in-the-wild (ABAW2) Competition relies on the Aff-Wild2 database, which is the first ever database annotated for all three main behavior tasks: valence-arousal estimation, action unit detection and expression classification. These three tasks constitute the basis of the four Challenges.

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. Valence-Arousal Estimation Challenge

This Challenge's corpora include 564 videos in Aff-Wild2 that contain annotations in terms of valence and arousal. Sixteen of these videos display two subjects, both of which have been annotated. In total, $2,816,832$ frames, with 455 subjects, 277 of which are male and 178 female, have been annotated by four experts using the method proposed in [4]. Valence and arousal values range continuously in $[-1, 1]$. Figure 1 shows the 2D Valence-Arousal histogram of annotations of Aff-Wild2.

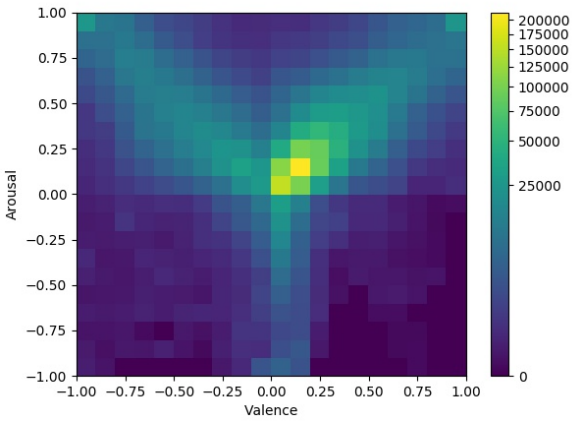

Figure 1. Valence-Arousal Estimation Challenge: 2D Valence-Arousal Histogram of Annotations in Aff-Wild2

Aff-Wild2 is split into training, validation and testing sets. Partitioning is done in a subject independent manner, in the sense that a person can appear strictly in only one of these sets. These sets consist of 341, 71 and 152 videos, respectively.

2.2. Expression Classification Challenge

This Challenge's corpora include 546 videos in Aff-Wild2 that contain annotations in terms of the the 6 basic expressions, plus the neutral state, plus a category 'other' that denotes expressions/affective states other than the 6 basic ones. Seven of these videos display two subjects, both of which have been annotated. In total, $2,624,160$ frames, with 437 subjects, 268 of which are male and 169 female,

[1] https://ibug.doc.ic.ac.uk/resources/iccv-2021-2nd-abaw/

[2] https://ibug.doc.ic.ac.uk/resources/fg-2020-competition-affective-behavior-analysis/

have been annotated by seven experts in a frame-by-frame basis. Table 1 shows the distribution of the expression annotations of Aff-Wild2.

Table 1. Expression Classification Challenge: Number of Annotated Images for each Expression

Expressions	No of Images
Neutral	468,069
Anger	36,627
Disgust	24,412
Fear	19,830
Happiness	245,031
Sadness	130,128
Surprise	68,077
Other	512,262

Aff-Wild2 is split into training, validation and testing sets, in a subject independent manner. These sets consist of 248, 70 and 228 videos, respectively.

2.3. Action Unit Detection Challenge

This Challenge's corpora include 541 videos that contain annotations in terms of 12 AUs, namely AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25 and AU26. Seven of these videos display two subjects, both of which have been annotated. In total, $2,627,632$ frames,with 438 subjects, 268 of which are male and 170 female, have been annotated in a semi-automatic procedure (that involves manual and automatic annotations). The annotation has been performed 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.

Table 2. Action Unit Detection Challenge: Distribution of AU Annotations in Aff-Wild2

Action Unit #	Action	Total Number of Activated AUs
AU 1	inner brow raiser	301,102
AU 2	outer brow raiser	139,936
AU 4	brow lowerer	386,689
AU 6	cheek raiser	619,775
AU 7	lid tightener	964,312
AU 10	upper lip raiser	854,519
AU 12	lip corner puller	602,835
AU 15	lip corner depressor	63,230
AU 23	lip tightener	78,649
AU 24	lip pressor	61,500
AU 25	lips part	1,596,055
AU 26	jaw drop	206,535

Aff-Wild2 is split into training, validation and testing sets, in a subject independent manner. These sets consist of 295, 105 and 141 videos, respectively.

2.4. Multi-Task-Learning Challenge

For this Challenge's corpora, we have created a static version of the Aff-Wild2 database, named s-Aff-Wild2. s-Aff-Wild2 contains selected-specific frames/images from Aff-Wild2. In total, 220,583 images are used that contain annotations in terms of valence-arousal; 6 basic expressions, plus the neutral state, plus the 'other' category; 12 action units (as described in the previous subsections).

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

At first, all videos are splitted into independent frames. Then they are passed through the RetinaFace detector [10] so as to extract, for each frame, face bounding boxes and 5 facial landmarks. The images were cropped according the bounding box locations; then the images were provided to the participating teams. The 5 facial landmarks (two eyes, nose and two mouth corners) were 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.

All cropped and cropped-aligned images were resized to $112 \times 112 \times 3$ pixel resolution and their intensity values were normalized to $[-1, 1]$.

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 performance measure is the average between the Concordance Correlation Coefficient (CCC) of valence and arousal. 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. CCC takes values in the range $[-1, 1]$; high values are desired. CCC is defined as follows:

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

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.

Therefore, the evaluation criterion for the Valence-Arousal Estimation Challenge is:

$$\mathcal{P}_{VA} = \frac{\rho_a + \rho_v}{2} \tag{2}$$

3.2. Expression Classification Challenge

The performance measure is the average F1 Score across all 8 categories (i.e., macro F1 Score). 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 takes values in the range $[0, 1]$; high values are desired. The F_1 score is defined as:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

Therefore, the evaluation criterion for the Expression Classification Challenge is:

$$\mathcal{P}_{EXPR} = \frac{\sum_{expr} F_1^{expr}}{8} \tag{4}$$

3.3. Action Unit Detection Challenge

The performance measure is the average F1 Score across all 12 AUs (i.e., macro F1 Score). Therefore, the evaluation criterion for the Action Unit Detection Challenge is:

$$\mathcal{P}_{AU} = \frac{\sum_{au} F_1^{au}}{12} \tag{5}$$

3.4. Multi-Task-Learning Challenge

The performance measure is the sum of: the average CCC of valence and arousal; the average F1 Score of the 8 expression categories; the average F1 Score of the 12 action units (as defined above). Therefore, the evaluation criterion for the Multi-Task-Learning Challenge is:

$$\begin{aligned} \mathcal{P}_{MTL} &= \mathcal{P}_{VA} + \mathcal{P}_{EXPR} + \mathcal{P}_{AU} \\ &= \frac{\rho_a + \rho_v}{2} + \frac{\sum_{expr} F_1^{expr}}{8} + \frac{\sum_{au} F_1^{au}}{12} \end{aligned} \tag{6}$$

4. Baseline Networks & Participating Teams' Methods and Results

All baseline systems rely exclusively on existing open-source machine learning toolkits to ensure the reproducibility of the results. All systems have been implemented in TensorFlow; training time was around six hours on a Titan X GPU, with a learning rate of 10^{-4} and with a batch size of 256.

In this Section, we present the top-performing teams per Challenge and we also describe the baseline systems developed for each Challenge; finally we report their obtained results, also declaring the winners of each Challenge.

4.1. Valence-Arousal Estimation Challenge

In total, 33 Teams participated in the VA Estimation Challenge. 16 Teams submitted their results. 7 Teams scored higher than the baseline.

The winner of this Challenge is: Situ-RUCAIM3 consisting of: Chuanhe Liu, Liyu Meng, Xiaolong Liu, Yuchen Liu, Zhaopei Huang, Meng Wang, Yuan Cheng, Qin Jin (Beijing Seek Truth Data Technology Services Co Ltd).

The runner up is: FlyingPigs (that also participated last year in the 2nd ABAW Competition) consisting of: Cuntai Guan, Ruyi An, Yi Ding, Su Zhang (Nanyang Technological University).

In the third place is: PRL consisting of: Soo-Hyung Kim, Hong-Hai Nguyen, Van-Thong Huynh (Chonnam National University).

Baseline Network The baseline network is a ResNet with 50 layers, pre-trained on ImageNet (ResNet50) and with a (linear) output layer that gives final estimates for valence and arousal.

Table 3 presents the leaderboard and results of the participating teams' algorithms that scored higher than the baseline in the Valence-Arousal Estimation Challenge. Table 3 illustrates the CCC evaluation of valence and arousal predictions on the Aff-Wild2 test set; it further shows the baseline network results (ResNet50). For reproducibility reasons, a link to a Github repository for each participating team's methodology exists and can be found in the corresponding leaderboard published in the official 3rd ABAW Competition's website [3]. It can be observed that Situ-RUCAIM3's method achieved the overall best performance (evaluation criterion is the mean CCC of valence and arousal) and the best performance in valence estimation. The FlyingPigs' method although ranked second in overall performance, it achieved the best performance in arousal estimation. Finally let us mention that the baseline network performance on the validation set is: 0.31 for valence and 0.17 for arousal (in terms of CCC).

4.2. Expression Classification Challenge

In total, 30 Teams participated in the Expression Classification Challenge. 14 Teams submitted their results. 7 Teams scored higher than the baseline.

The winner of this Challenge is: Netease Fuxi Virtual Human (that also participated last year in the 2nd ABAW Competition and was the winner of the respective Challenge) consisting of: Wei Zhang, Feng Qiu, Hao Zeng, Suzhen Wang, Zhimeng Zhang, Bowen Ma, Rudong An, Yu Ding (Netease Fuxi AI Lab).

[3] https://ibug.doc.ic.ac.uk/resources/cvpr-2022-3rd-abaw/

Table 3. Valence-Arousal Estimation Challenge Results: the evaluation criterion is the average CCC; CCC is displayed in %; the best performing submission is indicated in bold

Teams	CCC-V	CCC-A
Situ-RUCAIM3 [43]	56.05	51.65
	57.79	57.81
	60.6	**59.6**
	58.98	60.18
	59.29	59.85
FlyingPigs [63]	49.03	57.33
	49.03	58.42
	52.00	**60.16**
	47.94	58.00
	51.11	58.65
PRL [44]	37.00	38.14
	45.00	**44.48**
	38.45	39.51
	19.55	17.55
	20.00	18.00
HSE-NN [51]	40.14	42.78
	40.83	43.89
	40.61	43.49
	41.74	**45.38**
	41.35	44.88
AU-NO [25]	37.60	37.99
	39.57	37.56
	32.70	35.87
	41.82	**40.66**
LIVIA-2022 [48]	37.17	36.24
	32.08	32.95
	35.61	29.21
	37.42	**36.33**
Netease Fuxi Virtual Human [66]	11.86	23.96
	30.05	**24.42**
	21.61	25.09
	27.13	22.61
	24.88	25.63
baseline [27]	18.00	17.00

The runner up is: IXLAB consisting of: Jin-Woo Jeong, Jae-Yeop Jeong, Yeong-Gi Hong, Daun Kim (Seoul National University of Science and Technology), Yuchul Jung (Kumoh National Institute of Technology).

In the third place is: AlphaAff consisting of: Fanglei Xue, Zhongsong Ma (University of Chinese Academy of Sciences), Zichang Tan, Yu Zhu, Guodong Guo (Baidu Research).

Baseline Network The baseline network is a VGG16 network with fixed (i.e., non-trainable) convolutional weights (only the 3 fully connected layers were trainable), pre-trained on the VGGFACE dataset and with an output layer equipped with softmax activation function which gives the 8 expression predictions.

Table 4 presents the leaderboard and results of the participating teams' algorithms that scored higher than the baseline in the Expression Classification Challenge. Table 4 illustrates the average F1 score evaluation of predictions on the Aff-Wild2 test set; it further shows the baseline network results (VGG16). For reproducibility reasons, a link to a Github repository for each participating team's methodology exists and can be found in the corresponding leaderboard published in the official 3rd ABAW Competition's website. Finally let us mention that the baseline network performance on the validation set is: 0.23 (in terms of average F1 score).

Table 4. Expression Classification Challenge Results: the evaluation criterion is the average F1 Score, which is displayed in %; the best performing submission is indicated in bold

Teams	F1
Netease Fuxi Virtual Human [66]	33.43
	28.46
	35.87
	26.73
	33.97
IXLAB [21]	30.64
	30.24
	33.77
AlphaAff [59]	31.38
	31.73
	32.17
	31.85
	31.79
HSE-NN [51]	29.26
	29.73
	29.64
	30.25
	30.07
PRL [47]	26.86
	26.73
	26.32
	28.6
dgu [26]	**27.2**
USTC-NELSLIP [61]	**21.91**
	21.69
	21.80
	21.31
	21.89
baseline [27]	20.50

4.3. Action Unit Detection Challenge

In total, 38 Teams participated in the Action Unit Detection Challenge. 19 Teams submitted their results. 8 Teams scored higher than the baseline.

The winner of this Challenge is: Netease Fuxi Virtual Human (that also participated last year in the 2nd ABAW Competition and was the winner of the respective Challenge) consisting of: Wei Zhang, Feng Qiu, Hao Zeng, Suzhen Wang, Zhimeng Zhang, Bowen Ma, Rudong An, Yu Ding (Netease Fuxi AI Lab).

The runner up (with a very small difference from the winning team -49.89 vs 49.82-) is: SituTech consisting of: Chuanhe Liu, Wenqiang Jiang, Liyu Meng, Yannan Wu, Fengsheng Qiao, Yuanyuan Deng (Beijing Seek Truth Data Technology Services Co Ltd).

In the third place is: PRL consisting of: Soo-Hyung Kim, Hong-Hai Nguyen, Van-Thong Huynh (Chonnam National University).

In the fourth place is: STAR-2022 consisting of: Lingfeng Wang, Shisen Wang, Jin Qi (University of Electronic Science and Technology of China).

Baseline Network The baseline network is a VGG16 network with fixed convolutional weights (only the 3 fully connected layers were trained), pre-trained on the VGGFACE dataset and with an output layer equipped with sigmoid activation function which gives the 12 action unit predictions.

Table 5 presents the leaderboard and results of the participating teams' algorithms that scored higher than the baseline in the Action Unit Detection Challenge. Table 5 illustrates the average F1 score evaluation of predictions on the Aff-Wild2 test set; it further shows the baseline network results (VGG16). For reproducibility reasons, a link to a Github repository for each participating team's methodology exists and can be found in the corresponding leaderboard published in the official 3rd ABAW Competition's website. It is worth mentioning that, in this Challenge, the difference in the methods' performance between the winner of the Challenge and the team that ranked in the second place is only 0.07%. In general, the difference in the methods' performance between the top-four performing teams is less than 1.1%. Finally let us mention that the baseline network performance on the validation set is: 0.39 (in terms of average F1 score).

Table 5. Action Unit Detection Challenge Results: the evaluation criterion is the average F1 Score, which is displayed in %; the best performing submission is indicated in bold

Teams	F1
Netease Fuxi Virtual Human [66]	49.30
	49.89
	17.75
	16.38
	16.90
SituTech [23]	49.56
	49.39
	44.65
	49.82
	48.73
PRL [44]	47.9
	48.77
	48.26
	48.26
	49.04
STAR-2022 [55]	**48.83**
HSE-NN [51]	46.60
	47.18
	47.05
	47.13
	47.31
ISIR_DL [52]	44
	44.32
	42.09
	42.88
	44.30
SCPRLab@CNU [19]	**42.06**
	42.06
	42.06
USTC-AC [56]	**41.57**
	41.13
	41.39
	40.63
baseline [27]	36.50

4.4. Multi-Task-Learning Challenge

In total, 28 Teams participated in the Multi-Task Learning Challenge. 12 Teams submitted their results. 4 Teams scored higher than the baseline.

The winner of this Challenge is: NISL-2022 (that has participated both in the 1st and 2nd ABAW Competitions and has been the winner of multiple Challenges) consisting of: Didan Deng, Bertram Emil Shi (Hong Kong University of Science and Technology).

The runner up is: IMLAB consisting of: Sejoon Lim, Geesung Oh, Euiseok Jeong (Kookmin University).

In the third place is: HSE-NN consisting of: Andrey Savchenko (HSE University).

Baseline Network The baseline network is a VGG16 network with with fixed convolutional weights (only the 3 fully connected layers were trained), pre-trained on the

VGGFACE dataset. The output layer consists of 22 units: 2 linear units that give the valence and arousal predictions; 8 units equipped with softmax activation function that give the expression predictions; 12 units equipped with sigmoid activation function that give the action unit predictions.

Table 6 presents the leaderboard and results of the participating teams' algorithms that scored higher than the baseline in the Multi-Task Learning Challenge. Table 6 illustrates the evaluation of predictions on the Aff-Wild2 test set (in terms of the sum of the average CCC between valence-arousal, the average F1 score of the expression classes and the average F1 score of the action units); it further shows the baseline network results (VGG16). For reproducibility reasons, a link to a Github repository for each participating team's methodology exists and can be found in the corresponding leaderboard published in the official 3rd ABAW Competition's website. Finally let us mention that the baseline network performance on the validation set is: 0.30 (in terms of average F1 score).

Table 6. Multi-Task Learning Challenge Results: the evaluation criterion is the sum of the evaluation criteria of each task, which is displayed in %; the best performing submission is indicated in bold

Teams	Overall Metric
NISL 2022 [6]	110.38 **113.28**
IMLAB [20]	91.21 80.72 83.52 **95.31**
HSE-NN [51]	79.34 **80.90** 80.83 78.72
Netease Fuxi Virtual Human [66]	56.62 59.47 56.84 61.50 **67.50**
baseline [27]	28.00

5. Conclusion

In this paper we have presented the third Affective Behavior Analysis in-the-wild Competition (ABAW) 2022 held in conjunction with IEEE CVPR 2022. This Competition is a continuation of the first and second ABAW Competitions held in conjunction with IEEE FG 2020 and ICCV 2021, respectively. This Competition comprises four Challenges targeting: i) uni-task valence-arousal estimation, ii) uni-task expression classification (8 categories), iii) uni-task action unit detection (12 action units) and iv) multi-task-learning. The database utilized for this Competition has been derived from the Aff-Wild2, the first and large-scale database annotated for all these three behavior tasks.

The third ABAW Competition has been a very successful one with the participation of 33 Teams in the Valence-Arousal Estimation Challenge, 30 Teams in the Expression Classification Challenge, 38 Teams in the Action Unit Detection Challenge and 28 Teams in the Multi-Task Learning Challenge. All teams' solutions were very interesting and creative, providing quite a push from the developed baselines.

References

[1] Panagiotis Antoniadis, Ioannis Pikoulis, Panagiotis P Filntisis, and Petros Maragos. An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. *arXiv preprint arXiv:2107.03465*, 2021.

[2] 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.

[3] 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.

[4] 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.

[5] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[6] Didan Deng. Multiple emotion descriptors estimation at the abaw3 challenge. *arXiv preprint arXiv:2203.12845*, 2022.

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

[8] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020.

[9] Didan Deng, Liang Wu, and Bertram E Shi. Towards better uncertainty: Iterative training of efficient networks for multitask emotion recognition. *arXiv preprint arXiv:2108.04228*, 2021.

[10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multilevel face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020.

[11] Nhu-Tai Do, Tram-Tran Nguyen-Quynh, and Soo-Hyung Kim. Affective expression analysis in-the-wild using multi-task temporal statistical deep learning model. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 624–628. IEEE, 2020.

[12] Denis Dresvyanskiy, Elena Ryumina, Heysem Kaya, Maxim Markitantov, Alexey Karpov, and Wolfgang Minker. An audio-video deep and transfer learning framework for multimodal emotion recognition in the wild. *arXiv preprint arXiv:2010.03692*, 2020.

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

[14] Nico H Frijda et al. *The emotions*. Cambridge University Press, 1986.

[15] Darshan Gera and S Balasubramanian. Affect expression behaviour analysis in the wild using spatio-channel attention and complementary context information. *arXiv preprint arXiv:2009.14440*, 2020.

[16] Darshan Gera and S Balasubramanian. Affect expression behaviour analysis in the wild using consensual collaborative training. *arXiv preprint arXiv:2107.05736*, 2021.

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[18] 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.

[19] Duy Le Hoai, Eunchae Lim, Eunbin Choi, Sieun Kim, Sudarshan Pant, Guee-Sang Lee, Soo-Huyng Kim, and Hyung-Jeong Yang. An attention-based method for action unit detection at the 3rd abaw competition. *arXiv preprint arXiv:2203.12428*, 2022.

[20] Euiseok Jeong, Geesung Oh, and Sejoon Lim. Multitask emotion recognition model with knowledge distillation and task discriminator. *arXiv preprint arXiv:2203.13072*, 2022.

[21] Jae-Yeop Jeong, Yeong-Gi Hong, Daun Kim, Yuchul Jung, and Jin-Woo Jeong. Facial expression recognition based on multi-head cross attention network. *arXiv preprint arXiv:2203.13235*, 2022.

[22] Xianpeng Ji, Yu Ding, Lincheng Li, Yu Chen, and Changjie Fan. Multi-label relation modeling in facial action units detection. *arXiv preprint arXiv:2002.01105*, 2020.

[23] Wenqiang Jiang, Yannan Wu, Fengsheng Qiao, Liyu Meng, Yuanyuan Deng, and Chuanhe Liu. Facial action unit recognition with multi-models ensembling. *arXiv preprint arXiv:2203.13046*, 2022.

[24] 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.

[25] Vincent Karas, Mani Kumar Tellamekala, Adria Mallol-Ragolta, Michel Valstar, and Björn W Schuller. Continuous-time audiovisual fusion with recurrence vs. attention for in-the-wild affect recognition. *arXiv preprint arXiv:2203.13285*, 2022.

[26] Jun-Hwa Kim, Namho Kim, and Chee Sun Won. Facial expression recognition with swin transformer. *arXiv preprint arXiv:2203.13472*, 2022.

[27] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022.

[28] 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.

[29] 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, 2020.

[30] 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.

[31] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.

[32] 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.

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

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

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

[36] 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.

[37] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021.

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

[39] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[40] I Li et al. Technical report for valence-arousal estimation on affwild2 dataset. *arXiv preprint arXiv:2105.01502*, 2021.

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

[42] Shuyi Mao, Xinqi Fan, and Xiaojiang Peng. Spatial and temporal networks for facial expression recognition in the wild videos. *arXiv preprint arXiv:2107.05160*, 2021.

[43] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Wenqiang Jiang, Tenggan Zhang, Yuanyuan Deng, Ruichen Li, Yannan Wu, Jinming Zhao, et al. Multi-modal emotion estimation for in-the-wild videos. *arXiv preprint arXiv:2203.13032*, 2022.

[44] Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. An ensemble approach for facial expression analysis in video. *arXiv preprint arXiv:2203.12891*, 2022.

[45] Geesung Oh, Euiseok Jeong, and Sejoon Lim. Causal affect prediction model using a facial image sequence. *arXiv preprint arXiv:2107.03886*, 2021.

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

[47] Kim Ngan Phan, Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. Expression classification using concatenation of deep neural network for the 3rd abaw3 competition. *arXiv preprint arXiv:2203.12899*, 2022.

[48] Gnana Praveen Rajasekar, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro Koerich, Patrick Cardinal, and Eric Granger. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. *arXiv preprint arXiv:2203.14779*, 2022.

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

[50] Junya Saito, Xiaoyu Mi, Akiyoshi Uchida, Sachihiro Youoku, Takahisa Yamamoto, and Kentaro Murase. Action units recognition using improved pairwise deep architecture. *arXiv preprint arXiv:2107.03143*, 2021.

[51] Andrey V Savchenko. Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices. *arXiv preprint arXiv:2203.13436*, 2022.

[52] Gauthier Tallec, Edouard Yvinec, Arnaud Dapogny, and Kevin Bailly. Multi-label transformer for action unit detection. *arXiv preprint arXiv:2203.12531*, 2022.

[53] Manh Tu Vu and Marie Beurton-Aimar. Multitask multi-database emotion recognition. *arXiv preprint arXiv:2107.04127*, 2021.

[54] Lingfeng Wang and Shisen Wang. A multi-task mean teacher for semi-supervised facial affective behavior analysis. *arXiv preprint arXiv:2107.04225*, 2021.

[55] Lingfeng Wang, Shisen Wang, and Jin Qi. Multi-modal multi-label facial action unit detection with transformer. *arXiv preprint arXiv:2203.13301*, 2022.

[56] Shangfei Wang, Yanan Chang, and Jiahe Wang. Facial action unit recognition based on transfer learning. *arXiv preprint arXiv:2203.14694*, 2022.

[57] 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.

[58] Hong-Xia Xie, I Li, Ling Lo, Hong-Han Shuai, Wen-Huang Cheng, et al. Technical report for valence-arousal estimation in abaw2 challenge. *arXiv preprint arXiv:2107.03891*, 2021.

[59] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. *arXiv preprint arXiv:2203.13052*, 2022.

[60] Sachihiro Youoku, Yuushi Toyoda, Takahisa Yamamoto, Junya Saito, Ryosuke Kawamura, Xiaoyu Mi, and Kentaro Murase. A multi-term and multi-task analyzing framework for affective analysis in-the-wild. *arXiv preprint arXiv:2009.13885*, 2020.

[61] Jun Yu, Zhongpeng Cai, Peng He, Guocheng Xie, and Qiang Ling. Multi-model ensemble learning method for human expression recognition. *arXiv preprint arXiv:2203.14466*, 2022.

[62] 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.

[63] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3. *arXiv preprint arXiv:2203.13031*, 2022.

[64] Su Zhang, Yi Ding, Ziquan Wei, and Cuntai Guan. Audio-visual attentive fusion for continuous emotion recognition. *arXiv preprint arXiv:2107.01175*, 2021.

[65] 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.

[66] Wei Zhang, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. *arXiv preprint arXiv:2203.12367*, 2022.

[67] 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.