Video-based Frame-level Facial Analysis of Affective Behavior on Mobile Devices using EfficientNets

Andrey V. Savchenko
HSE University
Laboratory of Algorithms and Technologies for Network Analysis, Nizhny Novgorod, Russia
avsavchenko@hse.ru

Abstract

In this paper, we consider the problem of real-time video-based facial emotion analytics, namely, facial expression recognition, prediction of valence and arousal and detection of action unit points. We propose the novel frame-level emotion recognition algorithm by extracting facial features with the single EfficientNet model pre-trained on AffectNet. The predictions for sequential frames are smoothed using mean or median filters. It is demonstrated that our approach may be implemented even for video analytics on mobile devices. Experimental results for the large scale AffWild2 database from the third Affective Behavior Analysis in-the-wild Competition demonstrate that our simple model is significantly better when compared to the VggFace baseline. In particular, our method is characterized by 0.1-0.5 higher performance measures for test sets in the uni-task Expression Classification, Valence-Arousal Estimation, Action Unit Detection and Multi-Task Learning. Our team took the 3rd place in the multi-task learning challenge and 4th places in Valence-Arousal and Expression challenges. Due to simplicity, the proposed approach may be considered as a new baseline for all four sub-challenges.

1. Introduction

Affective computing, emotional intelligence [20] and, in particular, analysis of humans’ emotional states based on facial videos, is an essential task for many systems with man-machine interaction [4], education, health [28] and mobile services [2,7,25]. Many facial analysis tasks, such as face recognition, age and gender prediction, have reached high accuracy appropriate for many practical applications [1,23]. However, but the ability to understand human emotions is still far from maturity [8]. The personal bias and backgrounds increase the uncertainty of emotion perception and contextual information [4]. As a result, the datasets used to train FER (facial expression recognition) models are not very large and contain a lot of noise and inconsistencies in emotional labels of photos [19]. The video-based FER is even more complex task, because human emotions may change rapidly, and many frames do not contain enough information to reliably predict facial (macro) expression. Hence, the authors of the datasets are required to provide the labeling at frame level [22]. Thus, the number of video datasets for in-the-wild affective computing has been very limited.

The situation has changed with an appearance of the AffWild dataset [13, 30]. It has been recently extended in the AffWild2 database [14] with more videos and annotations for the following tasks: (1) frame-level FER; (2) detection of action units (AU), i.e., specific movements of facial muscles from Facial Action Coding System (FACS) [5]; and (3) prediction of valence and arousal, i.e., how active or passive, positive or negative is the human behavior.

Though FER has been a topic of major research [15], many models learn too many features specific for a concrete dataset, which is not practical for in-the-wild settings [8]. The development of in-the-wild affect prediction engines has been accelerated by a couple of ABAW (Affective Behavior Analysis in-the-wild) competitions [10, 16]. The third place in the first and second tasks was achieved by the authors of the paper [28] who proposed the multi-task learning (MTL) technique for the incomplete labels of these correlated tasks. The multi-modal audiovisual ensemble model [8] took the second place, while the winner of these two sub-challenges was a multi-task streaming network [32]. The latter captures identity-invariant emotional features using an advanced facial embedding. The valence-arousal challenge was won by deep ensembles with iterative distillation and pseudo-labeling [4].

As one can notice, most successful previous solutions use MTL [12, 32] to boost their performance. As a result, the authors of the third ABAW contest [9] decided to inspire researchers studying not only MTL, but also the uni-task models. The baseline uses the deep convolutional neural network (CNN), namely, VGG16, pre-trained on the VG-
GFACE dataset to make a decision in all tasks independently [9].

In this paper, we discuss our solution for all four tasks from the ABAW3 challenge. Most participants of such contests are mainly focused on improvement of accuracy metrics, so that they usually develop complex ensemble models [4, 32]. Hence, they cannot be implemented in real-time analysis of affective behavior in mobile or embedded systems. Hence, the main motivation of this paper is to develop a single [11] and lightweight model [24] that not only achieve high accuracy but may be used in mobile applications [7]. As a result, we contributed the novel model based on the EfficientNet architecture [27] that is much better than the baseline in terms of both size and performance. The weights of our CNN are tuned on external AffectNet dataset [19], so the facial embeddings extracted by this neural network do not learn any features that are specific to the Aff-Wild2 dataset. Thus, our method may become a new baseline for future challenges with the ABAW challenges.

This paper is structured as follows. Section 2 introduces our efficient model and the training procedure. Experimental results for all tasks of ABAW challenge are presented in Section 3. Finally, concluding comments and future work are discussed in Section 4.

2. Proposed approach

Three frame-level affective behavior analysis tasks [9] are considered for an input video \( X(t) \), \( t = 1, 2, ..., T \) with \( T \) frames, namely:

1. Facial expression recognition, in which it is required to assign each frame \( X(t) \) to one of \( C_{EXPR} > 1 \) categories (classes), such as happiness, fear, etc. It is a general multi-class classification problem.

2. AU analysis and recognition, in which the frame \( X(t) \) is associated with a subset of \( C_{AU} \geq 1 \) AUs. The task may be considered as a multi-label classification problem, i.e., prediction of a binary vector \( \textbf{AU}(t) = [AU_1(t), ..., AU_{C_{AU}}(t)] \). Here, \( AU_i(t) = 1 \) if the \( i \)-th action unit is detected in the \( t \)-th frame, otherwise \( AU_i(t) = 0 \).

3. Prediction of valence and arousal of an emotion. It is a type of regression tasks, but the values of valence and arousal are typically limited to a range \([-1, 1]\).

It is assumed that the facial regions are preliminarily extracted, so that \( X(t) \) contain cropped faces. The supervised learning scenario is considered, where a training set of \( N \) reference facial images \( \{X_n\}, n \in \{1, ..., N\} \) are given, and the facial expression \( e_n \in \{1, ..., C_{EXPR}\}, C_{AU}- \) dimensional binary vector \( \textbf{AU}_n \) of action units, valence \( V_n \), arousal \( A_n \) of the \( n \)-th reference image are known, though it is possible that some labels are unavailable.

In this paper, we will use conventional approach based on pre-training of deep CNN using large FER dataset. As we pay special attention to offline recognition on mobile devices [26], it is reasonable to use such architectures as MobileNets or EfficientNets [27]. The presented approach consists of the following steps:

1. Pre-training of a lightweight model on face identification task using very large facial dataset of celebrities [1].

2. Fine-tuning the model from item (1) on static photos from external dataset to obtain an emotional CNN [19].

3. The outputs of the emotional CNN (embeddings and expression scores) from item (2) are used to extract facial features of each video frame from the AffWild2 dataset [14].

4. These embeddings and scores are used to train simple frame-level MLP-based classification/regression models given the training set of each challenge.

5. Optional post-processing of frame-level outputs on models from item (4) computed for validation and test sets to make the predictions more smooth.

Let us consider the details of our approach. At first, a large external VGGFace2 facial dataset [1] with 9131 subjects is used to pre-train a CNN on face recognition task. The faces cropped by MTCNN (multi-task cascaded neural

Figure 1. Proposed model for the multi-task learning.
network) [31] detector without any margins were utilized for training, so that most parts of the background, hairs, etc. is not presented. As a result, the learned facial features are more suitable for emotional analysis. We trained the model totally of 8 epochs by the Adam optimizer and SAM (Sharpness-Aware Minimization) [6]. The models with the highest accuracy on validation set, namely, 92.1%, 94.19% and 95.49% for MobileNet-v1, EfficientNet-B0 and EfficientNet-B2, respectively, were used.

Second, the resulted CNN is fine-tuned on the training set of 287,651 photos from the AffectNet dataset [19] annotated with $C = 8$ basic expressions (Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness and Surprise). It is necessary to emphasize that the annotations of valence and arousal from the AffectNet dataset were not used in the pre-training. The last layer of the network pre-trained on VGGFace2 is replaced by the new head (fully-connected layer with $C$ outputs and softmax activation), so that the penultimate layer with $D$ neurons can be considered as an extractor of facial features. The weighted categorical cross-entropy (softmax) loss was optimized [19]. The new head was trained during 3 epochs with learning rate 0.001. Finally, the whole network is fine-tuned with a learning rate of 0.0001 at the last 5 epochs. The details of this training procedure are available in [24]. As a result, we fine-tuned three models, namely, MobileNet-v1, EfficientNet-B0 and EfficientNet-B2, that reached accuracy 60.71%, 61.32% and 63.03%, on the validation part of the AffectNet.

Third, such an emotional CNN was used as a feature extractor for frames $X(t)$ and reference images $X_n$. Though the cropped facial images provided by the organizers of the challenge have different (typically, low) resolution, they were resized to 224x224 pixels for the first two models, while the latter CNN requires input images with resolution 300x300. We examine two types of features: (1) facial image embeddings (output of penultimate layer) [24, 29]; and (2) scores (predictions of emotional class probabilities at the output of last softmax layer). As a result, $D$-dimensional embeddings $x(t)$ and $x_n$ and $C$-dimensional scores $s(t)$ and $s_n$ are obtained. Three kinds of features have been examined, namely: (1) embeddings only; (2) scores only; and (3) concatenation of embeddings and scores [21]. According to the rules of the uni-task challenges, the pre-trained model can be pre-trained on any task (e.g., VA estimation, Expression Classification, AU detection, Face Recognition), so that the expression scores returned by our model trained on the AffectNet can be used as facial features to predict Valence/Arousal and AUs. When we refined the model given the ABAW3 dataset, only the annotations available for a concrete challenge have been used to train a classification and regression models.

Fourth, we trained a shallow feed-forward neural network, such as multi-class logistic regression or MLP (multi-

![Figure 2. Sample screen of Android demo application.](image)
detected in a couple of frames, we perform the bilinear interpolation of the outputs of the model for two frames with detected faces. If face detection fails for several first or last frames of the video, we will simply use predictions for the closest frame with detected face.

Fifth, it is possible to smooth the predictions for \( k \geq 1 \) consecutive frames by using point-wise mean (box) or median filter with kernel size \( k \) [21]. If \( k \) is equal to 1, the frame-level predictions will be used. Otherwise, the slicing window with size \( k \) is processed for every \( t\)-th frame, i.e., we took the predictions at the output of our MLP classifiers for frames \( t - \frac{k}{2}, t - \frac{k}{2} + 1, ..., t - 1, t + 1, ..., t + \frac{k}{2} - 1, t + \frac{k}{2} \). The final decision function for the frame \( t \) is computed as a point-wise mean or median of these \( k \) predictions.

The training script for the presented approach is made publicly available\(^1\). The CNNs used for feature extraction, namely, MobileNet v1 (TensorFlow’s mobilenet_7.h5) and EfficientNets (PyTorch’s enet_b0_best_vgaf and enet_b2_8), are also available in this repository\(^2\). Finally, the possibility to use our model for mobile devices is demonstrated. The sample output of the demo Android application is presented in Fig. 2. It is possible to recognize facial expressions of all subjects in either any photo from the gallery or the video captured from the frontal camera.

### 3. Experimental results

In this section, four tasks from the third ABAW challenge are considered. We used the cropped images officially provided by the organizers of this challenge.

#### 3.1. Uni-task Expression Recognition

The first experiment is devoted to the uni-task FER with \( C_{EXPR} = 8 \) classes (anger, disgust, fear, happiness, sadness, surprise, neutral and other). The frame files missed in the cropped directory were ignored. As a result, the training and validation sets contains 585,317 and 280,532 files, respectively. Two performance metrics were computed, namely, (1) macro-averaged F1 score \( \bar{P}_{EXPR} \) [9]; and (2) top-1 unbalanced accuracy. The ablation study for several feature extractors and classifiers is presented in Table 1. Here, the absence of prefix “1 hidden layer” stands for the neural network without hidden layers. First, embeddings are classified more accurately when compared to emotional scores. Second, though EfficientNet-B2 has 2% greater accuracy than EfficientNet-B0 on the validation part of AffectNet [24], the latter model provides much better performance in this challenge. As a result, our best mean F1-score is 0.16 higher than \( \bar{P}_{EXPR} \) of the baseline VGGFACE provided by organizers of this challenge. However, even we used the weighted cross-entropy as a loss function, the imbalance of the dataset still influences the overall quality. Table 2 demonstrates the F1 scores of our best model for each class. As one can notice, the quality for anger and, especially, fear emotions is very low.

### 3.2. Uni-task Action Unit Detection

In the second experiment, we examine the uni-task Action Unit Detection problem. The training set contains

---

\(^1\)https://github.com/HSE-asavchenko/face-emotion-recognition/blob/main/src/abaw_train.ipynb

\(^2\)https://github.com/HSE-asavchenko/face-emotion-recognition/tree/main/models/affectnet_emotions

### Table 1. Expression Challenge Results on the Aff-Wild2’s validation set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>F1-score ( \bar{P}_{EXPR} )</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGGFACE Baseline [9]</td>
<td>embeddings</td>
<td>0.23</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>embeddings, 1 hidden layer</td>
<td>0.338</td>
<td>0.460</td>
</tr>
<tr>
<td>MobileNet [2]</td>
<td>scores</td>
<td>0.236</td>
<td>0.435</td>
</tr>
<tr>
<td></td>
<td>scores, 1 hidden layer</td>
<td>0.286</td>
<td>0.433</td>
</tr>
<tr>
<td>Our EfficientNet-B0 [24] embeddings, 1 hidden layer</td>
<td>0.307</td>
<td>0.428</td>
<td></td>
</tr>
<tr>
<td>Our EfficientNet-B2 embeddings, 1 hidden layer</td>
<td>0.305</td>
<td>0.412</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.317</td>
<td>0.435</td>
</tr>
</tbody>
</table>

### Table 2. F1 scores for FER with the best EfficientNet-B0 model.

<table>
<thead>
<tr>
<th>Expression</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>0.609</td>
</tr>
<tr>
<td>Anger</td>
<td>0.151</td>
</tr>
<tr>
<td>Disgust</td>
<td>0.516</td>
</tr>
<tr>
<td>Fear</td>
<td>0.016</td>
</tr>
<tr>
<td>Happiness</td>
<td>0.477</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.461</td>
</tr>
<tr>
<td>Surprise</td>
<td>0.303</td>
</tr>
<tr>
<td>Other</td>
<td>0.512</td>
</tr>
</tbody>
</table>
1,356,861 images and $C_{AU} = 12$ action units (AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25 and AU26), while 445836 facial frames are included into the validation set. The unweighted average F1-score $P_{AU}$ of our models (Table 3) is again up to 0.14 points greater than the baseline performance. This table contains the results when the threshold for each AU is equal to 0.5, and for different thresholds tuned for each AU separately. In the latter case, the following 12 thresholds were automatically found using the validation set: 0.8, 0.8, 0.7, 0.5, 0.5, 0.5, 0.6, 0.8, 0.8, 0.8, 0.3, and 0.7. The results are very similar to the first experiment: embeddings are classified more accurately, and EfficientNet-B0 is the best model.

### 3.3. Uni-task Valence-Arousal Prediction

In the third experiment, the uni-task Valence-Arousal Estimation is analyzed. The number of labeled images here is much higher, so that 1,555,919 and 338,755 frames were put into the training and validation sets. The estimates of CCC for valence and arousal together with their mean $P_{VA}$ are shown in Table 4. As one can notice, EfficientNet-B0 is still the best model, which has twice-higher $P_{VA}$ when compared to the baseline [9]. In contrast to the previous experiment, the greatest CCC is achieved by very fast regression models for $C = 8$ emotional scores of the AffectNet dataset. Conventional usage of embeddings leads to 0.02-0.09 lower CCCs.

### 3.4. Aggregation of Frame-Level Predictions

Next, we studied the impact of smoothing on the performance measures for EfficientNet-B0 and the best classifiers from the previous experiment. Five submissions have been prepared for AU, EXPR and VA challenges by using the frame-level predictions without smoothing and box (mean) and median filters with kernel sizes $k = 5$ and $k = 15$. The performance measures on the validation and test sets for each challenge are shown in Table 5. The best results are obtained by using the large slicing window ($k = 15$ frames). The mean filter is in most cases better than the median filter except the AU challenge. The proposed approach is much more accurate than the baseline on both validation and test sets. For example, our model has 10% greater F1-scores for the test set from the Expression and Action Unit Challenges. The most impressive is the increase of the CCC metric in the Valence-Arousal Challenge where we improved the baseline by more than 20%.

### 3.5. Multi-Task Learning

In the last experiment, we trained a complete multi-task model (Fig. 1) on the set of 142,225 images and computed the metric $P_{MTL} = P_{EXPR} + P_{VA} + P_{AU}$ using 26,876 validation frames. The results are reported in Table 6. It is important to notice two differences with previous experiments: (1) the best model here is EfficientNet-B2, and (2)
Table 5. Results of smoothing techniques on the Aff-Wild2’s validation and test sets.

<table>
<thead>
<tr>
<th>Model</th>
<th>Smoothing</th>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$P_{EXPR}$</td>
<td>$P_{AU}$</td>
</tr>
<tr>
<td>Baseline [9]</td>
<td>-</td>
<td>0.23</td>
<td>0.39</td>
</tr>
<tr>
<td>Proposed model</td>
<td>Frame-level ($k = 0$)</td>
<td>0.3807</td>
<td>0.5367</td>
</tr>
<tr>
<td></td>
<td>Mean ($k = 5$)</td>
<td>0.3914</td>
<td>0.5447</td>
</tr>
<tr>
<td></td>
<td>Median ($k = 5$)</td>
<td>0.3888</td>
<td>0.5430</td>
</tr>
<tr>
<td></td>
<td>Mean ($k = 15$)</td>
<td>0.4018</td>
<td>0.5445</td>
</tr>
<tr>
<td></td>
<td>Median ($k = 15$)</td>
<td>0.3996</td>
<td>0.5478</td>
</tr>
</tbody>
</table>

Table 6. Multi-Task-Learning Challenge Results on the Aff-Wild2’s validation set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>$P_{MTL}$</th>
<th>$P_{EXPR}$</th>
<th>$P_{VA}$</th>
<th>$P_{AU}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGGFACE</td>
<td>Baseline [9]</td>
<td>0.30</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Our MobileNet [2]</td>
<td>embeddings + scores, 1 hidden layer</td>
<td>1.123</td>
<td>0.386</td>
<td>0.283</td>
<td>0.455</td>
</tr>
<tr>
<td>EfficientNet-B0 [24]</td>
<td>embeddings + scores, 1 hidden layer</td>
<td>1.121</td>
<td>0.381</td>
<td>0.272</td>
<td>0.469</td>
</tr>
<tr>
<td>Our EfficientNet-B2</td>
<td>embeddings + scores, different AU thresholds</td>
<td>1.147</td>
<td>0.384</td>
<td>0.302</td>
<td>0.461</td>
</tr>
<tr>
<td></td>
<td>embeddings + scores, 1 hidden layer</td>
<td>1.135</td>
<td>0.378</td>
<td>0.298</td>
<td>0.458</td>
</tr>
</tbody>
</table>

MLP with 1 hidden layer is worse than a very simple logistic regression. In fact, the latter does not use knowledge about multiple tasks and makes predictions for each task independently. During the challenge, we made four submissions by using two CNNs (EfficientNet-B0 and B2) and training the feed-forward neural network classifier on the training set and concatenation of the training and validation sets. Despite the superiority of EfficientNet-B2 on the validation set (Table 6), the best performance metric $P_{MTL} = 0.809$ is obtained for EfficientNet-B0, though the difference with EfficientNet-B2 ($P_{MTL} = 0.8083$) is not significant. Anyway, our simple model is much more accurate when compared to the baseline. Indeed, its metric $P_{MTL}$ is equal to 0.3 and 0.28 for validation and test set, respectively. Thus, we improved the performance by more than 50% and took the third place in MTL sub-challenge.

4. Conclusion

In this paper, we have presented the frame-level facial emotion analysis model (Fig. 1) based on concatenation of embeddings and scores at the output of EfficientNet. The latter CNN was carefully pre-trained on the VGGFace2 [1] and AffectNet [24] datasets. Its experimental study for the tasks of the third ABAW challenge [9] have demonstrated that our technique is much more accurate than the baseline VGGFACE. Moreover, we have implemented an Android mobile application (Fig. 2) with publicly available source code to demonstrate the real-time efficiency of our approach and motivate practitioners to implement the facial emotion analytic engines on-device. Our approach is based on a single lightweight neural network, so that it may be not as accurate as ensembles of many CNNs [4].

Our best EfficientNet-B0 model is characterized by F1-score 0.38 for expression recognition, mean CCC 0.46 for valence and arousal prediction, and F1-score 0.54 for action unit detection. Our approach took the third place in the MTL challenge, fourth places in the Expression and Valence-Arousal Challenges and fifth place in the Action Unit Challenges. In average, there is only one team who took slightly lower average place in all four tasks [33]. Their transformer-based multimodal solution is the winner in the uni-task Expression Classification and Action Unit Detection, but our method is better in other two challenges. As the proposed model has not been fine-tuned on the AffWild2 dataset, we can claim that the facial features extracted by our networks lead to the most robust decisions. Due to simplicity, our approach may be considered as a new baseline for all four sub-challenges.

In the future, it is necessary to integrate our approach into more complex pipelines. For example, we process frames independently, so that any sequential and attention models can benefit from the usage of our facial emotional features [3,17,18]. Moreover, it is worth studying the combination of our models into a large ensemble with different representations of input videos. Finally, as our current best model for MTL solves each task independently, it is impor-
tant to properly use the correlation between different tasks in the multi-task learning scenario [32].

Acknowledgements. The work is supported by RSF (Russian Science Foundation) grant 20-71-10010.

References


