SN Computer Science (2020) 1:321
https://doi.org/10.1007/s42979-020-00325-6

ORIGINAL RESEARCH

# The FaceChannel: A Fast and Furious Deep Neural Network for Facial Expression Recognition

**Pablo Barros**[1] · **Nikhil Churamani**[2] · **Alessandra Sciutti**[1]

Received: 18 June 2020 / Accepted: 8 September 2020
© The Author(s) 2020

**Abstract**
Current state-of-the-art models for automatic facial expression recognition (FER) are based on very deep neural networks that are effective but rather expensive to train. Given the dynamic conditions of FER, this characteristic hinders such models of been used as a general affect recognition. In this paper, we address this problem by formalizing the FaceChannel, a light-weight neural network that has much fewer parameters than common deep neural networks. We introduce an inhibitory layer that helps to shape the learning of facial features in the last layer of the network and, thus, improving performance while reducing the number of trainable parameters. To evaluate our model, we perform a series of experiments on different benchmark datasets and demonstrate how the FaceChannel achieves a comparable, if not better, performance to the current state-of-the-art in FER. Our experiments include cross-dataset analysis, to estimate how our model behaves on different affective recognition conditions. We conclude our paper with an analysis of how FaceChannel learns and adapts the learned facial features towards the different datasets.

**Keywords** Facial expression recognition · Convolutional neural network · Affective computing

## Introduction

Evidence shows that humans can perceive, recognize, and commonly understand a set of 'basic' emotions from facial expressions across cultures and around the world [1]. Adapting an automatic facial expression recognition (FER) system to achieve such a capability, however, is still an open a difficult task. One of the most critical characteristics that yet needs to be addressed is how each person expresses the basic emotions differently, most of the time by combining different basic concepts or even shortly transitioning between them [2]. Understanding the compositionality of affect helps us to understand better each other, and to derive a larger comprehension of affect than the ones posted by most of the current automatic facial expression recognition (FER) systems [3, 4].

Given that most of the solutions for automatic facial expression recognition (FER) are difficult to adapt, due to their constrict computational nature, the most common solutions for this problem are to formalize affect into a way that bounds the categorization of such systems [5–7]. Most of these systems, thus, are extremely restricted on what they can recognize as affect, given the availability of data to train them. But most importantly, the current state-of-the-art on facial expression recognition (FER) is deep neural networks with millions of parameters to tune. This implies they are extremely difficult to adapt to novel stimuli and/or affective labels [8].

Deep learning models usually learn how to represent affective features by updating filters based on a large number of data samples, using strongly supervised learning methods [9–14]. As a result, these models can extract facial features for a collection of different individuals, which contributes to their generalization of expression representations enabling a universal and automatic FER machine. The development of such models was supported by the collection of several "in-the-wild" datasets [15–18] that provided large amounts of well-labeled data. These datasets usually contain emotion expressions from various multimedia sources ranging from single frames to a few seconds of video material. Because

✉ Pablo Barros
pablo.alvesdebarros@iit.it

[1] Cognitive Architecture for Collaborative Technologies Unit, Istituto Italiano di Tecnologia, Genoa, Italy

[2] Department of Computer Science and Technology, University of Cambridge, Cambridge, UK

Published online: 06 October 2020

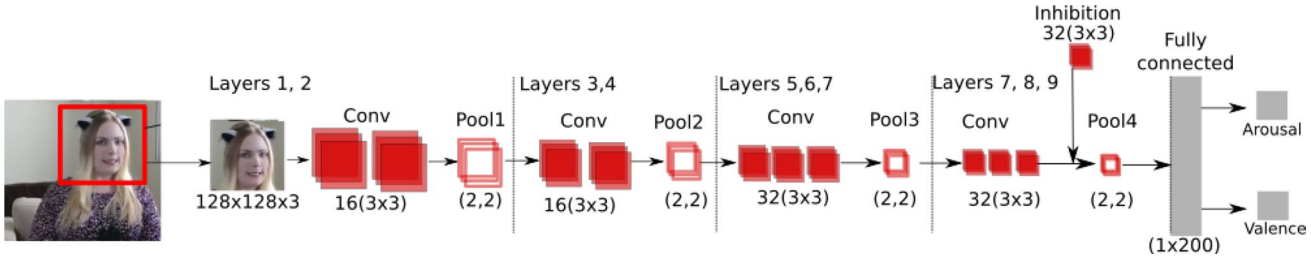


**Fig. 1** The FaceChannel and all the parameters and details of the neural architecture

of the availability of large amounts of training data, the performance of deep learning-based solutions forms the state-of-the-art in FER, benchmarked on these datasets [19–22].

Most of these models, however, employ large and deep neural networks that demand a high computational power for training and re-adapting [10, 23, 24]. As a result, these models specialize on recognizing emotion expressions under conditions represented in the datasets they are trained with [25, 26]. Thus, when these models are used to recognize facial expression under different conditions, not represented in the training data, they tend to perform poorly. This is the case on several social applications, such as human–robot interaction [27]. Retraining these networks to adapt to new application scenarios is usually the solution for adapting them. Yet, owing to the large and deep architecture of these models, retraining the entire network with changing conditions is rather expensive.

Furthermore, once trained, these deep neural models are relatively fast to recognize facial expressions when provided with rich computational resources. With reduced processing power, however, such as in robotic platforms, these models usually are extremely slow and do not support real-time application.

In this paper, we address the problem of very deep neural networks by formalizing the FaceChannel neural network. This model is an upgraded version of the Multi-Channel Convolution Neural Network, proposed in our previous work [28].

The FaceChannel is a light-weight convolution neural network, with around 2 million updatable parameters, that employs inhibitory layers trained from scratch. To evaluate our model, we perform a series of experiments on recognizing categorical and dimensional affect from different facial expressions datasets. We perform in-dataset and cross-dataset experiments, to evaluate in-depth the generalization of our model and its capability of adapting towards different scenarios. We also compare the performance of our model with current state-of-the-art deep neural networks for facial expression recognition, and demonstrate that the FaceChannel has a similar or better performance when compared to most of them, but with much less parameters to be updated.

We also provide a discussion on how the FaceChannel consumes low computational resources to both train and adapt towards the different datasets. Further, we provide an analysis on how the features of the convolution layers of the FaceChannel are affected by each datasets' specific characteristics, and how they are shared on the adapting scenarios to boost the network's performance.

The FaceChannel was initially presented in our previous workshop paper [29], and in this paper, we present an in-depth formalization of the model, extend the performance evaluations, including a novel set of cross-dataset assessments, and perform facial features learning analysis. Our goal with this paper is to complement our previous work with a detailed evaluation and understanding of the FaceChannel.

## The FaceChannel

The FaceChannel presented here is an updated version of our previous work on facial expression recognition [28]. We extend it by adapting the topology of the VGG16 model [30], but with much fewer parameters to be trained, as exhibited in Fig. 1. Our model has a total of 10 convolutional layers and 4 pooling layers. Batch normalization is used after each convolutional layer and a dropout of with a 50% chance is used after each pooling layer. Following our previous work, we adapted our last convolutional layer with shunting inhibitory connections [31] Each shunting neuron $S_{nc}^{xy}$ at the position ($x$, $y$) of the $n$th receptive field in the $c^{th}$ layer is activated as:

$$S_{nc}^{xy} = \frac{u_{nc}^{xy}}{a_{nc} + I_{nc}^{xy}}, \tag{1}$$

where $u_{nc}^{xy}$ is the activation of the convolution unit and $I_{nc}^{xy}$ is the activation of the inhibitory connections. A passive decay term, $a_{nc}$, which is learned, is shared among each shunting inhibitory connection. Each convolutional and inhibitory layer of the FaceChannel implements a ReLU activation function.

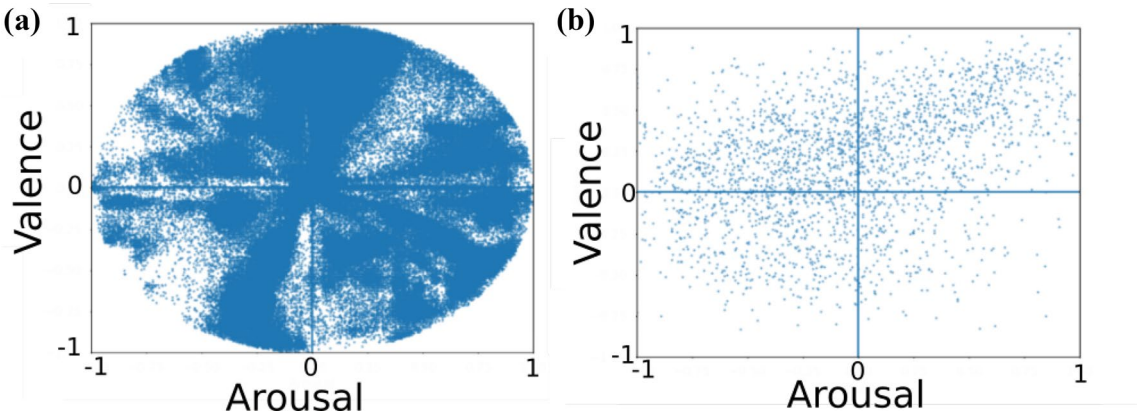


**Fig. 2** Annotation distributions for: **a** the AffectNet dataset [16] has a high variance on arousal and valence with a large number of data points and **b** the continuous expressions of the OMG-Emotion [12] videos cover high arousal and valence spread

Typically, in convolutional neural networks trained on images, the first layers learn how to represent and highlight edges, contours, and contrasts [32]. When trained to recognize faces, though, the last layers of such a network usually highlight facial characteristics, sometimes resembling even facial action units [33, 34]. The shunting neurons have a role of over-specifying the filters of the last layer. These neurons enhance the capability of the filters to extract strong and high-level facial representations, which improves the network capability of clustering database-specific features into the last layer [28]. This improves the network generalization, and makes it easier to be updated for a novel scenario, as and when needed.

The output of the convolutional layers is fed to a fully connected layer with 200 units, each one implementing a ReLU activation function, which is then fed to an output layer. Our model is trained using a categorical cross-entropy loss function.

As typical for most deep learning models, our FaceChannel has several hyper-parameters that need to be tuned. We optimized our model to maximize the recognition accuracy using a tree-structured Parzen estimator (TPE) [35] and use the optimal training parameters throughout all of our experiments. The entire network has around 2 million adaptable parameters, which makes it very light-weight as compared to commonly used VGG16-based networks.

## Experimental Setup

To evaluate the FaceChannel, we perform several intra- and inter-dataset experiments. As some of these datasets do not contain enough data to train a deep neural network properly, we repeat three evaluation routines for each dataset: (i) we train the model using the indicated experimental protocol given by each dataset; (ii) we pre-train the model using the AffectNet dataset [16] and evaluate it using the experimental protocol of each of the other datasets; and (iii) we pre-train the model using the AffectNet dataset, and fine-tune it using the training protocol for each of the evaluated datasets.

Our fine-tuning routine only trains the last fully connected layer of the FaceChannel. We performed empirical exploration experiments which demonstrated that re-training the entire network did not improve, and in some cases even decreased the networks' performance. Also, by fine-tuning only the last fully connected layer we guarantee that the facial features learned by the convolution layers are preserved, and only the decision-making on how to tune these features towards the labels of each dataset is affected.

Running these three experimental setups help us to investigate the capabilities of our model (a) to learn facial features from each specific dataset; (b) to learn general features from a vast number of examples from the AffectNet dataset; and (c) to adapt the learned features towards the individual specificity of each of the datasets.

### Datasets

#### AffectNet

The AffectNet [16] dataset consists of more than 1 million images obtained from web crawlers. Approximately half of them were manually annotated and contain a single label based on a continuous arousal and valence. All our experiments involving this dataset are performed using the training and validation subset separation, as the test-set labels are not publicly available.

The AffectNet has the most representative data distribution among all the datasets we experimented with (as illustrated in Fig. 2a). Given the amount of samples and the data distribution, we use it to pre-train the FaceChannel in two of our experimental routines. This guarantees that the

FaceChannel has enough data samples to learn general facial features.

### OMG-Emotion

The one-minute gradual emotion recognition (OMG-Emotion) dataset [12] is composed of 675 videos with around 10 h of data, with an average length of one minute per video of persons performing monologs. The videos were gathered using web crawlers and manually annotated based on a continuous arousal and valence scale. Evaluating the FaceChannel on this dataset helps us to assess how our model performs when recognizing affect from particular individuals in a continuous scenario. Each video displays one person, and it is annotated at utterance level. The emotion expressions displayed in the OMG-Emotion dataset are heavily impacted by person-specific characteristics that are highlighted by the gradual change of emotional behavior over the entire video. The videos in this dataset cover a diverse range of arousal and valence values as seen in Fig. 2b).

### FER+

We also evaluate the FaceChannel on the FER+ dataset [36]. The FER+ contains around 31,000 face images crawled from the internet. Although collected in a similar way as the AffecNet dataset, the labels of the FER+ were collected using a different strategy. Each image was annotated by 10 different annotators using a categorical selection based on the basic emotions (angry, disgust, fear, happy, sad, surprise) and neutral. The categorical distribution for all 10 annotators is used to create a single label per image. Each label, thus, represents a composition of the basic emotions. It is important for us to evaluate FaceChannel on this dataset to assess its capability of adapting to composed labels.

### FABO

To evaluate the FaceChannel in a controlled environment setting, which is not present in any of the previous datasets, we use the Bi-modal Face and Body benchmark dataset FABO [37]. This corpus is composed of images with the face and body posture of different subjects. In our experiments, we focus on the facial expressions only. The dataset is composed of different videos, and in each of them, one subject performs a pre-defined expression in a cycle of two to four expressions per video.

Each of the videos of the FABO dataset has an annotation of the apex of the expressions. Six individual observers annotated each video and a voting process was executed. A total number of 281 videos are used as elaborated in Table 1. Ten affective labels were used: "Anger", "Anxiety", "Boredom", "Disgust", "Fear", "Happiness", "Surprise", "Puzzlement", "Sadness" and "Uncertainty". We only used the apex frames for each of the expressions to train our model.

**Table 1** Number of videos available for each emotional state in the FABO dataset

| Emotional state | Videos | Emotional state | Videos |
|---|---|---|---|
| Anger | 60 | Happiness | 28 |
| Anxiety | 21 | Puzzlement | 46 |
| Boredom | 28 | Sadness | 16 |
| Disgust | 27 | Surprise | 13 |
| Fear | 22 | Uncertainty | 23 |

Each video has 2–4 executions of the same expression

## Training and Evaluation Protocol

We follow the training and test separation established by the authors each of the datasets in our experimental setup. We enforce this, as it is important to maintain the comparability with previously proposed models. For each frame, we run a face detection based on the DLib python[1] library and re-size it to a dimension of $128 \times 128$ pixels.

## Metrics

We use two different metrics to measure the performance of the FaceChannel in our experiments: *accuracy*, when recognizing emotion labels, and the concordance correlation coefficient (CCC) [38] between the outputs of the models and the true labels when recognizing arousal and valence. The CCC is computed as:

$$\mathrm{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{2}$$

where $\mu_x$ and $\mu_y$ represent the mean for model predictions and the annotations and $\sigma_x^2$ and $\sigma_y^2$, are the corresponding variances. $\rho$ is Pearson's Correlation Coefficient between model prediction labels and the annotations.

The CCC allows us to compare directly the annotations available in the AffectNet and OMG-Emotion datasets. While, the accuracy helps us to evaluate the performance of the FaceChannel on the FABO and FER+ datasets.

[1] https://pypi.org/project/dlib/.

**Table 2** Concordance correlation coeffcient (CCC), for arousal and valence when evaluating the FaceChannel with the AffectNet dataset

| Model | Arousal | Valence |
|---|---|---|
| AlexNet [23] | 0.34 | 0.60 |
| MobileNet [39] | 0.48 | 0.57 |
| VGGFace [25] | 0.40 | 0.48 |
| **VGGFace + GAN** [40] | **0.54** | **0.62** |
| Face Channel | 0.46 | 0.61 |

**Table 3** Concordance correlation coeffcient (CCC), for arousal and valence when evaluating the different versions of the trained FaceChannel with the OMG-Emotion dataset

| Model | Trained | Tuned | Arousal | Valence |
|---|---|---|---|---|
| FaceChannel | OMG-Emotion | – | 0.12 | 0.23 |
| FaceChannel | AffectNet | – | 0.25 | 0.31 |
| FaceChannel | AffectNet | OMG-Emotion | 0.32 | 0.46 |

**Table 4** Concordance correlation coeffcient (CCC), for arousal and valence when evaluating the best version of the FaceChannel with the OMG-Emotion dataset

| Model | Arousal | Valence |
|---|---|---|
| Zheng et al. [24] | 0.35 | 0.49 |
| Huang et al. [10] | 0.31 | 0.45 |
| Peng et al. [41] | 0.24 | 0.43 |
| Deng et al. [42] | 0.27 | 0.35 |
| FaceChannel | 0.32 | 0.46 |

## Results

### AffectNet

Although, the AffectNet corpus is very popular, not many researchers report the performance of arousal and valence prediction on its validation set. This is probably the case as most of the research uses the AffectNet dataset to pre-train neural models for generalization tasks in other datasets, without reporting the performance on the AffectNet itself. We report the final results on the AffectNet dataset in Table 2. The baseline provided by the authors uses an AlexNet-based convolutional neural network [23] re-trained to recognize arousal and valence. A similar approach is reported by Hewitt and Gunes [39], but using a much reduced neural network, to be deployed on a smart-device. Lindt et al. [25] report experiments using the VGGFace, a variant of the VGG16 network pre-trained for face identification. Kollias et al. [40] proposed a novel training mechanics, where it augmented the training set of the AffectNet using a generative adversarial network (GAN), and obtained the best reported accuracy on this corpus, achieving 0.54 CCC for arousal and 0.62 CCC for valence. Our FaceChannel provides an improved performance when compared to most of these results achieving a CCC of 0.46 for arousal and 0.61 for valence. Different from the work of Kollias et al. [40], we train our model using only the available training set portion, and expect these results to improve when training on an augmented training set.

### OMG-Emotion

Training the FaceChannel on different datasets makes it easier to visualize the impact of a large number of training samples, as reported in Table 3. When trained only with the OMG-Emotion dataset, the model achieved a CCC of 0.12 for arousal and 0.23 for valence, which is much lower than any other training configuration. Training the model with the AffectNet dataset increased drastically its performance, but still pre-training the model with the AffectNet and fine-tuning it with the OMG-Emotion train set to yield the highest performance. This demonstrates that although the features learned when training the model with the AffectNet are general and somehow more reliable than the ones learned with the OMG-Emotion alone, the specificities of the OMG-Emotion data are still beneficial to improve the models' performance.

The performance of the FaceChannel is very similar when compared to the current state-of-the-art results on the OMG-Emotion dataset, as reported by the winners of the OMG-Emotion challenge where the dataset was proposed [24, 41, 42] as exhibited in Table 4. All these models also reported the use of pre-training of uni-sensory convolutional channels to achieve such results, but employed deep networks with much more parameters to be fine-tuned in an end-to-end manner. The use of attention mechanisms [24] to process the continuous expressions in the videos presented the best results of the challenge, achieving a CCC of 0.35 for arousal and 0.49 for valence. Temporal pooling, implemented as bi-directional long short-term memories (LSTMs) , achieved the second place, with a CCC of 0.24 for arousal and 0.43 for valence. The late-fusion of facial expressions, speech signals, and text information reached the third-best result, with a CCC of 0.27 for arousal and 0.35 for valence. The complex attention-based network proposed by Huang et al. [10] was able to achieve a CCC of 0.31 in arousal and 0.45 in valence, using only visual information.

### FER+

Similar to what happened when training the model with the OMG-Emotion, the routine that includes pre-training the FaceChannel with the AffectNet and fine-tuning it with a

**Table 5** Accuracy when evaluating the different versions of the trained FaceChannel with the FER+ dataset

| Model | Trained | Tuned | Accuracy |
|---|---|---|---|
| FaceChannel | FER+ | – | 87.50% |
| FaceChannel | AffectNet | – | 88.20% |
| FaceChannel | AffectNet | FER+ | 90.50% |

**Table 6** Accuracy when evaluating the FaceChannel with the FER+ dataset

| Model | Accuracy |
|---|---|
| CNN VGG13 [36] | 84.98% |
| SHCNN [43] | 86.54 |
| TFE-JL [44] | 84.3 |
| ESR-9 [26] | 87.15 |
| FaceChannel | 90.50% |

specific dataset achieved the best results when evaluating the FER+ dataset, as reported in Table 5. In this case, however, the performance gain is not as underlined as it was in the OMG-Emotion dataset, probably as the FER+ dataset has already a large number of data samples for training. The most important difference here is on the label representation, as the FER+ represents the labels using a distribution of annotations over the entire category range. This is probably the most impacting change when fine-tuning the FaceChannel with the FER+ causes, which impacts directly on the performance of the model.

When trained and evaluated with the FER+ model, our FaceChannel provides improved results as compared to those reported by the dataset authors [36]. They employ a deep neural network based on the VGG13 model, trained using different label-averaging schemes. Their best results are achieved using the labels as a probability distribution, which is the same strategy we used. We outperform their result by almost 6% as reported in Table 6. We also outperform the results reported in Miao et al. [43], Li et al. [44], and Siqueira et al. [26] which employ different type of complex neural networks to learn facial expressions.

## FABO

Table 7 reports the experiments of the different versions of the FaceChannel on the FABO dataset. Similar to the experiments involving the FER+ dataset, the results of training the FaceChannel only with the FABO dataset are not much worst than when training with only the AffectNet dataset. The FABO dataset also contains enough data to tune the convolution layers towards learning facial features. The improvement obtained when training the model with the AffectNet and tuning it with the FABO dataset, however, demonstrates that the FABO dataset also has specific peculiarities which are not depicted by the FaceChannel when trained with the AffectNet alone. As in all of our previous experiments, the best results were obtained in this training configuration. Our model achieves higher accuracy for the experiments with the FABO dataset as well when compared with the state-of-the-art for the dataset [45]. They report an approach based on recognizing each video-frame, similar to ours. The results reported by Gunes et al. [46] for Adaboost and SVM-based implementations are reported using a frame-based accuracy. Our FaceChannel outperforms both models, as illustrated in Table 8.

To better understand how the FaceChannel classifies the emotions on the FABO dataset, we present as well the accuracy per classification class in Table 9. We observe that the network provides an stable classification over all the classes, including the ones with less examples such as the "Surprise" class.

**Table 7** Accuracy when evaluating the different versions of the trained FaceChannel with the FABO dataset

| Model | Trained | Tuned | Accuracy |
|---|---|---|---|
| Face channel | FABO | – | 76.2% |
| Face channel | AffectNet | – | 75.9% |
| Face channel | AffectNet | FABO | 80.54% |

**Table 8** Accuracy when evaluating the FaceChannel with the FABO dataset

| Model | Accuracy (%) |
|---|---|
| Temporal normalization [45] | 66.50 |
| Bag of words [45] | 59.00 |
| SVM [46] | 32.49 |
| Adaboost [46] | 35.22 |
| Face channel | 80.54 |

**Table 9** Class-specific accuracy when evaluating the FaceChannel with the FABO dataset

| Class | Accuracy |
|---|---|
| Anger | 75.8 |
| Anxiety | 77.8 |
| Boredom | 80.1 |
| Disgust | 78.3 |
| Fear | 85.1 |
| Happiness | 83.9 |
| Puzzlement | 85.4 |
| Sadness | 80.4 |
| Surprise | 75.8 |
| Uncertainty | 77.4 |

**Table 10** Training time, number of parameters and number of training samples for all of our experiments when training the FaceChannel for 100 epoches

| Dataset | N. samples | Parameters | GPU | Training time |
|---|---|---|---|---|
| AffectNet | 1 million | 2 millions | Yes | 6 h |
| AffectNet | 1 million | 2 millions | No | 10 h |
| FER+ | 97 thousand | 2 millions | Yes | 2 h |
| FER+ | 97 thousand | 2 millions | No | 6 h |
| FABO | 5 thousand | 2 millions | Yes | 45 min |
| FABO | 5 thousand | 2 millions | No | 3 h |
| OMG-Emotion | 10 thousand | 2 millions | Yes | 1 h |
| OMG-Emotion | 10 thousand | 2 millions | No | 5 h |

**Table 11** Training time, number of parameters and number of training samples for all of our experiments when fine-tuning the FaceChannel for 100 epoches

| Dataset | N. samples | Parameters | GPU | Training time |
|---|---|---|---|---|
| FER+ | 97 thousand | 800 thousand | Yes | 45 min |
| FER+ | 97 thousand | 800 thousand | No | 2 h |
| FABO | 5 thousand | 800 thousand | Yes | 10 min |
| FABO | 5 thousand | 800 thousand | No | 40 min |
| OMG-Emotion | 10 thousand | 800 thousand | Yes | 25 h |
| OMG-Emotion | 10 thousand | 800 thousand | No | 1 h |

## Discussions

Our experiments demonstrate how the FaceChannel can perform better than most deeper neural networks for recognizing facial expressions on the AffectNet, OMG-Emotion, FER+ and FABO datasets. Besides reaching a good performance when evaluating facial expressions alone, the advantage of our model is its smaller configuration when compared to the popular versions of convolution neural networks.

### Fast training

The final configuration of the FaceChannel has around 2 *million* parameters to be updated during training. The entire model was developed using Keras [47], and trained on a system with an Intel i7 CPU, 16 GB of RAM, and a Quadro RTX 4000 GPU with 8 GB of memory. During pre-training, all the 2 million parameters were updated; while, during fine-tuning, only the last fully connected layer was re-trained, reducing the number of trainable parameters to 800 thousand.

Table 10 reports the training time for all of our full-training experiments. We observe that, as expected, the training times when using the GPU are much smaller than when using the CPU. But the most important to note is that even when training with the 1 million examples of the AffectNet dataset, the FaceChannel took only 10 hours to train with the CPU, and 6 hours to train with the GPU. When compared to most of the state-of-the-art deep learning models [48], these numbers demonstrate a great benefit of a smaller and light-weighted deep neural architecture.

The training effort is even smaller when we compare the resources needed for fine-tuning the network, reported in Table 11. With fewer parameters to be updated, the network takes only 2 h to train on the 97 thousand examples of the FER+ dataset. Also, the fine-tuning achieved the highest results in each of these datasets, demonstrating the capability of the FaceChannel to adapt robustly, but also quickly, towards new representations present on affective label re-associations.

**Table 12** Number of trainable parameters for deep learning models discussed in our results section

| Model | Parameters |
|---|---|
| FaceChannel | 2 million |
| MobileNet | 4.9 million |
| VGG13+ | 34 mmillion |
| AlexNet | 60 million |
| VGGFace+ | 138 million |

We also provide on Table 12, a comparison of how many parameters the deep learning models presented on our experiments have. We observe that the FaceChannel has by far the lowest number of parameters although presenting a better or same performance as the other models, as exhibited in Sect. 4.

### Facial features adaptation

To better understand the impact that each of these datasets has on the convolution filters of the FaceChannel, we perform a visualization analysis, based on the GradCam method [49]. Figure 3 exhibits the neural activation of the last convolutional layer of the FaceChannel when trained with each of the datasets. We observe that the network trained with the AffectNet and FER+ datasets focuses mostly on features on the central area of the face, mostly encapsulating the eyes/nose/mouth region. This is mostly due to the images in these datasets having a centralized pre-processing. The networks trained with the FABO dataset, however, have a strong bias towards focusing on the eyes and the chin. This is mostly due to the training samples on this dataset which are composed of extremely exaggerated facial expressions, in particular on the eyes/chin areas. The network trained with OMG-Emotion dataset, however, does not present a unique pattern, but a much more spread neural activation. Combined with the performance results, we can affirm that this happens because the network was not able to learn any specific feature characteristics when trained with this dataset.

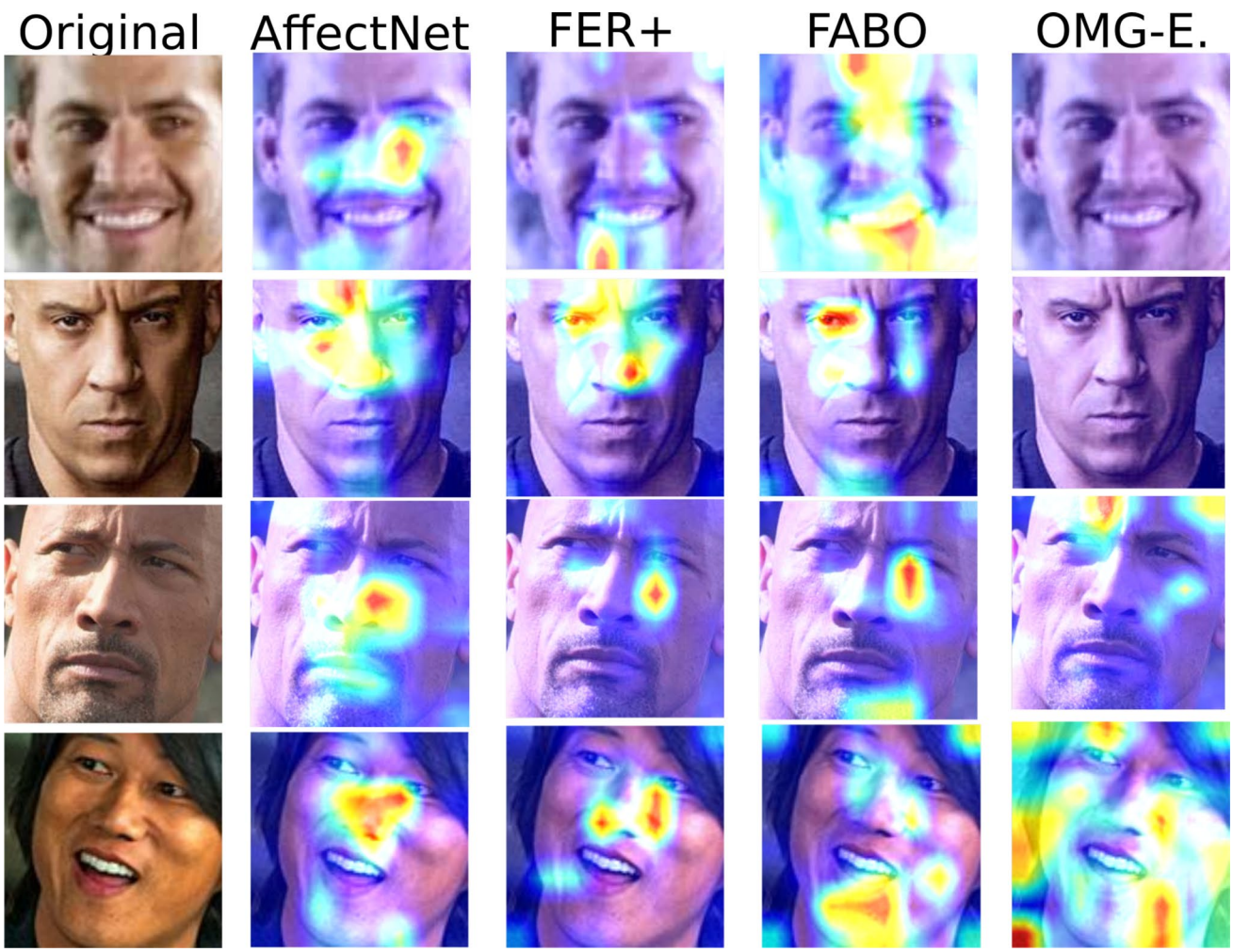


**Fig. 3** Feature-level analysis of how the different datasets impact the FaceChannel's capability of detecting facial features. The figure illustrates the results of the GradCam [49] visualization technique on the FaceChannel trained with the different datasets

Besides explaining the performance of the network, these analyses also help us understand why pre-training with the AffectNet achieves the best results on all of these datasets. The combination of the number of training samples, with a large variety of "in-the-wild" expressions for sure helped the network to tune its filters towards general facial features. This effect is observed also by recent research on training deep neural networks with facial expressions [33, 34, 50].

That our network was able to learn general features, even with a light-weighted architecture, is another testament to its strength at quick adaptation towards novel scenarios.

## Conclusions

We presented, in this paper, the formalization of the FaceChannel for automatic facial expression recognition (FER). Our neural network has an architecture based on the VGG16, but optimized to use much fewer trainable parameters. The reduction of the computational costs implies on a faster adaptation towards new scenarios, which is common when recognizing affect from different persons.

Our experiments demonstrate that our neural network has a compatible, and in most cases even better, performance than the current state-of-the-art models for automatic facial expression recognition (FER). We also demonstrate, using different FER datasets with specific data characteristics, how our model can be quickly adapted and fine-tuned for specific affective perception scenarios. To guarantee the reproducibility and dissemination of our model, we have made it fully available on GitHub[2].

We plan to study and deploy our model on platforms with reduced data processing capabilities, such as social robots.

[2] https://github.com/pablovin/FaceChannel.

Also, we believe that the light-weighted architecture will allow a quick adaptation towards individual aspects of affective performance, and thus, the development of personalized solutions is encouraged.

**Funding** Open access funding provided by Istituto Italiano di Tecnologia within the CRUI-CARE Agreement. This project is supported by a Starting Grant from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme. G.A. No 804388, wHiSPER". Open access funding provided by Istituto Italiano di Tecnologia within the CRUI-CARE Agreement.

**Availability of Data and Material** All the datasets used in our experiments are publicly available.

## Compliance with Ethical Standards

**Conflicts of interest** The authors declare that they have no conflict of interest.

**Code Availability** The entire code for the network's topology and the trained network are available here: shorturl.at/qtIT5

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

1. Ekman P, Friesen WV. Constants across cultures in the face and emotion. J Personal Soc Psychol. 1971;17(2):124–9.
2. Cavallo F, Semeraro F, Fiorini L, Magyar G, Sinčák P, Dario P. Emotion modelling for social robotics applications: a review. J Bionic Eng. 2018;15(2):185–203.
3. Hamann S, Canli T. Individual differences in emotion processing. Curr Opin Neurobiol. 2004;14(2):233–8.
4. Hess U, Blaison C, Kafetsios K. Judging facial emotion expressions in context: the influence of culture and self-construal orientation. J Nonverbal Behav. 2016;40(1):55–64.
5. Griffiths PE. Iii. basic emotions, complex emotions, machiavellian emotions 1. R Inst Philos Suppl. 2003;52:39–67.
6. Barrett LF. Solving the emotion paradox: categorization and the experience of emotion. Personal Soc Psychol Rev. 2006;10(1):20–46.
7. Afzal S, Robinson P. Natural affect data: Collection and annotation. New perspectives on affect and learning technologies. New York, NY: Springer; 2011. p. 55–70.
8. Mehta D, Siddiqui M, Javaid A. Facial emotion recognition: a survey and real-world user experiences in mixed reality. Sensors. 2018;18(2):416.
9. Hazarika D, Gorantla S, Poria S, Zimmermann R. Self-attentive feature-level fusion for multimodal emotion detection. In: 2018 IEEE conference on multimedia information processing and retrieval (MIPR). IEEE; 2018. p. 196–201.
10. Huang KY, Wu CH, Hong QB, Su MH, Chen YH. Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In: 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2019. p. 5866–70.
11. Kret ME, Roelofs K, Stekelenburg JJ, de Gelder B. Emotional signals from faces, bodies and scenes influence observers' face expressions, fixations and pupil-size. Front Hum Neurosci. 2013;7:810.
12. Barros P, Churamani N, Lakomkin E, Sequeira H, Sutherland A, Wermter S. The OMG-emotion behavior dataset. In: 2018 International joint conference on neural networks (IJCNN). IEEE; 2018. p. 1–7.
13. Kollias D, Tzirakis P, Nicolaou MA, Papaioannou A, Zhao G, Schuller B, Kotsia I, Zafeiriou S. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. Int J Comput Vis. 2019:1–23.
14. Kollias D, Schulc A, Hajiyev E, Zafeiriou S. Analysing affective behavior in the first abaw 2020 competition. 2020. arXiv :2001.11409.
15. Dhall A, Goecke R, Lucey S, Gedeon T, et al. Collecting large, richly annotated facial-expression databases from movies. IEEE Multimed. 2012;19(3):34–41.
16. Mollahosseini A, Hasani B, Mahoor MH. Affectnet: a database for facial expression, valence, and arousal computing in the wild. 2017. arXiv:1708.03985.
17. Zadeh AB, Liang PP, Poria S, Cambria E, Morency L-P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers), 2018. p. 2236–46.
18. Zafeiriou S, Kollias D, Nicolaou MA, Papaioannou A, Zhao G, Kotsia I. Aff-wild: valence and arousal 'in-the-wild' challenge. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017. p. 34–41.
19. Choi WY, Song KY, Lee CW. Convolutional attention networks for multimodal emotion recognition from speech and text data. In: Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML), 2018. p. 28–34.
20. Marinoiu E, Zanfir M, Olaru V, Sminchisescu C. 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 2158–67.
21. Du Z, Wu S, Huang D, Li W, Wang Y. Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. IEEE Trans Affect Comput. 2019. https://doi.org/10.1109/TAFFC.2019.2940224
22. Yang J, Wang K, Peng X, Qiao Y. Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In: Proceedings of the 20th ACM international conference on multimodal interaction. 2018. p. 594–98.
23. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012. p. 1097–1105.
24. Zheng Z, Cao C, Chen X, Xu G. Multimodal emotion recognition for one-minute-gradual emotion challenge. 2018. arXiv :1805.01060.
25. Lindt A, Barros P, Siqueira H, Wermter S. Facial expression editing with continuous emotion labels. In: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019). IEEE; 2019. p. 1–8.

26. Siqueira H, Magg S, Wermter S. Efficient facial feature learning with wide ensemble-based convolutional neural networks. 2020. arXiv:2001.06338.
27. Tapus A, Bandera A, Vazquez-Martin R, Calderita LV. Perceiving the person and their interactions with the others for social robotics-a review. Pattern Recognit Lett. 2019;118:3–13.
28. Barros P, Wermter S. Developing crossmodal expression recognition based on a deep neural model. Adapt Behav. 2016;24(5):373–96.
29. Barros P, Churamani N, Sciutti A. The facechannel: a light-weight deep neural network for facial expression recognition. In: 2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020) (FG), (Los Alamitos, CA, USA). IEEE Computer Society; 2020. p. 449–53.
30. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:1409.1556.
31. Fregnac Y, Monier C, Chavane F, Baudot P, Graham L. Shunting inhibition, a silent step in visual cortical computation. J Physiol. 2003;441–451.
32. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding neural networks through deep visualization. 2015. arXiv:1506.06579.
33. Mousavi N, Siqueira H, Barros P, Fernandes B, Wermter S. Understanding how deep neural networks learn face expressions. In: 2016 international joint conference on neural networks (IJCNN). IEEE; 2016. p. 227–34.
34. Zhou Y, Shi BE. Action unit selective feature maps in deep networks for facial expression recognition. In: 2017 international joint conference on neural networks (IJCNN). IEEE; 2017. p. 2031–38.
35. Bergstra JS, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In: Advances in neural information processing systems. 2011. p. 2546–54.
36. Barsoum E, Zhang C, Canton Ferrer C, Zhang Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016.
37. Gunes H, Piccardi M. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In 18th International conference on pattern recognition (ICPR'06) 2006. (Vol. 1, pp. 1148-1153). IEEE.
38. Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;255–268.
39. Hewitt C, Gunes H. Cnn-based facial affect analysis on mobile devices. 2018. arXiv:1807.08775.
40. Kollias D, Cheng S, Ververas E, Kotsia I, Zafeiriou S. Deep neural network augmentation: Generating faces for affect analysis. Int J Comput Vis . 2020;1–30.
41. Peng S, Zhang L, Ban Y, Fang M, Winkler S. A deep network for arousal-valence emotion prediction with acoustic-visual cues. 2018. arXiv:1805.00638.
42. Deng D, Zhou Y, Pi J, Shi BE. Multimodal utterance-level affect analysis using visual, audio and text features. 2018. arXiv:1805.00625.
43. Miao S, Xu H, Han Z, Zhu Y. Recognizing facial expressions using a shallow convolutional neural network. IEEE Access. 2019;7:78000–11.
44. Li M, Xu H, Huang X, Song Z, Liu X, Li X. Facial expression recognition with identity and emotion joint learning. IEEE Trans Affect Comput. 2018. https://doi.org/10.1109/TAFFC.2018.2880201
45. Chen S, Tian Y, Liu Q, Metaxas DN. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. Image vision comput. 2019;31(2):175–85.
46. Gunes H, Piccardi M. Automatic temporal segment detection and affect recognition from face and body display. IEEE Trans Syst Man Cybern Part B Cybern. 2009;39:64–84.
47. Gulli A, Pal S. Deep learning with Keras. Birmingham: Packt Publishing Ltd; 2017.
48. Li S, Deng W. Deep facial expression recognition: a survey. IEEE Trans Affect Comput. 2020.
49. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 618–26.
50. Patel K, Mehta D, Mistry C, Gupta R, Tanwar S, Kumar N, Alazab M. Facial sentiment analysis using AI techniques: state-of-the-art, taxonomies, and challenges. IEEE Access. 2020;8:90495–519.


**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.