Deep Learning based FACS Action Unit Occurrence and Intensity Estimation

Amogh Gudi, H. Emrah Tasli, Tim M. den Uyl, Andreas Maroulis
Vicarious Perception Technologies, Amsterdam, The Netherlands

Abstract— **Ground truth annotation of the occurrence and intensity of FACS Action Unit (AU) activation requires great amount of attention. The efforts towards achieving a common platform for AU evaluation have been addressed in the FG 2015 Facial Expression Recognition and Analysis challenge (FERA 2015). Participants are invited to estimate AU occurrence and intensity on a common benchmark dataset. Conventional approaches towards achieving automated methods are to train multi-class classifiers or to use regression models. In this paper, we propose a novel application of a deep convolutional neural network (CNN) to recognize AUs as part of FERA 2015 challenge. The 7 layer network is composed of 3 convolutional layers and a max-pooling layer. The final fully connected layers provide the classification output. For the selected tasks of the challenge, we have trained two different networks for the two different datasets, where one focuses on the AU occurrences and the other on both occurrences and intensities of the AUs. The occurrence and intensity of AU activation are estimated using specific neuron activations of the output layer. This way, we are able to create a single network architecture that could simultaneously be trained to produce binary and continuous classification output.**

I. INTRODUCTION

A picture is worth a thousand words, but how many words is the picture of a face worth? As humans, we make a number of conscious and subconscious evaluations of a person just by looking at their face. Identifying a person can have a defining influence on our conversation with them based on past experiences; estimating a person's age, and making a judgement on their ethnicity, gender, etc. makes us sensitive to their culture and habits. We also often form opinions about that person (that are often highly prejudiced and wrong); we analyse his or her facial expressions to gauge their emotional state (e.g. happy, sad), and try to identify non-verbal communication messages that they intent to convey (e.g. love, threat). We use all of this information when interacting with each other. In fact, it has been argued that neonates, only 36 hours old, are able to interpret some very basic emotions from faces and form preferences [1]. In older humans, this ability is highly developed and forms one of the most important skills for social and professional interactions. Indeed, it is hard to imagine expression of humour, love, appreciation, grief, enjoyment or regret without facial expressions.

Human face constantly conveys information, both consciously and subconsciously. However, as basic as it is for humans to visually interpret this information, it is quite a big challenge for machines. Such studies have been proven important in many different fields like psychology [2], humancomputer interaction [3], visual expression monitoring [4]- [5] and market research [6]. Conventional semantic facial feature recognition and analysis techniques mostly suffer from lack of robustness in real life scenarios.

This paper proposes a method to interpret semantic information available in faces in an automated manner without requiring manual design of feature detectors, using the approach of Deep Learning. Deep Learning refers to the use of Deep Neural Networks, an Artificial Neural Network with two or more hidden layers, for the task of recognizing patterns in data. The way Deep Learning works is intuitive and principle-oriented: higher level concepts can be seen as specific groupings of multiple lower level concepts. Considering the recent success of deep learning techniques, it is interesting to evaluate the performance of such a solution on the problem of FACS Action Unit estimation.

FG 2015 Facial Expression Recognition and Analysis challenge invites researchers to develop methods to estimate facial AU occurrence and intensity. The proposed network architecture in this study has been previously developed for detecting facial semantic features (like emotions, age, gender, ethnicity etc.) present in faces [7]. The same architecture is trained for the given tasks using the ground truth datasets.

II. RELATED WORK

There has been previous efforts on detecting the facial AU occurrence as well as the intensity. The study in [8] utilizes the facial landmark displacement as a measure of intensity dynamics. Regression based models [9] and multi-class classifiers are also [10] utilized for such purposes.

Latest developments in computational hardware have redirected attention to deep neural networks for many computer vision tasks. In the light of such studies deep learning methods have been proposed as a highly promising approach in solving such facial semantic feature recognition tasks. A recent work for facial recognition by Taigman et al. [11] has shown near-human performance using deep networks. Thanks to the use of pre-processing steps like face alignment and frontalization, and the use of a very large dataset, a robust and invariant classifier is produced that sets the state-of-the-art in the Labelled Faces in the Wild dataset [12].

In the task of emotion recognition from faces, Tang [13] sets the state-of-the-art on the Facial Expression Recognition Challenge-2013 (FERC-2013). This is achieved by implementing a two stage network: a convolutional network trained in a supervised way on the first stage, and a Support Vector Machine as the second stage, which is trained on

the output of the first stage. The initial global contrast normalization has proven to be beneficial and hence has also been adopted as a pre-processing step in this paper.

III. DATASETS

FERA 2015 challenge provides two ground truth datasets: the BP4D, and the SEMAINE Dataset. The BP4D Dataset contains occurrence annotations for Action Units 1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, as well as intensity annotations for Action Units 6, 10, 12, 14, 17. The SEMAINE data contains occurrence annotations for Action Units 2, 12, 17, 25, 28, 45. Details of the dataset are explained in the FERA 2015 baseline paper [14].

IV. METHOD

In this section, we describe the proposed method used to detect the occurrences and intensities of Facial AUs from face images. The algorithmic pipeline depicted in 2 is primarily based on the master's thesis work [7]. The whole pipeline is composed of the initial pre-processing step and followed by the classification step using the deep neural network.

A. Pre-Processing

The AU classifier (the deep network) is designed to accept frontal images of cropped faces as input. Therefore, we apply an initial pre-processing step to detect the face and handle pose variations. In an effort to normalize the input intensity differences due to lighting conditions, we also apply a global contrast normalization in this initial step. The basic pipeline of these pre-processing steps are as follows:

Face Location Normalization:

- We find faces in the images using a face detection algorithm (using the facial points based on the method described in [15], and as provided by the challenge) and extract the crop of the face such that the image is centered around the face.
- We perform in-plane rotation in order to remove tilt in the X-Y plane. This is achieved by enforcing the line connecting the two eyes to be horizontal.
- We resize the image such that the approximate scale of the face is constant. This is done by ensuring the distance between the two eyes to be constant. The final resolution of the face-crop is set to 48×48 pixels (this input resolution strikes a good balance between classification performance and computational complexity [7]).

Global Contrast Normalization:

- We convert the face-crop to 32-bit grayscale images to minimize the computational complexity of using multiple color channels.
- We normalize the pixel values by the standard deviation of pixel values (face-crop) in the whole video session.

This pre-processing pipeline is illustrated in Figure 1.

In addition, in order to increase the variance of training data (with the intention of making the network more robust), we add a horizontally mirrored copy of all training images.

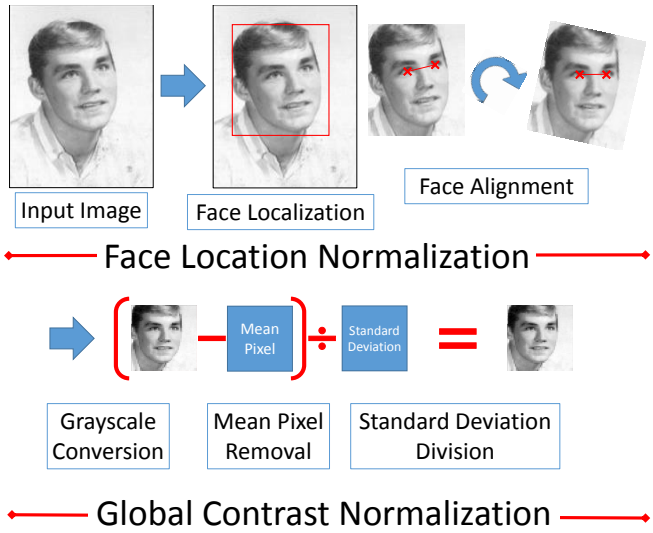


Fig. 1: The Pre-processing pipeline.

B. The Deep Neural Network

The estimation of AUs from the pre-processed face-crops is performed by a Deep Convolutional Neural Network. The choice of the hyper-parameters of this network are based on the work in [7].

In our framework, the input image size of the network is set to 48×48 grayscale pixels arranged in a 2D matrix. This pre-processed image is fed to the first hidden layer of the network: A convolutional layer with a kernel size of 5×5 having a stride of 1 both dimensions. The number of parallel feature-maps in this layer is 64. The 44×44 output image produced by this layer is then passed to a local contrast normalization and a max-pooling layer [16] of kernel size 3×3 with a stride of 2 in each dimension. This results in a sub-sampling factor of $1/2$, and hence the resulting image is of size 22×22. The second hidden layer is also a 64 feature-map convolutional layer with a kernel size of 5×5 (and stride 1). The output of this layer is a 18×18 pixel image, and this feeds directly into the third hidden layer of the network, which is again a convolutional layer of 128 feature maps with a kernel size of 4×4 (and stride 1). Finally, the output of this layer, which is of dimension 15×15, is fed into the last hidden layer of the network, a fully connected linear layer with 3072 neurons. Dropout technique [17], [18] is applied to this fully connected layer, with a dropout probability of 0.2. The output of this layer is connected to the output layer, which is composed of either 11 or 6 neurons (11 for the BP4D database, and 6 for the SEMAINE dataset), each representing one AU class label. All layers in the network are made up of ReLu units/neurons [19]. This architecture is illustrated in Figure 2.

Training of the deep neural network is done using stochastic gradient descent with momentum in mini-batch mode, with batches of 100 data samples. Negative log-likelihood is used as the objective function.

It should be noted that for the BP4D dataset, while the network was trained on occurrence of AUs 1, 2, 4, 7, 15, 23, it was also trained on the intensities of AUs 6, 10, 12, 14,

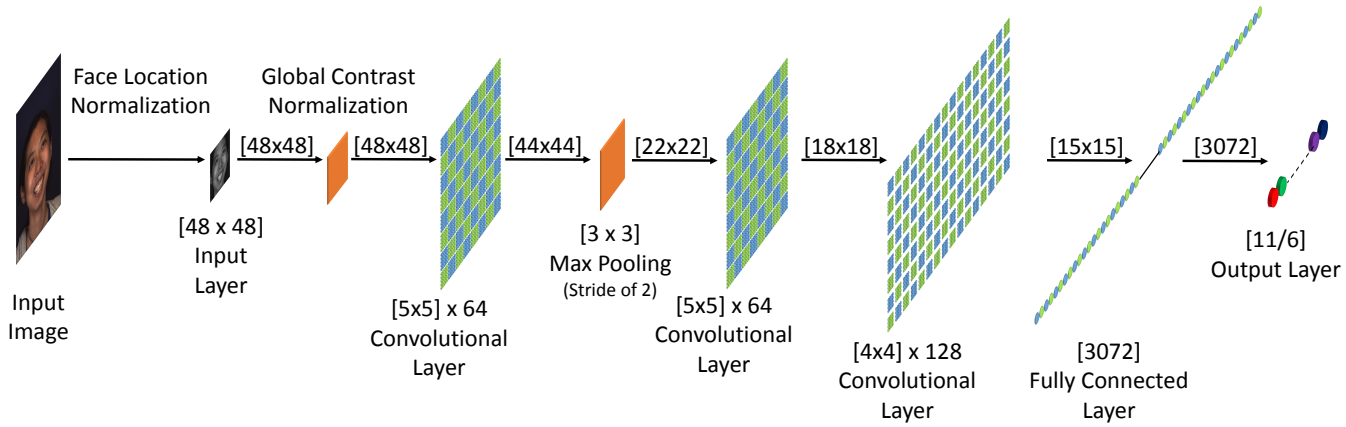


Fig. 2: The architecture of the deep convolutional neural network.

17. This way, the same network is able to participate in both the occurrence sub-challenge (for the BP4D dataset), as well as the two intensity sub-challenges.

The training of the deep network involved over 140,000 images of the BP4D dataset and 90,000 images of the SEMAINE Dataset (excluding mirrored copies). Initially, 50% of these were used in the training sets, while the rest (development sets) was used as the validation sets. However, in order to make full use of the annotated data given, 100% of the data was included in the training sets in the final versions of the networks, which were trained for a fixed number of epochs (determined by validation set performance in previous training experiments).

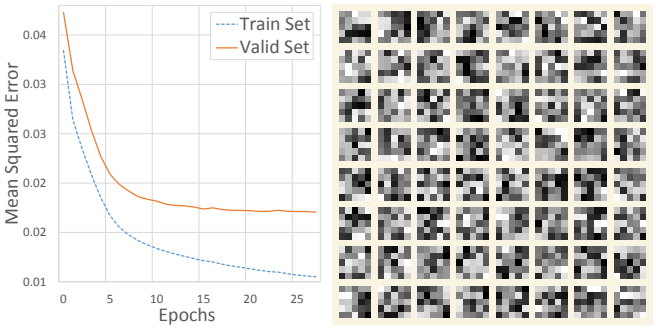
C. Post-Processing

The raw output of the deep network (the activations of the final layer) essentially denote the confidence scores for the presence/absence of each AU, where an extremely low value represents a strong absence of an AU, while a high value represents its strong presence. In an ideal case, the decision threshold of the deep network must lie on the mid-point of values representing the ground-truth presence and absence of an AU in the training set. However, many conditions can result in this threshold being skewed to one of the two extremes (e.g., uneven distribution of occurrences in the training set).

As a work-around to this problem, we optimize the decision threshold of the AUs on a validation set, i.e., we set the decision thresholds to a value that gives be highest F1 score, when tested on a validation set. In our case, we first separated the annotated dataset into two partitions: a training set and a validation set (same as the development set by the challenge). Next, we train the network only on the training set, and test all possible values of the decision threshold on the development validation set. Finally, we apply these best-performing thresholds as the decision thresholds for our final network, which is trained on the complete set of provided data (including the development set).

V. RESULTS

The training of these networks took roughly 15 hours each on a 1000+ core GPU. The learning curve and the weights



(a) Learning curve of the network w.r.t. mean squared error on the training and validation set.

(b) Visualization of the weights of the first-layer convolutional feature-maps.

Fig. 3: Deep neural network learning curve and weights trained on the BP4D Dataset.

BP4D Dataset		SEMAINE Dataset	
Action Unit	F1 Score	Action Unit	F1 Score
AU01	0.399	AU02	0.372
AU02	0.346	AU12	0.707
AU04	0.317	AU17	0.067
AU06	0.718	AU25	0.602
AU07	0.776	AU28	0.040
AU10	0.797	AU45	0.257
AU12	0.793		
AU14	0.681		
AU15	0.235		
AU17	0.368		
AU23	0.309		
Mean	**0.522**	**Mean**	**0.341**

TABLE I: Occurance sub-challenge results for BP4D and SEMAINE datasets.

Action Unit	MSE	PCC	ICC
AU06	1.287	0.664	0.663
AU10	1.249	0.735	0.733
AU12	0.928	0.788	0.788
AU14	1.686	0.591	0.549
AU17	0.757	0.329	0.329
Mean	**1.181**	**0.621**	**0.613**

TABLE II: Fully automated intensity estimation sub-challenge results for the BP4D dataset.

Action Unit	MSE	PCC	ICC
AU06	1.125	0.423	0.423
AU10	0.963	0.555	0.543
AU12	0.803	0.632	0.613
AU14	1.554	0.533	0.495
AU17	1.198	0.244	0.219
Mean	**1.129**	**0.478**	**0.459**

TABLE III: Pre-segmented intensity estimation sub-challenge results for the BP4D dataset.

of the trained network (on the BP4D Dataset) can be seen in Figure 3.

The performance of the network on the test set on the three sub-challenges of FERA-2015 are presented in Tables I,II and III. The evaluation methods used in these results are the F1 Score, the Mean Squared Error (MSE), Pearsons Correlation Coefficient (PCC) and the Intra-class Correlation Coefficient (ICC). The sub-challenges, test set and the evaluation measure are explained in [14].

As can be seen, our deep network performs with a mean F1 score of 0.522 on the BP4D test set, and 0.341 on the SEMAINE dataset in the occurrence sub-challenge. The method gives an average mean squared error of 1.181 and 1.129 on the fully automated and pre-segmented intensity sub-challenges respectively. Additionally, on the BP4D dataset, it was observed that video-segment global contrast normalization pre-processing step contributes an improvement of 0.051 to the final F1 score, while the post-processing step improves the final F1 score by 0.113.

It can be seen that the performance of our method on the SEMAINE dataset was much lower than on the BP4D dataset. One of the main contributing factors to this observation is the low number of individual faces included in the SEMAINE training data (31 people), as compared to the BP4D training data (41 individuals). In fact, we have been able to attain much higher performance for emotion and facial characteristics estimation in previous experiments [7], where the same network was trained on datasets containing more than 200 individuals. This suggests that the deep network ends up over-fitting by learning the identities of the individuals in the training set.

Some classification examples (from the validation set) by the network can be viewed in Figure 4.

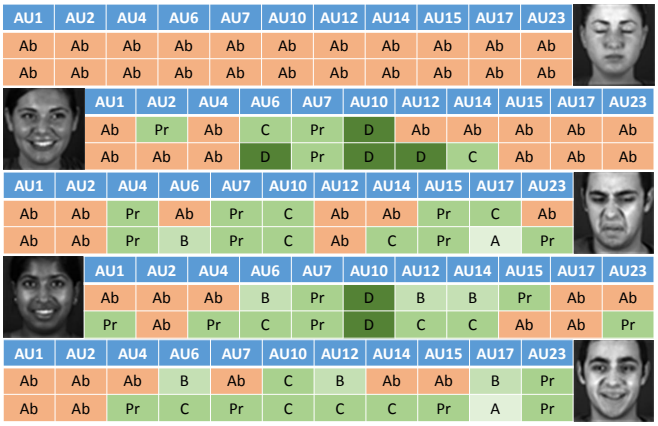


Fig. 4: Example AU estimation of images from the BP4D Dataset. The first row corresponds to the ground truth, and the second row corresponds to the AU estimation by the proposed method. Legend: Ab - Absent, Pr - Present, A,B,C,D,E - AU Intensity Levels

VI. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

In this paper, a deep learning approach has been demonstrated for detecting the occurrence and estimating the intensity of facial AUs. This approach is primarily based on the use of convolutional neural networks on two dimensional pre-processed and aligned images of faces. Deep convolutional networks have been recently very successful in very different visual tasks. That has been the main motivation of utilizing it in this specific challenge. However, the results have not been as promising as expected. We believe that this is due to the fact that the deep learning techniques depend on being trained on big datasets with high variation in the data. In the datasets supplied in the challenge, there is little variation in terms of the number of individuals and this can cause over fitting of the network on the individual faces in the dataset.

B. Future Works

- This framework proposed in this paper does not take into account the temporal dimension of the input data (frames of videos). Assuming time-domain information could improve the results, 3-D convolutional networks as proposed in [20], could readily fit into our framework.
- More extensive experimentation with alternative pre-processing techniques could be carried out. Methods like whitening are used in a wide range of machine learning tasks to reduce redundancy, and could be used to aid deep networks as well.
- A limiting factor in the conducted experiments is the available computational resources. This calls for more experimentation on larger networks, as the true optimal performance of these networks can only be achieved after extending the upper bound restrictions on the network size.

REFERENCES

[1] Teresa Farroni, Enrica Menon, Silvia Rigato, and Mark H Johnson, "The perception of facial expressions in newborns," *European Journal of Developmental Psychology*, vol. 4, no. 1, pp. 2–13, 2007.

[2] Paul Ekman, Wallace V Freisen, and Sonia Ancoli, "Facial signs of emotional experience.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1125, 1980.

[3] Maja Pantic and Leon JM Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[4] H. Emrah Tasli and Paul Ivan, "Turkish presidential elections trt publicity speech facial expression analysis," in *arXiv preprint arXiv:1408.3573*, 2014.

[5] H Emrah Tasli, Amogh Gudi, and Marten Den Uyl, "Integrating remote ppg in facial expression analysis framework," in *ACM International Conference on Multimodal Interaction (ICMI)*. ACM, 2014, pp. 74–75.

[6] Thales Teixeira, Michel Wedel, and Rik Pieters, "Emotion-induced engagement in internet video advertisements," *Journal of Marketing Research*, vol. 49, no. 2, pp. 144–159, 2012.

[7] Amogh Gudi, "Recognizing semantic features in faces using deep learning," M.S. thesis, Faculty of Science, University of Amsterdam, 2014, https://esc.fnwi.uva.nl/thesis/centraal/files/f2115870570.pdf.

[8] Michel Valstar and Maja Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*. IEEE, 2006, pp. 149–149.

[9] Arman Savran, Bulent Sankur, and M Taha Bilge, "Regression-based intensity estimation of facial action units," *Image and Vision Computing*, vol. 30, no. 10, pp. 774–784, 2012.

[10] Mohammad H Mahoor, Steven Cadavid, Daniel S Messinger, and Jeffrey F Cohn, "A framework for automated measurement of the intensity of non-posed facial action units," in *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*. IEEE, 2009, pp. 74–80.

[11] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1701–1708.

[12] Gary B Huang and Erik Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," .

[13] Yichuan Tang, "Deep learning using linear support vector machines," in *Workshop on Challenges in Representation Learning, ICML*, 2013.

[14] Michel François Valstar, "Fera 2015 - second facial expression recognition and analysis challenge," in *Automatic Face & Gesture Recognition and Workshops (FG 2015), 2015 IEEE International Conference on*. IEEE, 2015.

[15] Xuehan Xiong and Fernando De la Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 532–539.

[16] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet, "Convolutional networks and applications in vision," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*. IEEE, 2010, pp. 253–256.

[17] Nitish Srivastava, *Improving neural networks with dropout*, Ph.D. thesis, University of Toronto, 2013.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[19] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[20] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt, "Sequential deep learning for human action recognition," in *Human Behavior Understanding*, pp. 29–39. Springer, 2011.