Deep Learning for Facial Action Unit Detection under Large Head Poses

Zoltán Tősér[1], László A. Jeni[2], András Lőrincz[1], and Jeffrey F. Cohn[2,3]

[1] Faculty of Informatics, Eötvös Loránd University, Budapest, Hungary
[2] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[3] Department of Psychology, The University of Pittsburgh, Pittsburgh, PA, USA
{toserzoltan,lorincz}@inf.elte.hu, laszlojeni@cmu.edu, jeffcohn@pitt.edu

Abstract. Facial expression communicates emotion, intention, and physical state, and regulates interpersonal behavior. Automated face analysis (AFA) for the detection, synthesis, and understanding of facial expression is a vital focus of basic research with applications in behavioral science, mental and physical health and treatment, marketing, and human-robot interaction among other domains. In previous work, facial action unit (AU) detection becomes seriously degraded when head orientation exceeds 15° to 20°. To achieve reliable AU detection over a wider range of head pose, we used 3D information to augment video data and a deep learning approach to feature selection and AU detection. Source video were from the BP4D database (n = 41) and the FERA test set of BP4D-extended (n = 20). Both consist of naturally occurring facial expression in response to a variety of emotion inductions. In augmented video, pose ranged between −18° and 90° for yaw and between −54° and 54° for pitch angles. Obtained results for action unit detection exceeded state-of-the-art, with as much as a 10% increase in F_1 measures.

Keywords: deep learning, facial action unit detection, pose dependence

1 Introduction

The face is one of the most powerful channels of nonverbal communication [3, 5]. Facial expression provides cues about emotion, intention, alertness, pain, personality, regulates interpersonal behavior [4], and communicates psychiatric [8] and biomedical status [10] among other functions.

There has been increasing interest in automated facial expression analysis within the computer vision and machine learning communities. Several applications for related technologies exist: distracted driver detection [27], emotional response measurement for advertising [23, 25], and human-robot collaboration [2] are just some possibilities.

Given the time-consuming nature of manual facial expression coding and the alluring possibilities of the aforementioned applications, recent research has pursued computerized systems capable of automatically analyzing facial expressions. The dominant approach adopted by these researchers has been to identify

a number of fiduciary points on the face, extract hand-crafted or learned features that can characterize the appearance of the skin, and train classifiers in a supervised manner to detect the absence or presence of expressions.

Recently, deep learning based solutions have been proposed for coding holistic facial expressions and facial actions units. Li et al. [21] used a convolutional neural network (CNN) based deep representation of facial 3D geometric and 2D photometric attributes for recognizing holistic facial expressions. Liu et at. [22] proposed an Action Unit aware deep architecture to learn local appearance variations on the face and constructed a group-wise sub-network to code facial expressions. Xu et al. [28] explored transfer learning of high-level features from face identification data to holistic facial expression recognition. Only recently did Jaiswal and Valstar [12] propose a deep learning approach for recognizing facial action units under uncontrolled conditions. Action Units were coded using a memory network that jointly learns shape, appearance and dynamics in a deep manner.

Even though significant progress has been made [7], the current state-of-the-art science is still limited in several key respects. Stimuli to elicit spontaneous facial actions have been highly controlled and camera orientation has been frontal with little or no variation in head pose. Head motion and orientation to the camera are important if AU detection is to be accomplished in social settings where facial expressions often co-occur with head motion [17, 1]. As the head pose moves away from frontal, parts of the face may become self-occluded and the classifier's ability to measure expressions degrades. Here, we study the efficiency of a novel deep learning method for AU detection under large head poses.

This paper advances two main novelties:

AU Detection under Large Head Poses with 3D Augmentation
In our work we use the BP4D spontaneous dataset and its extension detailed in Sect. 2.2. An augmented dataset has been created using the 3D information and renderings of the faces with broad range of yaw and pitch rotations. We show that performance is high for the networks trained around different pose directions opening the door for a number of useful applications.

Selective Gradient Descent Optimization
Threshold performance metrics (such as the F_1 score) are piecewise-constant functions and including them directly in the CNN cost function would degrade the convergence of the optimization method. In our algorithm, we combined gradient descent with selective methods to overcome this issue. This approach results in a small but highly effective network that outperforms the more complex state-of-the-art systems.

The paper is organized as follows. The method section (Sect. 2) contains the overview of the architecture (2.1), the descriptions about database (2.2), its extension (2.3), the facial landmark tracking method (2.4) and the deep learning components (2.5). These descriptions are followed by our results (Sect. 3) and the related discussion (Sect. 4). We conclude in the last section (Sect. 5).

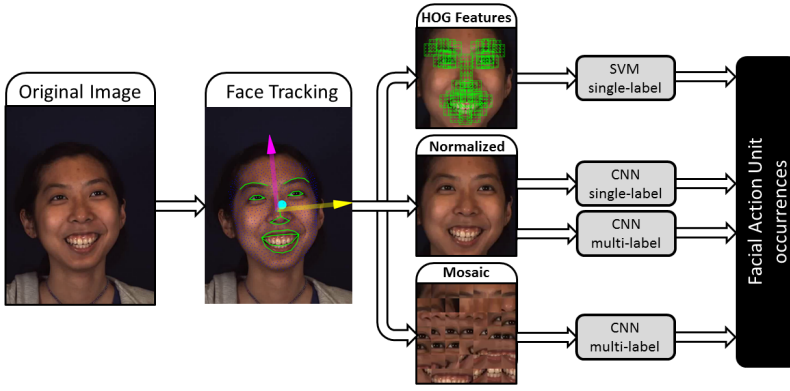

Fig. 1. Overview of the system. The original image underwent face tracking and was pre-processed in three different ways; Histogram of Gradients (HoG), similarity normalized (scaled and cropped), and cut and put together from patches around landmark positions (Mosaic). Training methods included Support Vector Machine (SVM), Convolutional Neural Networks (CNN) in single and multi-label versions.

2 Methods

2.1 Architecture

The main steps of pre-processing, such as face detection, mesh fitting, pose estimation are depicted in Fig. 1. Details are to follow below.

2.2 BP4D-Spontaneous Dataset

We used the BP4D-Spontaneous dataset [31] from the FERA 2015 Challenge [26]. This database includes digital video of 41 participants (56.1% female, 49.1% white, ages 18-29). These individuals were recruited from the departments of psychology and computer science and from the school of engineering at Binghamton University. All participants gave informed consent to the procedures and permissible uses of their data. Participants sat approximately 51 inches in front of a Di3D dynamic face capturing system during a series of eight emotion elicitation tasks. Target emotional expressions include anxiety, surprise, embarrassment, fear, pain, anger, and disgust. Example tasks include being surprised by a loud sound, submerging a hand in ice water, and smelling rotten meat. For each task, the 20-second segment with the highest AU density was identified; this segment then was coded for AU onset (start) and offset (end) by certified and reliable FACS coders.

The FERA 2015 Challenge [26] employed the 41 subjects from BP4D - Spontaneous dataset [31] as a training set. In this paper we refer this subset as "Train" set. Additional videos from 20 subjects were collected using the same setup and were used for testing in the challenge [26]. In this paper we refer this subset as "Test" set.

2.3 Database extension

The subjects in the BP4D-Spontaneous dataset exhibit only a moderate level of head movements in the video sequences. The dataset [31] comes with frame-level high-resolution 3D models. To validate the proposed method on larger viewpoint angles, an augmented dataset has been created using the 3D information and renderings of the faces with different yaw and pitch rotations. We used all the FACS coded data to synthesize the rotated views.

2.4 Facial Landmark Tracking and Face Normalization

The first step in automatically detecting AUs was to locate the face and facial landmarks. Landmarks refer to points that define the shape of permanent facial features, such as the eyes and lips. This step was accomplished using the ZFace tracker [14, 15], which is a generic tracker that requires no individualized training to track facial landmarks of persons it has never seen before. It locates the two- and three-dimensional coordinates of main fiducial landmarks in each image. These landmarks correspond to important facial points such as the eye and mouth corners, the tip of the nose, and the eyebrows. The moderate level of rigid head motion exhibited by the subjects in the BP4D-Spontaneous dataset was minimized as follows: facial images were warped to the average pose and face using similarity transformation on the tracked facial landmarks. The average face has been normalized to have 100 pixels inter-ocular distance and normalized images were cropped to 256x256 pixels. This procedure created a common space, where variation in head size and orientation would not confound the measurement of facial actions.

2.5 Deep learning

Deep learning aims to overcome the curse of dimensionality problem of MLPs via a number of architectural inventions. The increase of the number of layers lessens the transformational tasks of each layer. Rectified linear units (ReLUs) are favoured, since their sensitive range is large, the rectification can efficiently shatter the space, and supervised training does not require unsupervised pre-training (see [9] and the references therein).

Layers of the Network. Convolutional layers make another efficient innovation. They are particularly useful for images. One can view each layer as a set of trainable template matchings [6]. It has the following attractive properties: (a) The templates (also called filters) can be matched at each pixel of the image relatively quickly due to the convolution operation itself [20]. The result for each filter is called the feature map. (b) While the number of neurons can be large, still the number of variables, the weights, is kept low, saving in memory requirements and decreasing the curse of dimensionality problem. (c) Each convolutional layer may be followed by a subsampling layer. The role of this step

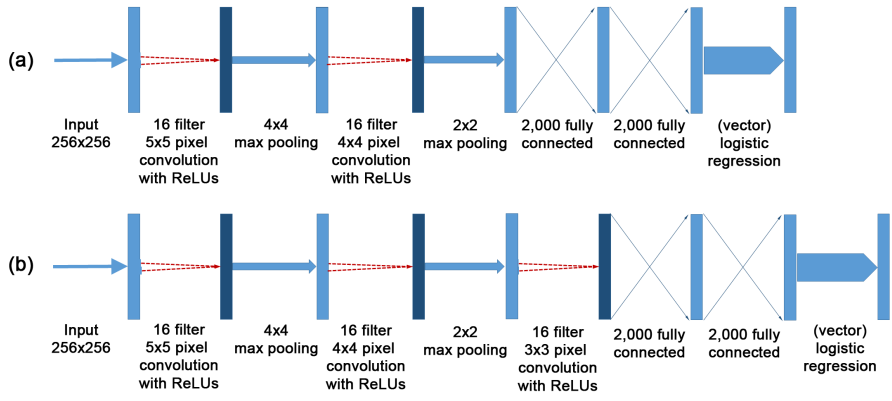

Fig. 2. Deep neural network, main components: convolutional layers with ReLUs (CL), pooling layers (PL), fully connected layers (FC), output layer with logistic regression (OL). There are two versions. (a): CL-PL-CL-PL-FC-FC-OL, (b): additional CL between the second PL and the first FC layer.

is to decrease the number of units that scale as the product of the dimension of the input of that layer and the number of filters. Max-pooling that solicits the largest response in each pooling region is one of the preferred methods. The effective result of pooling is that the precision of the feature map degrades, which is nicely compensated by the number of feature maps and the option of further convolutional processing steps without explosion in the number of units. Subsampling also reduces overfitting. For more details, see [19] and the references therein.

Convolutional networks typically add densely connected layers after the convolutional layers, often made of ReLUs. Our architecture is sketched in Fig. 2.

We used typical regularization, stabilization, early stopping, and local minima avoiding procedures [24] with a reasonably small network and we found that larger networks would not improve performance considerably. The parameters and some procedures of the architecture are as follows:

(a) The dimension of the input layer is 256×256. The original three channel color images were converted to a single grayscale channel and the values were scaled between 0 and 1.

(b) The first and second convolutional cascades have 16 filters each, with 5×5 in the first and 4×4 pixels in the second cascade. The stride was 1 in both cases. Max pooling was 4×4 and 2×2 applied with a stride of 4 and 2, respectively. Occasionally a third convolutional layer with 16 filters of 3×3 pixels each was added when the representation power of the architecture was questioned (Fig. 2).

(c) There are two densely connected layers of 2,000 ReLU units in each.

(d) The output is a sigmoid layer for the action units. Special procedures include dropout before the two dense and the sigmoid layers with 50% rate. Gradient training is controlled by Adamax (see later). Minibatch size is 500.
(e) The cost function to be minimized has two components, the sum of two terms, a regularizing ℓ_2 norm for the weights and the binary cross-entropy cost on the outputs of the network. This latter takes the average of all cross entropies in the sample: assume that we have $1 \leq n \leq N$ samples with binary labels $y_n \in \{-1, +1\}$ and network responses $\hat{y}_n$ for all n. The loss function is

$$J(\hat{y}_1, \ldots, \hat{y}_n) = \frac{1}{N} \sum_{n=1}^{N} y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n). \tag{1}$$

where the proper range of estimation is warranted by the logistic function: $\hat{y}_n(z) = 1/(1 + e^{-\theta z})$ with z being the input to the n^{th} output unit and θ being a trainable parameter.

Early stopping Training stops early if performance over the validation set is not improving over, say m epochs. This way overfitting becomes less probable. In our case, $m = 5$ was chosen. F_1 score is the typical measure for face related estimations. However, F_1 score has discontinuities and constant regions making it dubious for gradient based methods. Our approach that aims to overcome this problem is the following: we computed the gradient for the binary cross-entropy, *but* used the F_1 score as performance measure in the validation step. This way, gradient descent was guided by the F_1 score itself. The high quality results that we reached with a relatively simple network may be partially due to this procedure.

Cross-validation All the results and methods reported on the "Train" set have been validated with a 10-fold, subject-independent cross-validation. In the other experiments we trained on the "Train" set and reported performance measures on the "Test" set, following the challenge protocol [26].

Details of the Backpropagation Algorithm. Beyond the advances of GPU technology and deep learning architectures, error backpropagation also underwent fast and efficient changes. We used one of the most recent methods called Adamax [18]. It is a version of the Adam algorithm, a first-order gradient-based optimization, designed for stochastic objective functions exploiting adaptive estimates of lower-order moments. Adam estimates the ℓ_2 norm of the current and past gradients. If the gradients are small, the step size is made larger; inverse proportionality is applied. Adamax generalizes the ℓ_2 norm to ℓ_p norm and suggests to take the $p \to \infty$ limit. For more details, see [18].

Applied Software. There are many implementations of deep learning, mostly based on Python or C++. For a comprehensive list of software tools, today,

the link `http://deeplearning.net/software_links/` is a good starting point. We used Lasagne, a lightweight library built on top of Theano. Theano (`http://deeplearning.net/software/theano`) has been developed by the Montreal Institute for Learning Algorithms. It is a symbolic expression compiler that works both on CPU and on GPU and it is written in Python.

3 Results

First, we evaluated the performance on the FERA Train set, employing a 10-fold, subject independent CV. According to Table 1, HoG based SVM is the best for AU14 and AU15, and performance is superior for AU15. The representation at around the decision surface seems superior for these AUs. For the other AUs, SNI based CNNs with single AU classification are better. Multi-label classification is somewhat worse for almost all AUs, but let us note that these evaluations are faster, time scales linearly with the number of AUs for the single AU case.

Table 1. F_1 measures on the FERA BP4D Train set with different classifiers (C), input features (IF) and output label (OL) structures. The input features are Histogram of Gradients (HOG), mosaic images (MI), and similarity normalized images (SNI). The output structures are either single- (S) or multi-label (M).

C	IF	OL	Action Units											
			1	2	4	6	7	10	12	14	15	17	23	Mean
SVM	HOG	S	0.44	0.29	0.45	0.77	0.75	0.81	0.87	**0.62**	**0.39**	0.58	0.41	0.58
CNN	MI	M	0.22	0.01	0.43	0.76	0.64	0.77	0.85	0.47	0.00	0.27	0.00	0.40
	SNI	S	**0.63**	**0.44**	**0.54**	**0.82**	**0.80**	**0.85**	**0.90**	0.58	0.27	**0.60**	**0.45**	**0.63**
		M	0.55	0.38	0.53	0.80	0.75	0.83	**0.90**	0.55	0.23	0.59	0.37	0.59

Table 2. Results on the FERA BP4D Test set with multi-label CNN and SNI. Performance measures include F_1 score, its skew normalized version ($F_1^{s.n.\cdot}$) [13], and area under ROC curve (AUC). The table shows the degree of skew (ratio of negative and positive labels) for each AU.

	Action Units											
	1	2	4	6	7	10	12	14	15	17	23	Mean
skew	15.31	20.02	12.23	2.11	0.66	1.01	1.37	0.98	11.78	6.81	6.78	7.19
F_1	0.26	0.23	0.27	0.76	0.75	0.8	0.8	0.64	0.26	0.38	0.3	0.50
$F_1^{s.n.\cdot}$	0.74	0.67	0.69	0.84	0.7	0.8	0.83	0.64	0.43	0.68	0.38	0.67
AUC	0.84	0.79	0.76	0.92	0.77	0.88	0.92	0.72	0.75	0.77	0.74	0.81

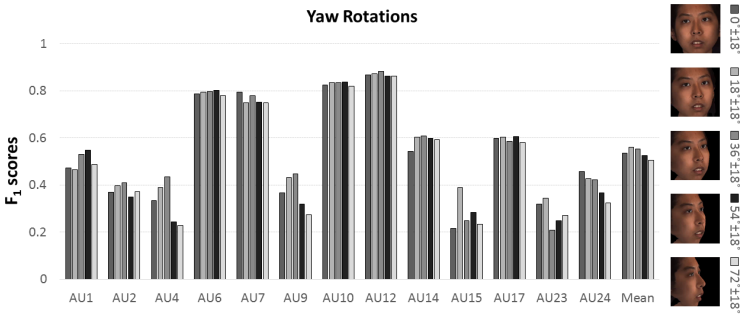

Fig. 3. F_1 measures as a function of yaw rotation on the augmented BP4D Train set, using the single-label classifier.

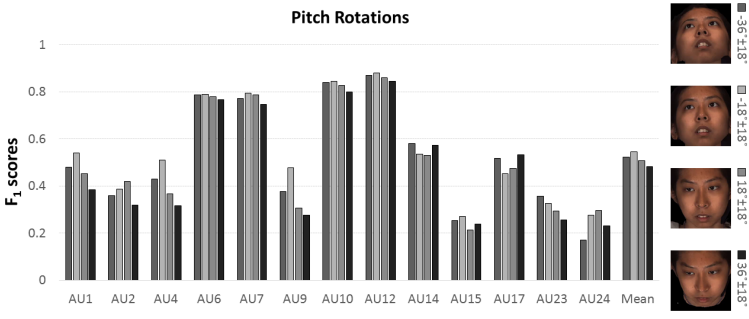

Fig. 4. F_1 measures as a function of pitch rotation on the augmented BP4D Train set, using the single-label classifier.

In the next experiment we trained the system on the FERA 2015 Train set, and tested it on the Test set. The AU base-rates are significantly different on these subsets [26] and F_1 score is attenuated by skewed distributions [13]. For this reason we report the degree of skew, F_1 score, its skew normalized version ($F_1^{s.n.}$) [13], and area under the receiver operating characteristic (ROC) curves. The AUC values are shown in Table 2 for the FERA BP4D test set, where skew parameters range between 1 and 20.

Head pose has three main angles, roll, yaw and pitch. Roll can be compensated in the frontal view by the normalization step. The case is more complex for non-frontal views. We studied yaw and pitch angles around the frontal view. Yaw is symmetric in this case and we show data for $(-18°, +18°)$ ranges around head poses 0, 18, 36, 54 and 72 degrees that covers the full frontal–to–profile view range. Angle dependence is relatively large for AU4, AU15, and AU23, but the mean F_1 score is a weak function of the head pose angle (Fig 3).

We studied the asymmetric pitch around the frontal view for $(-18°, +18°)$ ranges around -36, -18, +18, and +36 degrees. The mean F_1 score is also a weak function of the pitch angle. AU1, AU4, and AU23 are affected by this angle more

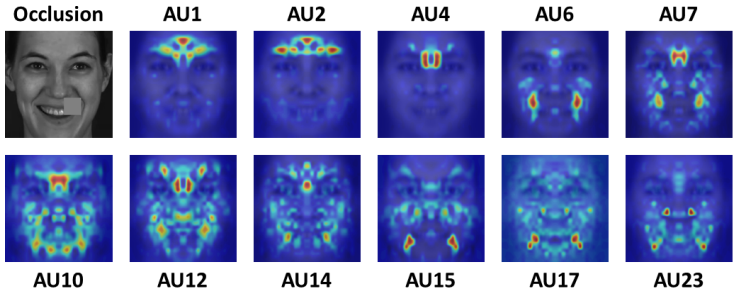

Fig. 5. Occlusion Sensitivity Maps [30]. (a): cropped 256×256 pixel images are covered by uniform grey 21×21 pixel patches at around pixels of a 32×32 pixel grid uniformly placed over the image. (b)-(n): the modified images are evaluated for binary cross-entropy performance. Performance is color coded at the central pixel of the patch and the 32×32 image is depicted for the different AUs.

strongly than the other AUs (AU2, AU6, AU7, AU10, AU12, AU14, AU15, and AU17), see, Fig. 4.

Occlusion sensitivity maps [30] were generated for the different action units. We used 200 images for each subject, giving 8,200 images for the map generations. At around certain pixels, the pixels of the 21×21 sized patches were set to 0.5, the middle of the normalized range, $[0, 1]$. Central pixels were laid uniformly on each image at $32 \times 32 = 1,024$ positions. The modified images, more than 8 million, were then tried on the trained network for each AU and the binary cross-entropy measure was computed. Results are shown in condensed form, the value is color coded on a 32×32 occlusion sensitivity map in Fig. 5.

We end the result section by comparing our results with the most recent ones reported in the literature, the Local Gabor Binary Pattern (LGBP) [26], the geometric feature based deep network (GDNN) [12], the Discriminant Laplacian Embedding (DLE) [29], Deep Learning with Global Contrast Normalization (DL) [11], and the Convolutional and Bi-directional Memory Neural Networks (CRML) [12] methods. DLE wins for AU15, CMLR is the best for AU10, and AU 14, and DL performs the best for AU1 and AU2. Our architecture comes first for the other AUs, with one exception, the single label case wins. Since the multi-label case is considerably harder, we suspect that better training can improve the results further, e.g., by adding noise to the input on top of the dropout and/or increasing the database.

The single label case produced the best mean value. A special difference between the CMRL method and ours is that we can work on single images, whereas CMRL requires frame series. Furthermore, the inclusion of temporal information should improve performance for our case, too.

Table 3. Comparison of the single-label (SL) and multi-label (ML) version with other methods in the literature.

	Action Units											
	1	2	4	6	7	10	12	14	15	17	23	Mean
LGBP [26]	0.18	0.16	0.22	0.67	0.75	0.8	0.79	0.67	0.14	0.24	0.24	0.44
GDNN [12]	0.33	0.25	0.21	0.64	0.79	0.8	0.78	0.68	0.19	0.28	0.33	0.48
DLE [29]	0.25	0.17	0.28	0.73	0.78	0.8	0.78	0.62	**0.35**	0.38	0.44	0.51
DL [11]	**0.40**	**0.35**	0.32	0.72	0.78	0.80	0.79	0.68	0.24	0.37	0.31	0.52
CRML [12]	0.28	0.28	0.34	0.7	0.78	**0.81**	0.78	**0.75**	0.2	0.36	0.41	0.52
this (ML)	0.26	0.23	0.27	**0.76**	0.75	0.8	0.8	0.64	0.26	0.38	0.3	0.50
this (SL)	0.34	0.21	**0.40**	0.74	**0.82**	**0.81**	**0.83**	0.73	0.25	**0.44**	**0.47**	**0.55**

4 Discussion

Recent progress in convolutional neural networks (e.g., [30, 22, 28, 12] and see also the general review [24] and the cited references therein) shows that Deep Neural Networks, including CNNs are flexible enough to compete with hand-crafted features, such as HoG, SIFT, Gabor filters, LBP, among many others. The adaptivity of the CNN structure tunes the convolutional layers of the CNN to the database according to the statistics of the data. The fully connected layers, on the other hand, serve to collect, combine and exclude certain portions of the image.

The big progress is due to the tricks of avoiding local minima during the training procedure and the collection of such methods keeps increasing. We used high dropout rates, early stopping, and rectified linear units to overcome the danger of falling into one of the local minima too early during training. We have no doubt that this quickly developing field will come up superior solutions and performance will increase further. The maturation of deep learning neural network technologies offer great promises in the field of facial expression estimations.

The success of our relatively small network is most probably due to another additional trick; we combined gradient descent with selective methods. Although the contribution of this trick that we detail below is hard to grab quantitatively, we should note that we used no binary mask [12], no temporal extensions [12, 16], known to have a considerable impact on performance.

The problem of optimization lies in the dubious F_1 score, which is not a good cost function, due to its discontinuities and flat, constant regions. Instead, a closely related quantity, the binary cross-entropy is preferred for gradient computations. Selection does not require well behaving, smooth costs and it can be introduced into the procedure at the validation step that guides early stopping. If performance is not improving on the validation set for a number of steps, in spite of the fact that it still does on the training set, then the gradient procedure should be stopped, since a local minimum of the training set is approached. Upon early stopping a new minibatch can be used for improving the performance.

This validation step can serve the selective process if gradient descent is stopped according to a different measure instead of the cost function. In our case, this measure was the F_1 score. It should be noted that the ideal values for the F_1 score and the binary cross entropy are the same, although they are rarely reached for real problems.

Clearly, special procedures, such as binary masks and temporal information should improve our results further, alike to performance increases in the studies mentioned previously.

Our main finding is that performance is a weak function of the head pose for CNNs and it remains high for a broad variety of angles. This opens the possibility of many real-life applications from cyber-physical systems with human in the loop, including smart factories, medical cyber-physical systems, independent living situation among many others. Furthermore, insights, sometimes of diagnostic value can be gained for affective disorders, addiction, and social relations. The progress of GPU technology will provide further gains in evaluation time that will decrease training time and evaluation frequency, too. The single-label version of our system runs at 58 FPS, while the multi-label version reaches over 600 FPS on a Titan X GPU.

Real life applications may require "in the wild" databases. This point remains to be seen.

5 Conclusions

Recent progress in deep learning technology and the availability of high quality databases enabled powerful learning methods to enter the field of face processing. We used these deep learning methods and the BP4D database for training an architecture for action unit recognition. Our results surpassed the state-of-the-art for images and could be further improved if temporal information is available. The main result is that angle dependence is minor, a large yaw and pitch range can be covered without considerable deterioration in performance. In turn, relevant applications from human-computer interaction to psychiatric interviews may gain momentum by applying such tools.

6 Acknowledgements

This work was supported in part by US National Institutes of Health grant MH096951 to the University of Pittsburgh and by US National Science Foundation grants CNS-1205664 and CNS-1205195 to the University of Pittsburgh and the University of Binghamton. Neither agency was involved in the planning or writing of the work.

References

1. Ambadar, Z., Cohn, J.F., Reed, L.I.: All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. Journal of nonverbal behavior 33(1), 17–34 (2009)

2. Bauer, A., Wollherr, D., Buss, M.: Human–robot collaboration: a survey. International Journal of Humanoid Robotics 5(01), 47–66 (2008)
3. Ekman, P., Rosenberg, E.L.: What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS). Oxford University Press, New York, NY, 2nd edn. (2005)
4. Fairbairn, C.E., Sayette, M.A., Levine, J.M., Cohn, J.F., Creswell, K.G.: The effects of alcohol on the emotional displays of whites in interracial groups. Emotion 13(3), 468–477 (2013)
5. Fridlund, A.J.: Human facial expression: An evolutionary view. Academic Press (1994)
6. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics 36(4), 193–202 (1980)
7. Girard, J.M., Cohn, J.F., Jeni, L.A., Sayette, M.A., De La Torre, F.: Spontaneous facial expression in unscripted social interactions can be measured automatically. Behavior Research Methods pp. 1–12 (2014), `articles/Girard14BRM.pdf`
8. Girard, J.M., Cohn, J.F., Mahoor, M.H., Mavadati, S.M., Hammal, Z., Rosenwald, D.P.: Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. Image and Vision Computing 32(10), 641–647 (2014)
9. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: International Conference on Artificial Intelligence and Statistics. pp. 315–323 (2011)
10. Griffin, K.M., Sayette, M.A.: Facial reactions to smoking cues relate to ambivalence about smoking. Psychology of Addictive Behaviors 22(4), 551 (2008)
11. Gudi, A., Tasli, H.E., den Uyl, T.M., Maroulis, A.: Deep learning based facs action unit occurrence and intensity estimation. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. vol. 6, pp. 1–5. IEEE (2015)
12. Jaiswal, S., Valstar, M.F.: Deep learning the dynamic appearance and shape of facial action units. In: Applications of Computer Vision, Winter Conference on, (WACV). IEEE (March 2015)
13. Jeni, L.A., Cohn, J.F., De La Torre, F.: Facing imbalanced data–recommendations for the use of performance metrics. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII) (2013)
14. Jeni, L.A., Cohn, J.F., Kanade, T.: Dense 3d face alignment from 2d videos in real-time. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) (2015), `http://zface.org`
15. Jeni, L.A., Cohn, J.F., Kanade, T.: Dense 3d face alignment from 2d video for real-time use. Image and Vision Computing (2016)
16. Jeni, L.A., Lőrincz, A., Szabó, Z., Cohn, J.F., Kanade, T.: Spatio-temporal event classification using time-series kernel based structured sparsity. In: Computer Vision–ECCV 2014, pp. 135–150. Springer (2014)
17. Keltner, D., MOffitt, T.E., Stouthamer-Loeber, M.: Facial expressions of emotion and psychopathology in adolescent boys. Journal of Abnormal Psychology 104(4), 644 (1995)
18. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (2015)
20. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)

21. Li, H., Sun, J., Wang, D., Xu, Z., Chen, L.: Deep representation of facial geometric and photometric attributes for automatic 3d facial expression recognition. arXiv preprint arXiv:1511.03015 (2015)
22. Liu, M., Li, S., Shan, S., Chen, X.: Au-inspired deep networks for facial expression feature learning. Neurocomputing 159, 126–136 (2015)
23. McDuff, D., el Kaliouby, R., Demirdjian, D., Picard, R.: Predicting online media effectiveness based on smile responses gathered over the internet. In: International Conference on Automatic Face and Gesture Recognition (2013)
24. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Networks 61, 85–117 (2015)
25. Szirtes, G., Szolgay, D., Utasi, A.: Facing reality: An industrial view on large scale use of facial expression analysis. Proceedings of the Emotion Recognition in the Wild Challenge and Workshop pp. 1–8 (2013)
26. Valstar, M.F., Almaev, T., Girard, J.M., McKeown, G., Mehu, M., Yin, L., Pantic, M., Cohn, J.F.: Fera 2015-second facial expression recognition and analysis challenge. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. vol. 6, pp. 1–8. IEEE (2015)
27. Vicente, F., Huang, Z., Xiong, X., De la Torre, F., Zhang, W., Levi, D.: Driver gaze tracking and eyes off the road detection system. IEEE Transactions on Intelligent Transportation Systems 16(4), 2014–2027 (2015)
28. Xu, M., Cheng, W., Zhao, Q., Ma, L., Xu, F.: Facial expression recognition based on transfer learning from deep convolutional networks. In: Natural Computation (ICNC), 2015 11th International Conference on. pp. 702–708. IEEE (2015)
29. Yuce, A., Gao, H., Thiran, J.P.: Discriminant multi-label manifold embedding for facial action unit detection. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. vol. 6, pp. 1–6 (2015)
30. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer vision–ECCV 2014, pp. 818–833. Springer (2014)
31. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing 32(10), 692–706 (2014)