Facial Expression Recognition using Convolutional Neural Networks: State of the Art

Christopher Pramerdorfer, Martin Kampel
Computer Vision Lab, TU Wien
Vienna, Austria
Email: {cpramer,kampel}@caa.tuwien.ac.at

***Abstract*—The ability to recognize facial expressions automatically enables novel applications in human-computer interaction and other areas. Consequently, there has been active research in this field, with several recent works utilizing Convolutional Neural Networks (CNNs) for feature extraction and inference. These works differ significantly in terms of CNN architectures and other factors. Based on the reported results alone, the performance impact of these factors is unclear. In this paper, we review the state of the art in image-based facial expression recognition using CNNs and highlight algorithmic differences and their performance impact. On this basis, we identify existing bottlenecks and consequently directions for advancing this research field. Furthermore, we demonstrate that overcoming one of these bottlenecks – the comparatively basic architectures of the CNNs utilized in this field – leads to a substantial performance increase. By forming an ensemble of modern deep CNNs, we obtain a FER2013 test accuracy of 75.2%, outperforming previous works without requiring auxiliary training data or face registration.**

I. Introduction

Being able to recognize facial expressions is key to non-verbal communication between humans, and the production, perception, and interpretation of facial expressions have been widely studied [1]. Due to the important role of facial expressions in human interaction, the ability to perform *Facial Expression Recognition* (FER) automatically via computer vision enables a range of novel applications in fields such as human-computer interaction and data analytics [2].

Consequently, FER has been widely studied and significant progress has been made in this field. In fact, recognizing basic expressions under controlled conditions (e.g. frontal faces and posed expressions) can now be considered a solved problem [1]. The term *basic expression* refers to a set of expressions that convey universal emotions, usually anger, disgust, fear, happiness, sadness, and surprise. Recognizing such expressions under naturalistic conditions is, however, more challenging. This is due to variations in head pose and illumination, occlusions, and the fact that unposed expressions are often subtle, as Fig. 1 illustrates. Reliable FER under naturalistic conditions is mandatory in the aforementioned applications, yet still an unsolved problem [1], [2].

Convolutional Neural Networks (CNNs) have the potential to overcome these challenges. CNNs have enabled significant performance improvements in related tasks (e.g. [4]–[6]), and several recent works on FER successfully utilize CNNs for feature extraction and inference (e.g. [7]–[9]). These works differ significantly in terms of CNN architecture, preprocessing, as well as training and test protocols, factors that all affect performance. It is therefore not possible to assess the impact of the CNN architecture and other factors based on the reported results alone. Being able to do so is, however, required in order to be able to identify existing bottlenecks in CNN-based FER, and consequently for improving FER performance.

Fig. 1. Example images from the FER2013 dataset [3], illustrating variabilities in illumination, age, pose, expression intensity, and occlusions that occur under realistic conditions. Images in the same column depict identical expressions, namely anger, disgust, fear, happiness, sadness, surprise, as well as neutral.

The aim of this paper is to shed light on this matter by reviewing existing CNN-based FER methods and highlighting their differences (Section II), as well as comparing the utilized CNN architectures empirically under consistent settings (Section III). On this basis, we identify existing bottlenecks and directions for improving FER performance. Finally, we confirm empirically that overcoming one such bottleneck improves performance substantially, demonstrating that modern deep CNNs achieve competitive results without auxiliary data or face registration (Section IV). An ensemble of such CNNs obtains a FER2013 [3] test accuracy of 75.2%, outperforming existing CNN-based FER methods.

In this paper, we consider the task of predicting basic expressions from single images using CNNs. For more general surveys, we refer to [1], [2]. We note that it is straight-forward to adapt image-based methods to support image sequences by integrating per-frame results using graphical models. The conclusions drawn in this paper are thus relevant for sequence-based FER as well.

II. State of the Art in CNN-Based FER

We review six state-of-the-art methods for CNN-based FER, highlight methodological differences, and discuss the reported

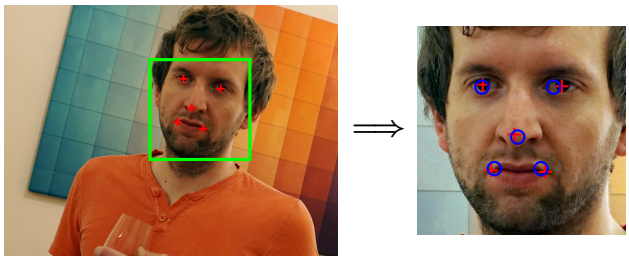

Fig. 2. Illustration of a standard preprocessing pipeline, which involves face detection (green square), facial landmark detection (red crosses), registration to reference landmarks (blue circles), and illumination correction.

performances. Most of these methods were evaluated on several databases in the original papers, the most common dataset being FER2013 [3]. For consistency, we study the methods as they were used for this dataset, and summarize and discuss the reported performances on this dataset.

FER2013 is a large, publicly available FER dataset consisting of 35,887 face crops. The dataset is challenging as the depicted faces vary significantly in terms of person age, face pose, and other factors (Fig. 1), reflecting realistic conditions. The dataset is split into training, validation, and test sets with 28,709, 3,589, and 3,589 samples, respectively. Basic expression labels are provided for all samples. All images are grayscale and have a resolution of 48 by 48 pixels. The human accuracy on this dataset is around 65.5% [3].

A. Overview

Image-based FER under naturalistic conditions has been an active research field for years, and several public challenges have been held to promote progress in this field. One such challenge was FER2013 [3], which was won by one of the first CNN-based FER methods [7]. The method uses an ensemble of CNNs trained to minimize the squared hinge loss.

In a more recent work, Yu and Zhang [8] also utilize an ensemble of CNNs, and employ data augmentation at both training and test time in order to improve performance. Instead of performing ensemble voting via uniform averaging as in [7], ensemble predictions are integrated via weighted averaging with learned weights. The method ranked second in the recent EmotiW2015 challenge [10].

The winner [11] of this challenge employs a large committee of CNNs. Certain properties of the individual networks (e.g. input preprocessing and receptive field size) vary in order to obtain more diverse models. The ensemble predictions are integrated in a hierarchical fashion, with network weights assigned according to validation set performance.

Mollahosseini et al. [12] trained a single CNN based on the Inception architecture [13] on data compiled from multiple posed and naturalistic datasets in an effort to obtain a model that generalizes well across datasets.

In [14] Zhang et al. present a method for inferring social relation traits from images using a Siamese network. In order to increase the amount of available training data, they utilize multiple datasets with heterogeneous labels. The authors present a patch-based registration and feature extraction technique, and perform feature integration via early fusion.

A recent work by Kim et al. [9] shows that it is beneficial to use both unregistered and registered versions of a given face image during both training and testing. In order to prevent registration errors from affecting FER performance, registration is performed selectively based on the results of facial landmark detection. The authors show that registration can also be performed by deep networks, and that utilizing pose information captured by such networks leads to a small increase in FER performance (about 0.4%).

TABLE I
PREPROCESSING OPERATIONS. FD: FACE DETECTION, LM: FACIAL LANDMARK EXTRACTION., HISTEQ: HISTOGRAM EQUALIZATION, LPF: LINEAR PLANE FITTING.

Method	FD	LM	Registration	Illumination
[7]	no	no	no	normalize
[11]	several	[15]	rigid (LM)	several
[8]	several	no	no	histeq, lpf
[12]	no	[15]	affine (LM)	no
[14]	no	[15]	indirect	no
[9]	no	[15]	rigid (LM)	several

B. Methodological Differences

In order to highlight the methodological differences between these works, we break down each method into the three components (i) preprocessing, (ii) CNN architecture, and (iii) CNN training and inference.

1) Preprocessing: Preprocessing entails operations that are applied once to each image. This typically includes face detection, face registration to compensate for pose variations, and means for correcting for illumination variations. Fig. 2 illustrates these steps.

Table I summarizes the preprocessing steps of every method. Only [11] and [8] perform face detection; all other methods rely on face crops provided by the datasets.

Face registration is common, with rigid or affine transformations based on extracted facial landmark locations being the most popular approach. This form of registration has the potential to improve FER performance [9] provided that landmarks can be detected reliably. However, this is not always the case in practice due to challenging face poses and/or partial occlusions [9]. There are different approaches to account for this problem; [11] perform landmark detection on multiple versions of a given face image and utilize the detections with the highest detector confidence. [9] perform alignment only if this confidence exceeds a threshold.

The majority of existing methods uses some form of illumination correction. In [7] every image is normalized to have a mean of 0 and a norm of 100. [8] employ histogram equalization and linear plane fitting. In [11] and [9] the methods vary between individual CNNs in the ensembles.

2) CNN Architecture: Table II compares the utilized CNN architectures and their depths (number of layers with weights)

TABLE II
CNN ARCHITECTURES. C, P, N, I, AND F STANDS FOR CONVOLUTIONAL, POOLING, RESPONSE-NORMALIZATION, INCEPTION, AND FULLY CONNECTED LAYERS, RESPECTIVELY.

Method	Architecture	Depth	Parameters
[7]	CPCPFF	4	12.0 m
[11]	CPCPCPFF	5	4.8 m
[8]	PCCPCCPCFFF	8	6.2 m
[12]	CPCPIIPIPFFF	11	7.3 m
[14]	CPNCPNCPCFF	6	21.3 m
[9]	CPCPCPFF	5	2.4 m

TABLE III
DIFFERENCES IN TERMS OF CNN TRAINING AND INFERENCE. AD: ADDITIONAL TRAINING DATA, AF: ADDITIONAL FEATURES, +: DATA AUGMENTATION, S,A: SIMILARITY/AFFINE TRANSFORM, T: TRANSLATION, M: HORIZONTAL MIRRORING, REG: FACE REGISTRATION.

Method	AD	AF	+ Train	+ Test	Ensemble
[7]	no	no	S,M	–	average
[11]	no	no	T,M	–	hierarchy
[8]	no	no	A,M	A	weighted
[12]	yes	no	ten-crop	–	–
[14]	yes	yes	–	–	–
[9]	no	yes	T,M,reg	ten-crop,reg	average

and parameter counts. The counts were calculated assuming single-channel input images of size 48 by 48 pixels (the size of images in the FER2013 dataset). In some works, the CNNs operate on images of different sizes.

The table highlights that the individual architectures vary significantly in terms of layer composition, depth, and number of parameters. Most architectures are shallow compared to architectures in related fields [5], [13], [16]. As the corresponding papers lack details on architecture selection, the reason for this discrepancy is unknown.

The small size of available FER datasets such as FER2013 is not the limiting factor. First, deeper networks do not necessarily have more parameters, as shown in Table II. Second, deeper networks impose a stronger prior on the structure of the learned decision function, and this prior effectively combats overfitting [17]. Third, modern deep CNNs achieve impressive results on datasets with a similar size, such as CIFAR10 [5]. A possible explanation is that CNNs do not have to be as deep for FER; [18] shows that a CNN with depth 5 is already able to learn discriminative high-level features. We postpone further discussions on this matter to Section III.

3) CNN Training and Inference: Table III highlights the differences in terms of CNN training and inference. Of the six works compared, four use only the FER2013 training set for CNN training. [12] and [14] instead train on an union of seven and three datasets, respectively, in order to compensate for the fact that available FER image datasets are comparatively small.

[14] and [9] use additional features. In [14] a vector of HoG features is computed from face patches and processed by the first fully connected layer of the CNN (early fusion). In [9] the these features encode face pose information, and classifiers are trained to perform FER on this basis. Integration is performed via ensemble voting (late fusion).

All works except [14] employ data augmentation during training in order to increase the amount of available data. Most works use standard augmentation methods that are not specific to FER, including horizontal mirroring and random cropping. The exception is [9], which additionally augments the training set by a registered version of every image.

In [8] and [9] augmentation is also performed at test time. In the former work, multiple perturbed versions of each test image are generated by applying affine transformations randomly, and the CNN output probabilities are averaged. The latter work follows the same approach but uses ten-crops (center and corner crops and mirrored versions) of a given image before and after face registration.

Most works use an ensemble of CNNs, whose predictions are integrated via different forms of averaging.

C. Reported Results on FER2013

Fig. 3 compares the reported FER2013 test accuracies. Based on the methodological differences and these results, we draw the following remarks.

The three best-performing works use comparatively shallow CNNs (depths of 5 and 6). The work utilizing the deepest and most modern (in terms of layer types and arrangement) CNN [12] performs worst on this dataset. However, a direct comparison of the results is not possible because the CNN was trained on a superset of FER2013; in case of [12], this presumably has a negative impact on FER2013 test performance. On the other hand, Zhang et al. [14] demonstrate that utilizing additional training data in a way that accounts for dataset bias improves the performance on FER2013.

The three best-performing methods use face registration, suggesting that registration is beneficial even under challenging conditions (according to [14], facial landmark extraction is inaccurate for about 15% of images in the FER dataset).

Data augmentation and ensemble voting are important in order to improve generalization performance. The best-performing work that was trained on FER2013 alone uses the most comprehensive form of data augmentation during both training and testing. Ensemble voting reportedly improves the test accuracy by 2-3% [8], [9].

III. EMPIRICAL COMPARISON

Im summary, Fig. 3 shows that the best-performing FER methods utilize CNNs that are shallow and basic (in terms of layer types and arrangement) compared to the state of the art in related fields [5], [6], contradicting the general trend towards deeper and deeper networks.

This, however, does not mean that such CNNs are inapplicable for FER because the CNN architecture is one of many factors that influence FER performance. In order to obtain more information on this matter, we perform an empirical comparison of the utilized CNN architectures.

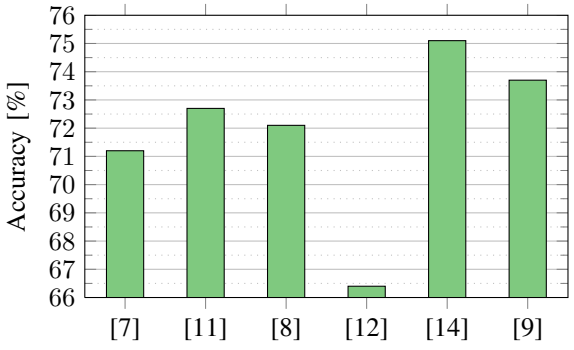

Fig. 3. Reported FER2013 test results.

A. Experiments

We train and test all CNN architectures utilized in the compared works using the same protocols.

1) CNN Architectures: We test all architectures as they are described in the corresponding papers, apart from the following differences. We add batch normalization [19] layers after every convolutional and fully connected (fc) layer for robustness to suboptimal network initialization. Furthermore, we add a dropout layer [20] after the first fc layer.

[11] uses an ensemble of CNNs with different receptive fields and numbers of neurons in the first fc layer. We use the configuration with 3×3 receptive fields and 2,048 neurons.

2) Dataset and Preprocessing: In order to enable comparison with reported results, we perform all experiments on the FER2013 dataset [3], adhering to the official training, validation, and test sets. We use the face crops as provided by the dataset and employ histogram equalization for illumination correction. This follows subtraction and division by the mean and standard deviation over all training pixels, respectively.

We do not perform landmark-based registration for two reasons. First, this prevents registration errors from affecting the results. Second, this forces the CNNs to learn to compensate for pose variations, potentially leading to more general models.

3) CNN Training and Inference: We train every architecture for up to 300 epochs, optimizing the cross-entropy loss using stochastic gradient descent with a momentum of 0.9. The initial learning rate, batch size, and weight decay are fixed at 0.1, 128, and 0.0001, respectively. The learning rate is halved if the validation accuracy does not improve for 10 epochs.

For training data augmentation we use horizontal mirroring and random crops of size 48 by 48 pixels after zero-padding (as the input images are already of this size). We ensure that all CNNs see the same training samples (after augmentation) in the same order, thereby enabling a fair comparison.

As the individual architectures were designed for training on larger datasets and/or other forms of data augmentation, training using the same regularization parameters would be unfair. In order to account for this, we perform a grid-search to find out an optimal dropout rate for every architecture.

Finally, the best model obtained for each architecture in terms of validation accuracy is tested on the test set using standard ten-crop oversampling [5].

4) Feature Comparison: Furthermore, we empirically compare the quality of the features learned by the models with the highest validation accuracy. This is accomplished by replacing the fc backends of the trained CNNs with a two-layer MLP with 1,024 hidden units, which is then trained using the above protocol. All other parameters are fixed, effectively causing the MLP to learn to perform FER using the features extracted by the pretrained frontend of the network.

Doing so allows a more direct comparison of the results because the backends of the resulting models are no longer different. With different backends, a more powerful backend could mask limitations in the learned representations. Fixing the backend enables us to study the impact of the network depth on the capabilities of the learned representations.

B. Results and Discussion

Fig. 4 summarizes the results. All results, except those for [12], are lower than the reported ones (Fig. 3). The main reason for this is the lack of ensemble voting, as most results are comparable to reported results of single networks [8], [9].

In some cases, the differences cannot be explained by the lack of ensemble voting; in case of [14], the reason for the lower performance is the lack of auxiliary training data. The reported accuracy in [12] is lower than the measured accuracy. This is also explained by the additional training data, which in this case has a negative effect on FER2013 performance. This shows that auxiliary training data has the potential to significantly improve FER performance, provided that care is taken in order to address dataset bias.

In case of [8], the measured accuracy is about 3% lower than the reported one using a single CNN. We tested both stochastic pooling (as used in [8]) and max pooling, but in both cases were unable to reach the reported accuracy.

Overally, shallower CNN architectures again perform better than deeper ones (cf. Table II). This also applies to the learned features. This, however, does not confirm that modern deep networks are not suitable for FER; there is only one architecture in the comparison that qualifies as such [12], and some architectural choices of this network are questionable (initial convolution with a 7×7 receptive field, which appears too large given the input resolution, and a wide backend in the form of a three-layer MLP with 4,096 units in the first layer).

On the contrary, we postulate that modern deep networks can outperform the shallow architectures of current works, based on findings in related research fields [5], [6]. In Section IV we perform experiments to confirm this hypothesis.

C. Current Bottlenecks

In this section, we highlight major bottlenecks in CNN-based FER based on the reported and measured results. We postulate that overcoming these bottlenecks will lead to substantial improvements in FER performance.

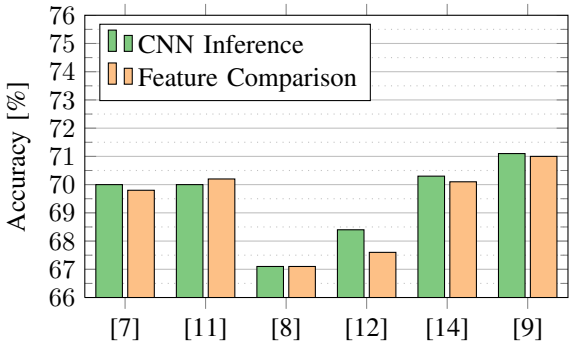

Fig. 4. Ten-crop test results of the different network architectures (green), and when using pretrained frontends as feature extractors (orange).

The CNN architectures used are basic and shallow compared to state-of-the-art architectures in related fields [5], [13], [16].

Most works use general data augmentation techniques such as random crops and mirroring, which are not optimized for the task at hand. Kim et al. [9] show that face-aware data augmentation via face registration improves performance. This approach can be extended, for instance by data augmentation via frontalized samples [21], or by synthesizing faces in various poses via 3D face pose estimation [22], [23]. Such data augmentation techniques, utilized for both training and inference, have the potential to effectively compensate for the limited size of current FER datasets.

For the same reason, training on a combination of multiple datasets can lead to significant improvements, provided that dataset bias is accounted for. [14] show that this is beneficial even with datasets with heterogeneous labels.

Lastly, we put forward that the biggest bottleneck that currently hinders FER performance is the fact that there is no publicly available dataset that is large by current deep learning standards. The introduction of datasets with hundreds of thousands or millions of images has enabled significant performance gains in related research fields such as face recognition [6], [16], [22]. In contrast, FER2013, one of the largest FER image datasets available, has only 35,887 images. Compiling a large FER dataset is a laborious task due to the challenging annotation process; assigning correct expression labels in presence of subtle expressions, partial occlusions, and pose variations is a challenging task for humans [1].

IV. Deep CNNs for FER

In this section, we confirm experimentally that overcoming one of these bottlenecks, the comparatively shallow and basic CNN architectures of current FER methods, leads to a substantial improvement in accuracy on FER2013.

A. Experiments

In order to enable a fair comparison with the results in Section III, we use the exact same dataset, preprocessing, as well as training and testing protocols.

TABLE IV
Tested deep architectures and their ten-crop test accuracy on FER2013. 3R means group of three residual blocks.

Name	Architecture	Depth	Parameters	Accuracy
VGG	CCPCCPCCPCCPFF	10	1.8 m	72.7%
Inception	CIPIIPIIPIIPF	16	1.2 m	71.6%
ResNet	3R4R6R3RPF	33	5.3 m	72.4%

1) CNN Architectures: We consider three CNNs whose architectures are summarized in Table IV. All are inspired by current state-of-the-art architectures in related fields:

VGG. An architecture similar to VGG-B [24] but with one CCP block less. We also use dropout after each such block (this improved the validation accuracy by around 1%). For consistency with Section III, the backend consists of a single hidden layer with 1024 units.

Inception. An architecture similar to GoogLeNet [13], but with a more consistent structure and without initial strided convolutions or pooling (the input images are already small enough). The net uses a consistent distribution of feature map sizes in a given Inception layer that is based on the number of 3×3 features maps n; the $1 \times 1, 3 \times 3$ reduce, 5×5 reduce, 5×5, and pool projection layers have $¾n$, $½n$, $⅛n$, $¼n$, and $¼n$ feature maps, respectively. n is initialized to 32 and increased by 32 after every Inception layer.

ResNet. Our architecture is identical to the 34-layer ResNet from [5], but without the initial CP block. Our network is also more narrow, having 256 feature maps in the final residual group to reduce the number of parameters. We use dropout after the final pooling layer.

VGG and Inception have less parameters than any of the architectures used in the pertinent literature, despite being significantly deeper (cf. Tables IV. and II). Even the very deep ResNet has fewer parameters than most of these architectures.

We did not specifically search for architectures that perform well on FER2013. Our goal is to confirm that modern deep architectures generally perform well, not to obtain the absolute best accuracies on this dataset.

2) CNN Ensembles: In order to demonstrate the potential of an ensemble of such deep CNNs, we perform an exhaustive search to identify optimal ensembles of up to 8 models in terms of FER2013 validation accuracy.

B. Results and Discussion

The test accuracies of the individual models with the best validation accuracies are given in Table IV. The best modern deep model outperforms the best shallow model by almost 2% under identical conditions (cf. Fig. 4). All considered architectures outperform the best shallow model, including Inception, which has only half as many parameters. These results confirm that utilizing modern deep architectures has the potential to substantially improve FER performance.

Our individual CNNs already perform competitively to previous works that utilize ensemble voting (Fig. 3). By forming an ensemble of 8 such CNNs, we achieve a FER2013 test

accuracy of 75.2%, performing comparably to the current best method we are aware of [14].

Our ensemble of deep models obtains state-of-the-art performance without utilizing additional training data or features, comprehensive data augmentation, or requiring face registration. By not requiring face registration, our FER method is conceptually simpler than previous methods and not affected by registration errors. We expect that utilizing auxiliary training data and comprehensive, FER-specific data augmentation would improve the performance further.

This paper has been studying the FER performance by means of the FER2013 dataset, which is the most common image dataset in CNN-based FER and one of the largest publicly available datasets in this field. (There are several video datasets that contain a much larger number of frames, but these frames are naturally highly correlated and the number of subjects in such datasets is small.) Still, results obtained on this and other FER datasets are only indicative of real-world FER performance due to dataset bias. This limitation applies not only to this study but to FER research in general.

V. Conclusions

In this paper we have reviewed the state of the art in CNN-based FER, highlighted key differences between the individual works, and compared and discussed their performance with a focus on the underling CNN architectures. On this basis, we have identified existing bottlenecks and consequently means for advancing the state of the art in this challenging research field. Furthermore, we have shown that overcoming one such bottleneck by employing modern deep CNNs leads to a significant improvement in FER2013 performance. Finally, we have demonstrated that an ensemble of such CNNs outperforms state of the art methods without the use of additional training data or requiring face registration.

We expect that overcoming the remaining bottlenecks identified in this paper will result in further substantial performance improvements. For the future, we plan to investigate ways for overcoming these bottlenecks, with a focus on FER-specific data augmentation. Furthermore, we will study the bias that affects FER2013 and other datasets, and investigate the possibility of creating a new, more comprehensive, and publicly available FER dataset.

References

[1] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 37, no. 6, pp. 1113–1133, 2015.

[2] M. V. B. Martinez, "Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition," in *Advances in Face Detection and Facial Image Analysis*. Springer, 2016, pp. 63–100.

[3] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[5] K. He, X. Zhang, haoqing Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR*, vol. 1512, 2015.

[6] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[7] Y. Tang, "Deep Learning using Support Vector Machines," in *International Conference on Machine Learning (ICML) Workshops*, 2013.

[8] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *ACM International Conference on Multimodal Interaction (MMI)*, 2015, pp. 435–442.

[9] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing Aligned and Non-Aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops*, 2016, pp. 48–57.

[10] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and Image Based Emotion Recognition Challenges in the Wild: EmotiW 2015," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 423–426.

[11] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces*, pp. 1–17, 2016.

[12] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going Deeper in Facial Expression Recognition using Deep Neural Networks," *CoRR*, vol. 1511, 2015.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[14] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang, "Learning Social Relation Traits from Face Images," in *Proc. IEEE Int. Conference on Computer Vision (ICCV)*, 2015, pp. 3631–3639.

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

[16] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[17] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. D. Shet, "Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks," *CoRR*, vol. 6082, 2013.

[18] P. Khorrami, T. Paine, and T. Huang, "Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?" in *IEEE International Conference on Computer Vision (ICCV) Workshops*, 2015, pp. 19–27.

[19] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *CoRR*, vol. 1502, 2015.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective Face Frontalization in Unconstrained Images," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[22] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1701–1708.

[23] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3D Face Alignment from 2D Videos in Real-Time," in *IEEE Int. Conf. Automatic Face and Gesture Recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–8.

[24] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR*, vol. 1409, 2014.