Facial Expression Recognition using

Convolutional Neural Networks

Abstract

The ability to recognize facial expressions automat-

ically 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 bot-

tlenecks 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.

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 70.47%, outperforming previous works without requiring auxiliary training data or face registration.

INTRODUCTION

Being able to recognize facial expressions is key to non-verbal communication between humans, and the production,p erception, 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 ofthe 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.

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, weconﬁrm 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.

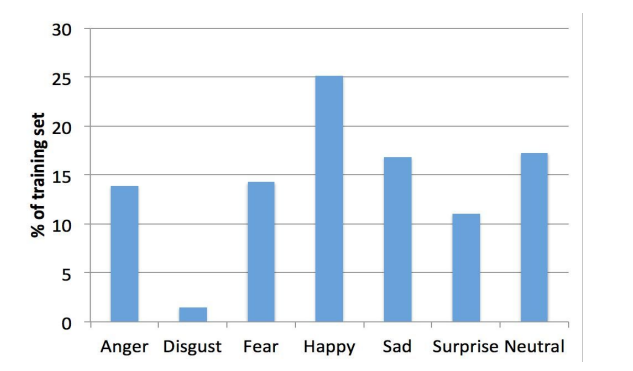
In this paper, we consider the task of predicting basic expressions from single images using CNNs. For more general surveys, we refer to [1], [2]. We note that it is straightforward to adapt image-based methods to support image sequences by integrating per-frame results using graphical models. The conclusions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

RELATED WORKS

In FER2013 challenge of the ICML 2013 Representation Learning [4], Tang introduced a CNN jointly learned with linear support vector machine (SVM) for facial expression recognition [11]. With a simple CNN and a SVM instead of softmax classiﬁer, the model outperformed the others and won the ﬁrst place in the challenge. Inspiring by the success of GoogLeNet [8], Mollahosseini et al. proposed an architecture containing four Inception modules [12]. However, their research cannot lead to a better performance on the FER2013dataset. In 2016, Zhou et al. proposed the multi-scale CNNs[13]. This model consists of three other networks with different input sizes. In addition, they used late fusion technique to get the ﬁnal classiﬁcation results. By combining multiple CNNs and modifying the loss function, Yu et al. obtained a higher accuracy compared to the previous approaches [14]. Similarly, Kim et al. introduced a multiple CNNs for facial expressionrecognition in the wild [15]. Another multi-scale CNN was proposed by Wang et al. [16]. In this work, the authors use the entire feature maps in the network for classiﬁcation. However, using all generated features without selection may reduce the overall performance due to trivial information in shallow layers of the network. To the best of our knowledge, this work is the current state-of-the-art method on the FER2013 dataset.

PROPOSED METHOD

**Dataset:** In this project we used FER-2013 dataset, which consists of about 37,000 well structured 48 × 48 pixel gray-scale images of faces. The images are processed in such a way that the faces are almost centered and each face occupies about the same amount of space in each image. Each image has to be categorized into one of the seven classes that express different facial emotions. These facial emotions have been categorized as: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, and 6=Neutral. Figure 1 depicts one example for each facial expression category. In addition to the image class number (a number between 0 and 6), the given images are divided into three different sets which are training, validation, and test sets. There are about 29,000 training images, 4,000 validation images, and 4,000 images for testing



**Figure 1: Distribution of different emotions across the FER-2013 dataset**

**CNN**

We evaluated both a variety of preprocessing techniques as well as several model architectures, ultimately developing a custom CNN model capable of attaining near-state-of-the-art accuracy of 70.47% on the FER-2013 test set. For preprocessing, we experimented with centering (i.e., subtracting mean) and scaling data. We found it generally helpful to subtract the mean found in the train distribution from all sets before training/evaluating. We also implemented data augmentation: we randomly rotate, shift, flip, crop, and sheer our training images. This yielded about a 10 p.p. increase in accuracies. We implemented several CNN architectures from papers applying emotion recognition to these and other datasets. Ultimately, what yielded the best performance was our custom developed CNN architecture (left). Analyzing error in neural networks is infamously difficult. We analyzed our error across different classes, as well as by visual inspection of images we classified correctly and incorrectly. One early observation was that we fail much more at certain emotions, and that we were failing to classify images where it was necessary to rely on fine details in the images (e.g., small facial features or curves). Due to this, we increased the number of layers and decreased filter sizes to increase the number of parameters in our network, which had a clear effect in allowing us to fit the dataset better. This led to some overfitting, which we addressed by using dropout, early stopping around 100 epochs, and augmenting our training set. Given this, we only start learning training set noise after achieving approx. 70% dev set accuracy; this is clear from plotting accuracy during training. This leaves us with some suggestions for future work, which largely focus on enabling increased parameterization of the network.

**Real-Time Classification**

We used OpenCV’s Haar cascades to detect and extract a face region from a webcam video feed, then classified it using our CNN model. We found it best to neither subtract the training mean nor normalize the pixels in the detected face region before classifying it. Real-time classification better exposed our model’s strengths: neutral, happy, surprised, and angry were generally well-detected. Illumination was a very important factor in the model’s performance. This suggests that out training set may not truthfully represent the distribution of lighting conditions.

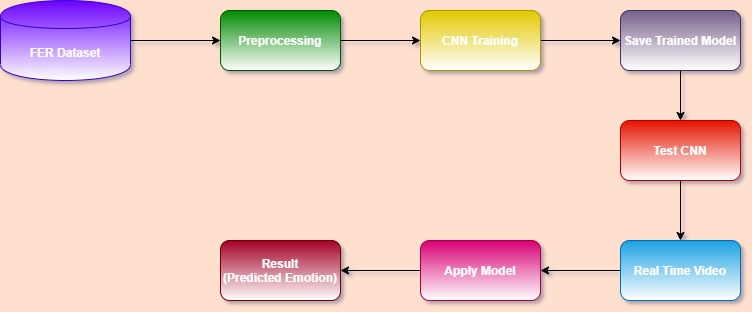


Figure:- System Work-Flow

FUTURE WORK/CONCLUSION

For continued work on this project, we believe there are two major areas of focus that would improve our real-time emotion recognition system. First, we suggest fine tuning the architecture of the CNN used for the model to fit perfectly with the problem at hand. Some examples of this fine tuning include finding and removing redundant parameters, adding new parameters in more useful places in the CNN’s structure, adjusting the learning rate decay schedule, adapting the location and probability of dropout and experimenting to find ideal stride sizes.

A second area of focus lies in adapting the datasets to more closely reflect real-time recognition conditions. For example, simulating low light conditions and “noisy” image backgrounds, could help the model become more accurate in real-time recognition. Additionally making sure that the distribution of models in the training dataset accurately reflects the distribution of subjects that the system will see when running in real-time.

REFRENCES

[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 Auto-matic 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 inrepresentation 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 Interna-

tional 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.

[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 Learningfor 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 Conferenceon 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 inthe 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, “Videoand Image Based Emotion Recognition Challenges in the Wild: EmotiW2015,” 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 expressionrecognition,” Journal on Multimodal User Interfaces, pp. 1–17, 2016.

[12] A. Mollahosseini, D. Chan, and M. H. Mahoor, “Going Deeper in FacialExpression 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.