SN Computer Science (2020) 1:79
https://doi.org/10.1007/s42979-020-0114-9

ORIGINAL RESEARCH

Detecting Affect States Using VGG16, ResNet50 and SE-ResNet50 Networks

Dhananjay Theckedath[1] · R. R. Sedamkar[2]

Abstract

Affect detection is a key component in developing intelligent human computer interface systems. State-of-the-art affect detection systems assume the availability of full un-occluded face images. This work uses convolutional neural networks with transfer learning to detect 7 basic affect states, viz. Angry, Contempt, Disgust, Fear, Happy and Sad. The paper compares three pre-trained networks, viz. VGG16, ResNet50 and a SE-ResNet50, in which a new architectural block of squeeze and excitation has been integrated with ResNet50. Modified VGG-16, ResNet50 and SE-ResNet50 networks are trained on images from the dataset, and the results are compared. We have been able to achieve validation accuracies of 96.8%, 99.47%, and 97.34% for VGG16, ResNet50 and SE-ResNet50, respectively. Apart from accuracy, the other performance matrices used in this work are precision and recall. Our evaluation, based on these performance matrices, shows that accurate affect detection is obtained from all the three networks with Resnet50 being the most accurate.

Keywords Convolutional neural network · Transfer learning · Affect states

Introduction

Affect describes the experience of feeling or emotion. It mediates a person's interaction with stimuli. The expression of emotion is achieved through a complex combination of information produced from the body and the brain. It has been believed that there exists a correlation between facial expressions and the mental state. There have been several attempts to quantify this relationship.

Effective affect analysis hugely depends upon accurate identification of facial features. A lot of ongoing research is in the area of applying artificial intelligence (AI) and deep learning algorithms to perform affect detection. AI aims to narrow the communicative gap between humans and computers by developing systems which are capable of recognizing and responding to the affect states [1].

Deep learning models extract relevant features from the data automatically and have shown to achieve state-of-the-art accuracy, sometimes exceeding human-level performance [2]. With the progress in deep learning algorithms and advancements in GPU technology, detecting of affect states has become a widely tackled research problem [3].

Convolutional neural network (CNN) is one of the most popular deep learning architectures. Due to its practical effectiveness, there has been an increased interest in deep learning. The interest in CNN started with AlexNet which won the ImageNet competition in 2012 [4].

CNN automatically detects important features without any human supervision and is also computationally efficient. The CNN consists of three basic layers, viz. convolution layer, sub-sampling layer and fully connected layer. Convolution and sub-sampling layers are repeated depending on the application and are finally connected to the fully connected layer. The convolution layer performs 2-dimensional convolution on the images from the dataset using a set of filter masks. A nonlinearity function called rectified linear unit (ReLU) is used which converts all negative values to zero. The sub-sampling layer performs down sampling of the convolved images, thereby reducing the spatial resolution.

This article is part of the topical collection "Advances in Computational Intelligence, Paradigms and Applications" guest edited by Young Lee and S. Meenakshi Sundaram.

✉ Dhananjay Theckedath
dhananjay.kishore@gmail.com

[1] Biomedical Engineering Department, Thadomal Shahani Engineering College, Mumbai, India

[2] Computer Engineering Department, Thakur College of Engineering and Technology, Mumbai, India

Published online: 11 March 2020

The most commonly used sub-sampling technique is Max pooling which uses the maximum value in a neighborhood to represent the neighborhood. The task of sub-sampling is to make the network more robust and invariant. Several repetitions of convolution with ReLU and sub-sampling are used. The final convolution layer is flattened and is connected to the fully connected layer. This is then given to the classification layer. A typical block diagram of CNN architecture is shown in Fig. 1 [5].

CNNs are known to give excellent results only when provided with enormous amounts of data, running into thousands of samples. Since this is not practically possible in most of the cases, transfer learning is used where a pre-trained network is trained using data from the new dataset. The main benefit of transfer learning is that it reduces the time taken to develop and train a model by reusing weights of already developed models.

This paper is organized as follows. Section 2 discusses related works. Section 3 explains the dataset that is used to carry out our experiments. In Sect. 4, we discuss the proposed methodology used in our work. Experimental results are provided in Sect. 5. This is followed by discussion in Sect. 6 and conclusion in Sect. 7.

Related Work

A number of studies have been reported in the literature which has focused on automatic detection of affect states. Most of the papers available use systematic measurements created by Ekman and Friesen [6], which have proved to be the standard for subsequent studies in affect analysis.

Arushi and Vivek [7] have used CNN to classify human faces. They developed a 5-layer CNN and used fractional Max pooling. They obtained a validation accuracy of 47%. They also fine-tuned the VGG-16 network and obtained an accuracy of 38%.

The paper by Happy et al. [8] proposes a new technique for expression recognition by using selected facial patches. Eye and nose localization is performed along with lip corner detection. They use Sobel filtering followed by Otsu thresholding to achieve this. They use the local binary patterns (LBP) for feature extraction and classification. They have achieved results varying from 87.8% for Angry to 98.46 for Surprise using the CK+ dataset.

While there has been some work carried out in detecting the standard affect states, there is not much research reported on detection and classification of learning-centered affective states.

Paper [9] deals with classifying frustrated and delighted smiles. The authors have successfully extracted the local and global features which are related to the dynamics of human smile. In the paper, the authors have analyzed and compared two variants of support vector machines (SVM), hidden Markov models (HMM) and hidden-state conditional random fields (HCRF) for binary classification. They have achieved an accuracy of 92 percent in distinguishing smiles under frustrating and delighted stimuli.

In [10], authors introduce a new configuration which the call SPLITFACE. This is a deep convolutional neural network-based method which successfully performs attribute detection in faces which are partially occluded. In this paper, several segments of the face are taken and they have identified which attributes are localized in which part of the face. The authors have successfully demonstrated that their method significantly outperforms other recent methods.

Large convolutional network models have exhibited very good classification performances on the ImageNet benchmark. In spite of their impressive results, there is still a lack of understanding of why they perform so well. Paper [11] discusses a novel visualization technique which gives an insight into the function of intermediate feature layers and the operation of the classifier.

VGG-16 is a deep CNN network which was published in 2015 [12]. In this work, the Visual Geometry Group (VGG) explores the effect of increasing the depth of the convolutional network on its accuracy. They use architecture with very small convolution filters of size 3 × 3, which show

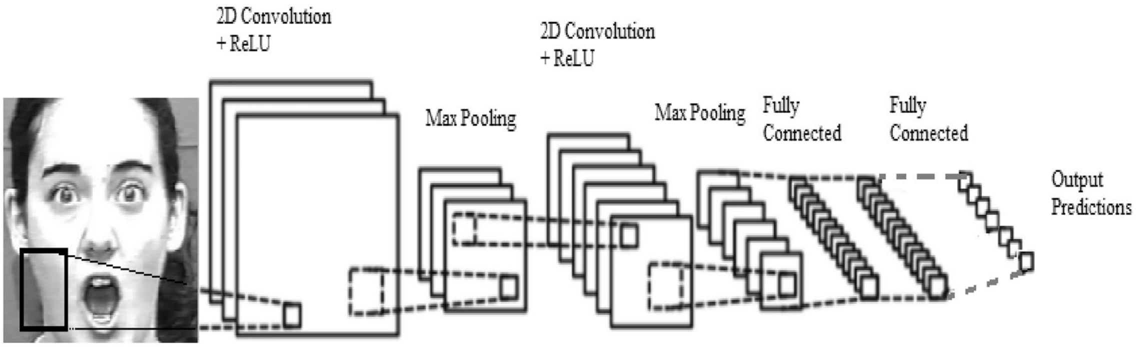

Fig. 1 Schematic of CNN architecture

significant improvement compared to the state-of-the-art configurations. These findings were submitted in the ImageNet Challenge 2014.

Squeeze-and-excitation networks are a recent addition to the growing advancements in the area of deep learning. Paper [13] introduces a new architectural unit, which they call the squeeze-and-excitation (SE) block. The goal of the SE block is to improve the quality of representations produced by a network. This they do by modeling the interdependencies between the channels of its convolutional features They have shown that it is not only possible to construct an SE network by simply stacking a collection of SE blocks but can also be integrated into standard architectures such as VGG, ResNet and Inception.

Dataset

We use the extended Cohn–Kanade dataset (CK +) [14]. Facial behavior of adults ranging from 18 to 50 years is recorded in the dataset and consists of 593 sequences across 123 subjects. Each sequence consists of images from the onset to the peak expression. The start is the neutral expression, and the last frame is the target expression. The target expression for each sequence is fully FACS coded.

The database consists of 7 standard expressions, viz. Disgust, Happy, Surprised, Fear, Angry Contempt and Sad. The peak expressions are shown in Fig. 2.

In our work, we have used the target frame along with a few extra frames prior to it. This gave us an increased number of images to carry out our experiments. A total of 2502 images are used. The distribution of labeled images used is shown in Table 1.

Proposed Methodology

The paper attempts to compare three different networks in relation to detection of standard affect states. It tries to understand which of the networks is most effective in detecting affect states.

The images in the CK+ dataset are against a constant background. The first step is to locate the face and eliminate the background. This is done using the Viola–Jones algorithm [15]. The extracted face image was resized to 256×256. This is shown in Fig. 3.

Of the total 2502 images selected from the dataset, 70% were used for training, while 30% were used for validation. Hence, 1751 images were used to train the network and 751

Table 1 Distribution of labeled images

Affect states	Number of images
Angry	412
Contempt	105
Disgust	325
Fear	220
Happy	604
Sad	313
Surprise	523
Total	2502

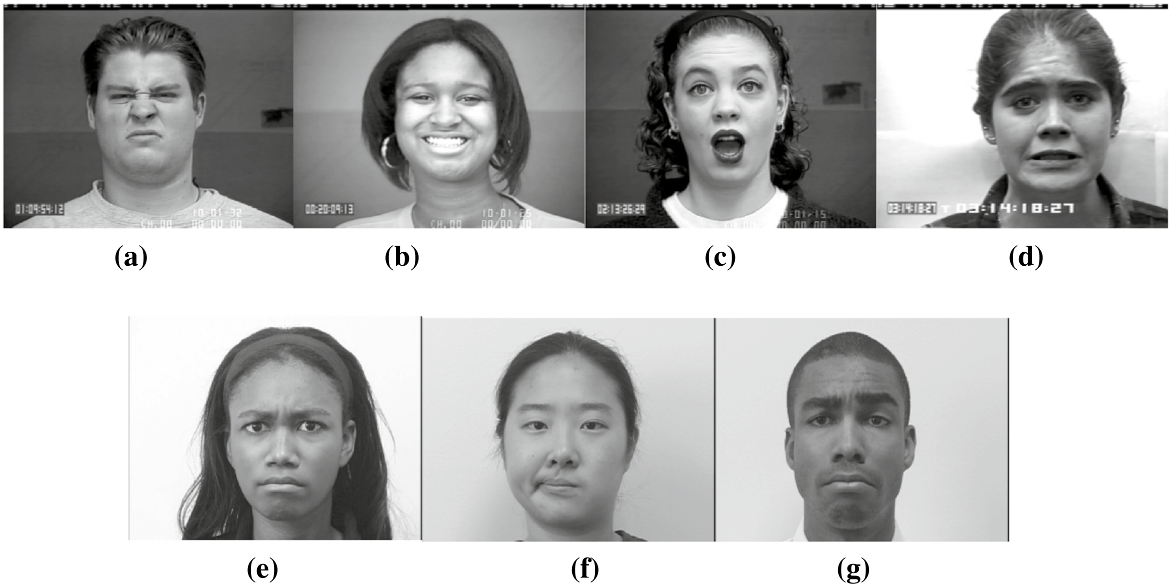

Fig. 2 **a** Disgust, **b** Happy, **c** Surprised, **d** Fear, **e** Angry, **f** Contempt and **g** Sad

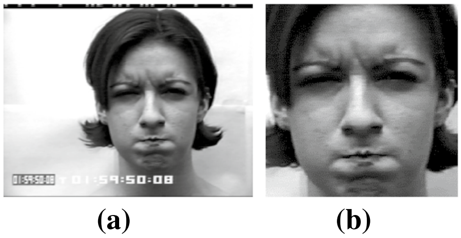

Fig. 3 **a** Original image from the database, **b** cropped image using Viola–Jones algorithm

Table 2 Distribution of labeled images in the validation set

Affect states	Number of images
Angry	114
Contempt	29
Disgust	95
Fear	60
Happy	189
Sad	99
Surprise	165
Total	751

images were used for validation. The distribution of affect states in the validation set is shown in Table 2. All images of the training set and the validation set were normalized.

Keras deep learning framework [16] is used, and it includes various pre-trained deep learning models along with their weights. The networks used in this work are VGG16, ResNet50 and SE-ResNet50.

VGG-16 Network

The VGG-16 network was trained on the ImageNet database [17, 18]. Because of the extensive training that the VGG-16 network has undergone, it gives excellent accuracies even when the image data sets are small.

The VGG-16 network consists of 16 convolution layers and has a small receptive field of 3 × 3. It has a Max pooling layer of size 2 × 2 and has a total of 5 such layers. There are 3 fully connected layers after the last Max pooling layer.

This is followed by three fully connected layers. It uses the softmax classifier as the final layer. ReLu activation is applied to all hidden layers. A schematic of the VGG-16 architecture is shown in Fig. 4 [19].

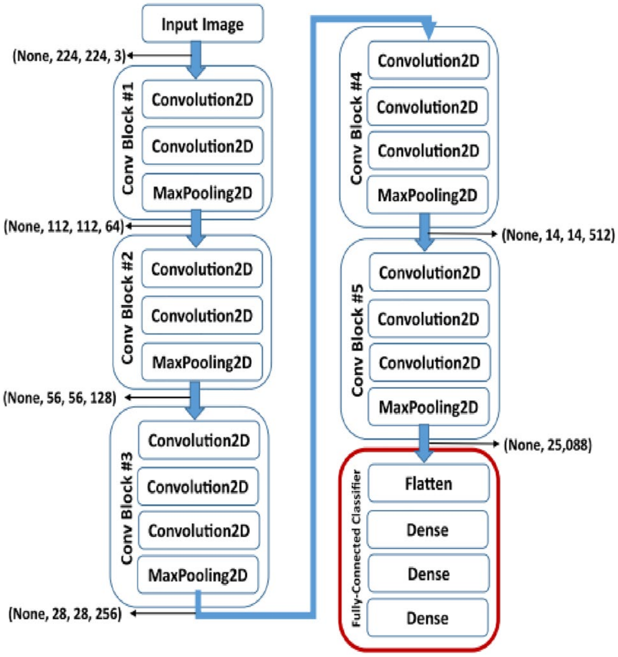

Fig. 4 Block diagram of VGG-16 network

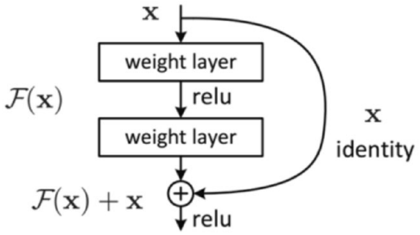
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Fig. 5 ResNet50

ResNet50 and SE-Resnet50

ResNet50 is a short form of residual networks having 50 layers. It is comparable to VGG-16 accept that Resnet50 has an additional identity mapping capability. This is shown in Fig. 5.

ResNet predicts the delta that is required to reach the final prediction from one layer to the next [20]. ResNet reduces the vanishing gradient problem by allowing this alternate shortcut path for gradient to flow through. The identity mapping used in ResNet allows the model to bypass a CNN weight layer if the current layer is not necessary. This helps in avoiding the over fitting problem to the training set. ResNet50 has 50 layers.

Squeeze-and-excitation (SE) block is a new architectural unit with the goal of improving the quality of representations produced by a network [13]. The SE block is integrated into existing architecture of ResNet50 by inserting it after the nonlinearity following each convolution. Here, the SE block transformation is taken to be the non-identity branch of a residual module. Squeeze and excitation both act before summation with the identity branch. The SE block improves channel interdependencies at almost no computational cost. The inclusion of the SE block into the ResNet50 network is shown in Fig. 6.

Results

All our experiments were developed in Keras and trained using Intel Core i5-7200U (7th Gen) CPU on 64-bit Window 10 OS. In this section, we explore the results obtained using the three pre-trained networks, viz. VGG-16, Resnet50 and SE-Resnet50 as classifiers to categorize an image into the seven affect states, viz., Anger, Contempt, Disgust, Fear, Happy, Sad and Surprise. We have compared the results obtained from them. The classifier used in all the three networks is softmax. The networks are trained using the images from the training set and tested using images from the validation set. Validation set comprises of images that the network has not seen before. The data were shuffled in order to train the network better. The image batch size was set to 32 for VGG16 and ResNet50 and 16 for SE ResNet50. In each case, the network was used as a classifier and only the last layer was modified to suit the classes under study.

The training and validation accuracy at the end of 25 epochs is given Table 3.

Table 3 shows that all the three networks perform exceptionally well. The training accuracy is the highest for VGG16 network with an accuracy of 0.9994. Training accuracy indicates how successful the network is in correctly classifying the data it is being trained on. Hence, a training accuracy of 0.9994 means the network is successful in classifying 99.94% of the images from the training set.

The validation accuracy is highest for ResNet50 network with an accuracy of 0.9947. Validation accuracy is more important as it indicates how successful the network is in correctly classifying the data it has not seen before. Hence, a validation accuracy of 0.9947 means the network is successful in classifying 99.47% of the images from the validation set which, unlike the training set, the network has not seen before.

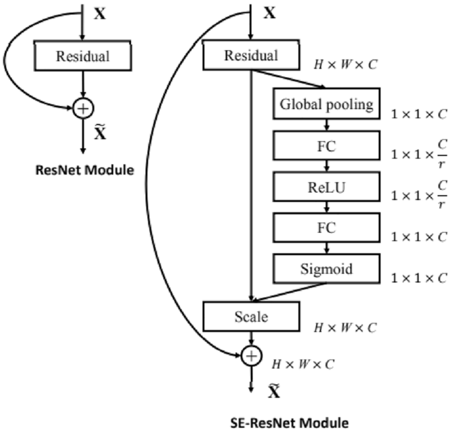

Fig. 6 SE-ResNet50

The validation accuracy curves of each of the networks are shown in Fig. 7. It is observed that ResNet50 achieves high validation accuracy with the least number of epochs. SE-ResNet50 takes the most number of epochs to achieve high validation accuracy.

While accuracy is one of the measures used to test a network, it tends to be misleading. Apart from accuracy, we have in this paper used confusion matrix, precision and recall as performance matrices to evaluate the results obtained. Precision indicates the false positives obtained, while recall gives us the false negatives. Figures 8, 9 and 10 give us the confusion matrix obtained using a VG16, ResNet50 and SE-ResNet50, respectively.

Discussion

Confusion matrices give us a pictorial representation of precision and recall. Along with validation accuracy, we have also successfully computed precision and recall values. The precision and recall values obtained using the three networks for the different affect states are given in Tables 4 and 5.

From Table 4, we observe that ResNet50 gives a high precision value of 1 for affect states of Angry, Contempt,

Table 3 Training and validation accuracy

	Training accuracy	Validation accuracy
VGG16	0.9994	0.968
ResNet50	0.9954	0.9947
SE ResNet50	0.9829	0.9734

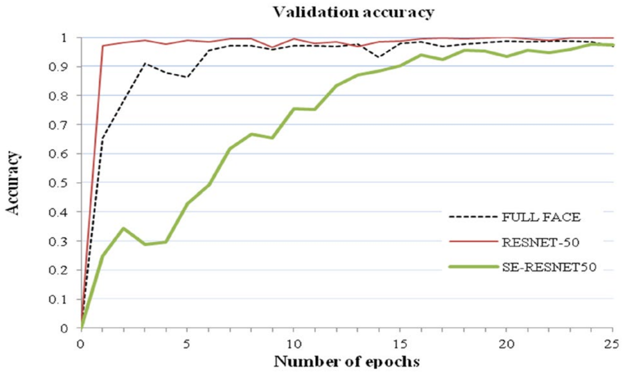

Fig. 7 Validation accuracy curves with 25 epochs

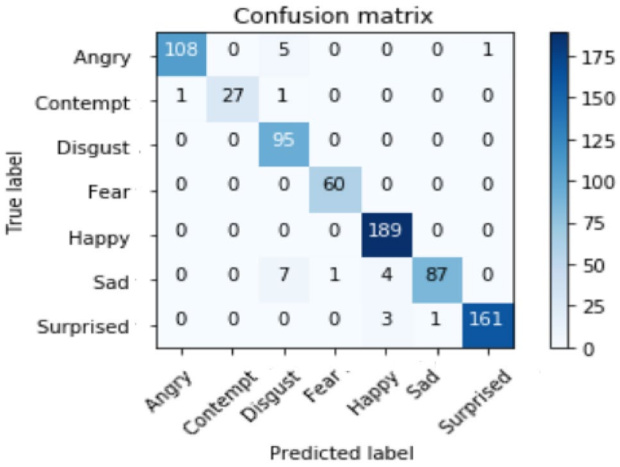

Fig. 8 Confusion matrix using VGG16

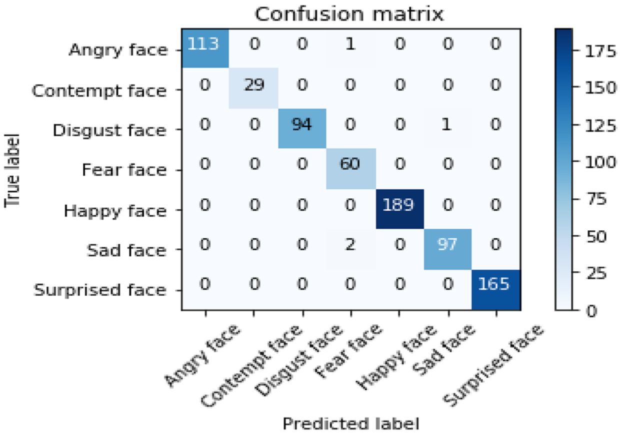

Fig. 9 Confusion matrix using ResNet50

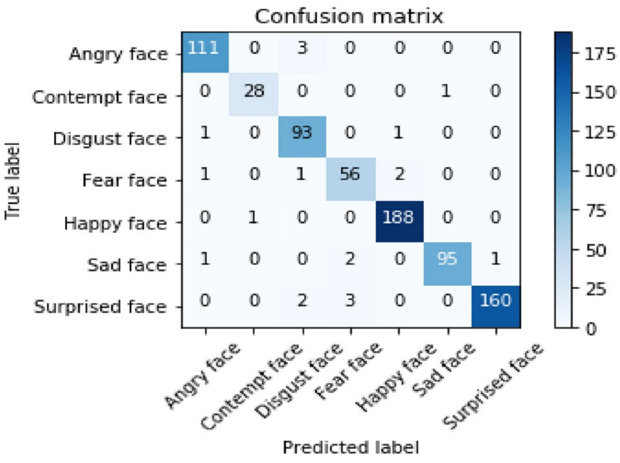

Fig. 10 Confusion matrix using SE-ResNet50

Table 4 Precision values obtained for the three networks

	VGG16	ResNet50	SE-ResNet50
Angry	0.99	1.00	0.97
Contempt	1.00	1.00	0.97
Disgust	0.88	1.00	0.94
Fear	0.98	0.95	0.92
Happy	0.96	1.00	0.98
Sad	0.99	0.99	0.99
Surprise	0.99	1.00	0.99

Table 5 Recall values obtained for the three networks

	VGG16	ResNet50	SE-ResNet50
Angry	0.95	0.99	0.97
Contempt	0.93	1.00	0.97
Disgust	1.00	0.99	0.98
Fear	1.00	1.00	0.93
Happy	1.00	1.00	0.99
Sad	0.88	0.98	0.97
Surprise	0.98	0.91	0.98

Disgust, Happy and Surprise. This implies that there are zero false positives obtained while detecting the above-mentioned affect states. VGG16 gives us a precision of 1 for Contempt. From Table 5, we observe that ResNet50 gives a high recall value of 1 for affect states of Contempt, Fear and Happy indicating zero false negatives for these states, while VGG16 gives a recall of 1 for affect states Disgust, Fear and Happy. ResNet50 achieves these results at low number of epochs compared to the other networks.

Based on all the performance matrices mentioned here, we have shown that all the 3 networks are capable of detecting various affect states with very high accuracy, precision and recall, with ResNet50 performing slightly better than the other two.

Conclusion

All experiments were performed on the CK+ facial database. Three pre-trained networks, viz. VVG-16, ResNet50 and SE-ResNet50, were used, and their top fully connected layer was modified. We have successfully compared VGG16, ResNet50 and SE-ResNet50 and demonstrated that using transfer learning, all the three pre-trained networks give very high accuracy, precision and recall values. Of these, ResNet50 achieves the highest precision and recall and also takes much lesser epochs to train.

Hence, CNN with transfer learning helps to identify the various affect states very accurately thereby improving the interaction between humans and computers. We could further probe if the same technique of using CNN with transfer learning can help solve the problem of occlusions.

References

1. Calvo RA, D'Mello S. Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans Affect Comput. 2010;1(1):18–37.
2. Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. In: 2012 IEEE conference on computer vision and pattern recognition 2012 Jun 16, IEEE, pp 3642–3649.
3. Dachapally PR. Facial emotion detection using convolutional neural networks and representational autoencoder units. arXiv preprint arXiv:1706.01509 (2017).
4. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems 2012, pp 1097–1105.
5. https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/. Accessed 4 May 2019.
6. Ekman P, Friesen W. Facial action coding system: a technique for the measurement of facial movement. Palo Alto: Consulting Psychologists Press; 1978.
7. Raghuvanshi A, Choksi V. Facial expression recognition with convolutional neural networks. CS231n Course Projects (2016).
8. Happy SL, Routray A. Automatic facial expression recognition using features of salient facial patches. IEEE Trans Affect Comput. 2015;6(1):1–12.
9. Hoque ME, McDuff DJ, Picard RW. Exploring temporal patterns in classifying frustrated and delighted smiles. IEEE Trans Affect Comput. 2012;3(3):323–34.
10. Mahbub U, Sarkar S, Chellappa R. Segment-based methods for facial attribute detection from partial faces. IEEE Transactions on Affective Computing. 2018 Mar 27.
11. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision 2014 Sep 6. Cham: Springer; 2014. pp 818–833.
12. Simonyan K, Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
13. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018.
14. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW). IEEE, pp. 94–101.
15. Viola P, Jones MJ. Robust real-time face detection. Int J Comput Vision. 2004;57(2):137–54.
16. Chollet F. Keras. GitHub; 2015.
17. Pal S. Transfer learning and fine tuning for cross domain image classification with Keras. GitHub: transfer learning and fine tuning for cross domain image classification with Keras; 2016.
18. Deng J, et al. Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009. IEEE; 2009.
19. Gopalakrishnan K, et al. Deep convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection. Constr Build Mater. 2017;157:322–30.
20. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778; 2016.