International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Affect State Classification from Face Segments Using Resnet-50 and SE-Resnet-50

Dhananjay Theckedath, R.R Sedamkar

Abstract: ***One of the important components of an intelligent Human computer Interface system is accurate classification of the various affect states. Such interface systems are however plagued by a recurring problem of image occlusion. The challenge hence is to be able to classify the various affect states accurately from whatever portions of the face are available to the system. This paper attempts to investigate if there are segments within the facial region which carry sufficient information about the affect states. In this paper we have used two pre-defined Convolutional Neural networks (CNN). We have implemented a ResNet-50 network and a modified version of ResNet-50 which has a Squeeze and Excitation network connected to ResNet-50. This is called SE-ResNet-50. We use these two networks to classify seven basic affect states of Angry, Contempt, Disgust, Fear, Happy, Sad and Surprise from various segments of the face. We partition the face into four regions with each region comprising of only 50% of the original data. The results obtained are compared with that obtained using the full face. The validation accuracy values are obtained for full face as well as the four segments of the face. The paper also calculates precision and recall for each partitioned area for each of the affect states using the two networks. Our evaluation shows that both, ResNet-50 as well as SE-ResNet-50 are successful in accurately classifying all the 7 affect state from the Right segment, Left segment Lower segment and Upper segment of the face. While ResNet-50 performs marginally better compared to the SE-ResNet-50 in identifying the various affect states form the right, left and lower segments of the face, SE-ResNet-50 performs better in identifying the affect states from the upper segment of the face. We can thus conclude that right segment, left segment, lower segment and upper segments of the face contain sufficient information to correctly classify the seven affect states. The experimental results presented in this paper show that pre-defined Convolutional Neural Networks gives us very high accuracy, precision and recall values and hence can be used to accurately classify affect states even when there are occlusions present in the image and only certain portions of the face are available for analysis.***

Keywords : Affect States, Convolutional neural networks, ResNet-50, SE- ResNet-50;

I. INTRODUCTION

Facial expression analysis is a major challenge in the field of affect computing.

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Dhananjay Theckedath*, Assistant Professor, Biomedical Engineering department, Thadomal Shahani Engineering college, India.
R.R. Sedamkar, Professor, Computer Engineering department, Thakur college of Engineering & Technology, India.

Affect describes the experience of feeling or emotion. We deduce affect states of individual using facial expressions.
The various affect states are achieved through a complex combination of information produced by the brain and the facial muscles.
Affect analysis depends upon accurate detection and classification of facial features. The complexity of the task is increased by variations in head rotation, illumination, facial poses and occlusion of the face. Image occlusion is a recurring issue while acquiring images for an intelligent human computer interface.
Detection of the affect states becomes a challenging problem in the absence of a full face. Occlusions can arise due to posture of sitting, palms on the face, reading glasses etc. Designing models for automatic expression recognition systems, which are tolerant to partial occlusion, remain a challenging task.
A lot of ongoing research is being carried out in the area of Deep Learning. One of the applications where deep learning is now being used is in developing systems which are capable of recognizing and responding to the affect states [1].
Deep learning models extract relevant features from large datasets automatically and have shown to achieve very high accuracies [2]. Deep learning models have turned out to be very good in discovering intricate structures and are therefore applicable to many domains. With the advances in Graphics Processing Units (GPU), emotion recognition has become a widely-tackled research problem [3].
Convolutional Neural Network (CNN) is one of the commonly used deep learning networks. Traditional neural networks only contain 2-3 hidden layers, while deep networks can have as many as 150. The interest in CNN started with AlexNet which won the ImageNet competition in 2012 [4] .
Arushi and Vivek [5] developed a 5 layer CNN to classify human expressions 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 S L Happy et al [6] performs eye and noce localization alon with lip corner detection. All these three are know to carry expression information. They use sobel filtering and follow it up by otsu thresholding. Local binary patterns (LBP) of these imaes are computed for feature extraction. They have achieved results varying from 87.8% for angry to 98.46 for Surprise using the CK+ dataset.

In [7], authors introduce a new configuration which they call SPLITFACE which is used for partially occluded images. This is a deep convolutional neural network-based method which successfully performs attribute detection in faces which are partially occluded. The paper has successfully shown that the methods mentioned perform better than the recent methods.

Paper [8] discusses a novel visualization technique which gives an insight into the function of intermediate feature layers of large convolutional network models as well as the operation of the classifier.

ResNet50 which is a short form of Residual Networks is deelp learning architecture [9]. The ResNet is similar to other deep networks but has an additional identity mapping capability . Res-Net models fit a residual mapping to predict the delta needed to reach the final prediction from one layer to the next. This has shown to address the vanishing gradient problem.
Squeeze and Excitation Networks are a recent addition to the growing advancements in the area of deep learning. [10] 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 .
In paper [11], a method is presented to recognize facial expressions from time sequential facial expression. A technique known as Enhanced Independent Component Analysis (EICA) is implemented to extract the locally independent component features. Using the features obtained, discrete Hidden Markov models (HMMs) are used to model different facial expressions.
Due to high variation in subjects, Facial expression recognition is a challenging task. To thwart problems arising due to variations introduced by personal attributes and achieve better facial expression recognition performance, paper [12] proposes a novel identity-aware convolutional neural network (IACNN). Extensive experiments on three public datasets including a spontaneous facial expression datasets have shown that the proposed IACNN achieves promising results in real world.

Paper [13] proposes a deep convolutional network by using various facial parts. In this work, a combination of face detection, feature extraction and classification algorithms are discussed. The results achieved shows that the system provides improved classification accuracy when compared to other methods.
Authors of this paper [14] propose a novel deep learning framework, called spatial-temporal recurrent neural network (STRNN). This framework integrates learning from both, spatial as well as temporal information into a unified spatial-temporal dependency model. In this network, a multidirectional Recurrent Neural Network (RNN) layer is used to obtain contextual cues by moving across the spatial regions of each temporal slice along different directions
Transfer learning is a technique used in CNN in which a model trained on one set of data is used to identify features in a second set of data. In order to achieve very high accuracies, CNN's need to be trained using thousands of samples. This as we know, is not always possible in most of the cases. In transfer learning a pre-trained network which has been already trained using a very large data set is used. This pre trained network is then trained using data from the new dataset. In other words, we transfer the weights that a Network has learned from task comprising of a huge dataset to a new task. 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. The pre trained networks used in this paper are ResNet-50 and SE-ResNet-50.

This paper is organized as follows. Section II discusses the experimental set up used. Section III explains the proposed methodology. Results are shown in section IV and are discussed in Section V. Conclusion is given in section VI.

II. EXPERIMENTAL SETUP

A. Dataset

We use the Extended Cohn-Kanade dataset (CK +) [15] . The dataset consists of 593 sequences across 123 subjects. Each of these sequence consists of a series of images starting from the neutral expression to the peak expression also known as the target expression. All the target expressions are fully FACS coded and Correlation between Action units and affect states is provided in the paper. The seven standard expression in the dataset are Disgust, Happy, Surprised, Fear, Angry Contempt and Sad.

B. Data Preparation

As seen in Fig.1 (a), the dataset consists of frontal face images along with the background which contains certain text information. The first step was to extract the face from the background. We use the Viola & Jones algorithm to achieve this. [16]. The cropped images were then resized to size 256 × 256. An original image from the dataset and the cropped image using Viola& Jones algorithm is shown in Fig. 1. The image shown is of the final target expression from the dataset.

(a) **(b)**

Fig. 1. a) Original image from the dataset ,b) Cropped image using Viola Jones algorithm

While creating the dataset for our work, we include not just the target expression but also a few more frames prior to it. This is done in order to have an increased number of images for our experiments and also to make our network robust and invariant to minor variations in the expression. Hence the total number of images used is 2502.

Fig. 2. Various stages leading to the target expression.

Since each of these images is labeled, the distribution of labeled images used is shown in Fig. 3.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

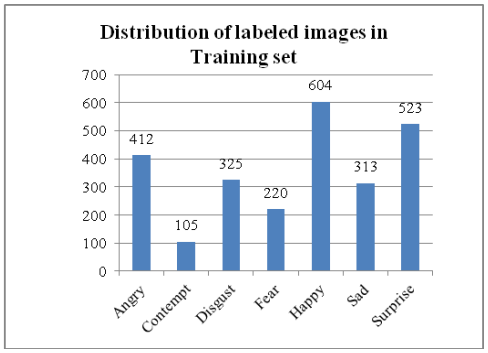


Fig. 3. Distribution of labeled images in the training set

The next step was to partition each of 256 × 256 images o into 4 segments viz. Right segment, Left segment, Upper segment and Lower segment. This was done by selecting half part of the image and replacing the pixels belonging to the remaining part of the image with zero. The result of this partioning on a single image is shown in Fig. 4.

(a) (b) (c) (d) (e)

Fig. 4. (a) Full Face, (b) Right segment, (c) Left segment, (d) Upper segment, (e) Lower segment

As is evident, each image now contains only 50% of the original information. This partioning is done on every image of the dataset. At the end of this procedure we had successfully partitioned each of the 2502 images into 4 separate segments. We created 5 main folders which were named as Full Face folder, Right segment folder, Left segment folder, Upper segment folder and Lower segment folder and each of these folders comprised of 2502 images. These 5 folders had 7 sub- folders comprising of the seven standard affect states viz. Disgust, Happy, Surprised, Fear, Angry Contempt and Sad.

Out of a total of 2502 images in each main folder , we used 70% of the images for training and the remaining 30% of the images for validation. Hence the network was trained using 1751 images and validated using 751 images. This distribution and of 70% and 30% is done randomly and all images of the training set and the validation set were normalized. The distribution of images in the validation set is shown in Fig. 5.

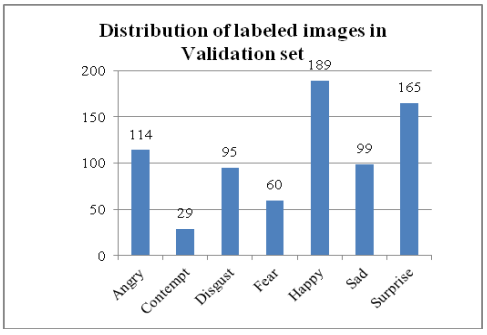


Fig. 5. Distribution of images in the validation set

III. PROPOSED METHODOLOGY

The objective of selecting only a segment of the face was to mimic the effect of occlusions where only certain sections of the full face are available. In this paper we show how each segment of the face contributes to the detection of expression. The paper tries to identify which of the four segments carry the most relevant information required for affect detection. In this paper we use the Convolutional Neural network along with transfer learning to perform our experiments. We use two trained networks in this paper viz. ResNet-50 and a modification of ResNet-50 known as SE ResNet-50. We will briefly discuss them.

A. ResNet-50

ResNet50 which is a short form of Residual Networks is similar in architecture to networks such as VGG-16 but has an additional identity mapping capability. Instead of fitting the latent weights to predict the final emotion at each layer, Res-Net models fit a residual mapping to predict the delta needed to reach the final prediction from one layer to the next. ResNet diminish the vanishing gradient problem by allowing this alternate shortcut path for gradient to flow through. The identity mapping permits the model to bypass a CNN weight layer if the current layer is not necessary. This further helps the model to avoid over fitting to the training set. ResNet50, there are 50 layers.

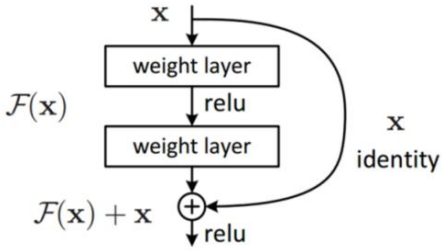


Fig. 6. (a) Full Face, (b) Right segment, (c) Left segment, (d) Upper segment, (e) Lower segment

B. SE-ResNet-50

Squeeze and excitation (SE) block is a new architectural unit which is aimed at improving the quality of representations produced by a network .The SE block can be integrated to any existing network. In this work, we integrated the SE block with ResNet 50 by inserting it after the non-linearity following each convolution. Squeeze and Excitation both act before summation with the identity branch. This is shown in Fig. 7

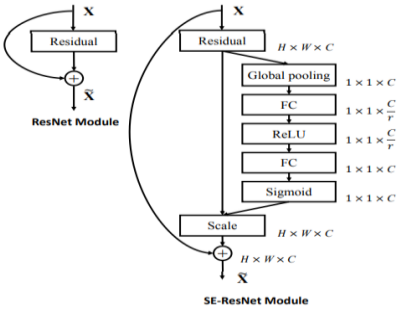


Fig. 7. SE Resnet50

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

IV. RESULTS

We have used the Nvidia Titan Xp GPU and all our experiments were developed in Keras [17].

We use categorical cross-entropy loss function, Softmax classifier and rmsprop as the optimizer.

Both the networks are trained on each segment of the face including the full face. The networks are trained using the images from the training set and tested using images from the validation set. The data in the training set is shuffled so as to randomize the order of the images. This procedure helps in training the network in a better way. The image batch size was set at 32 and we ran 25 epochs.

The training and validation accuracy at the end of 25 epochs for each of the expressions using ResNet-50 is given in Table-I.

Table- I: Training and Validation accuracy using ResNet-50

ResNet50	Training accuracy	Validation accuracy
Full Face	1.00	0.99
Right segment	0.99	0.99
Left segment	1.00	1.00
Lower segment	1.00	0.99
Upper segment	0.99	0.91
Average	0.99	0.98

The training and validation accuracy at the end of 25 epochs for each of the expressions using SE-ResNet-50 is given in Table-II.

Table- II: Training and Validation accuracy using SE-ResNet-50

SE- ResNet50	Training accuracy	Validation accuracy
Full Face	0.9	0.98
Right segment	0.98	0.98
Left segment	0.98	0.94
Lower segment	0.98	0.97
Upper segment	0.96	0.97
Average	0.96	0.97

Validation accuracy is more important as it is the accuracy of the network on images it has not encountered before. A comparison of the performance of ResNet-50 and SE-ResNet-50 on the various facial segments is shown in Fig. 8.

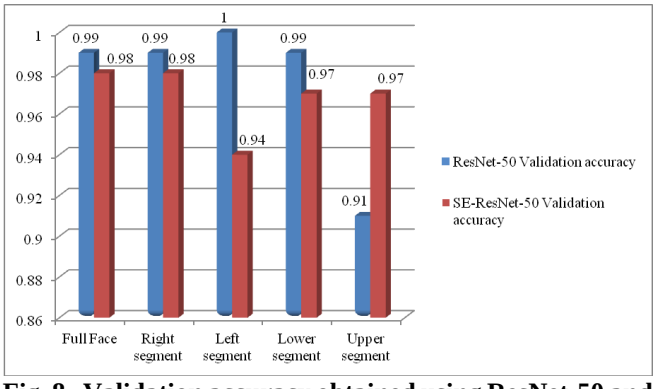
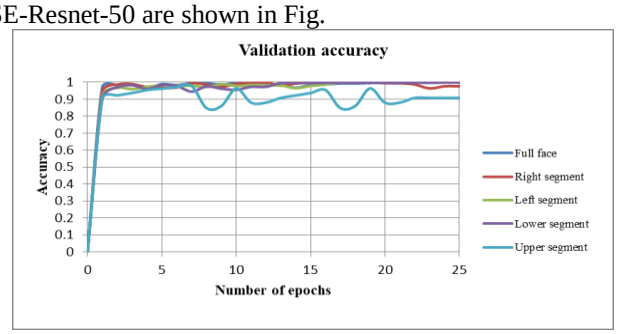


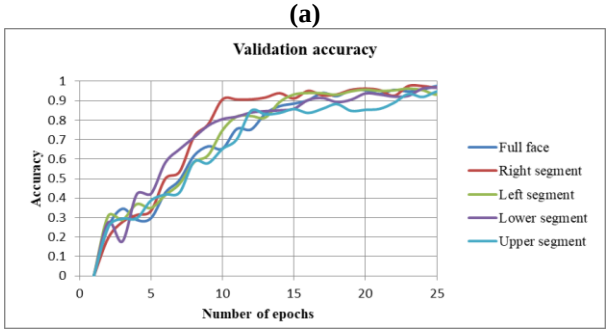
Fig. 8. Validation accuracy obtained using ResNet-50 and SE-ResNet-50 at the end of 25 epochs.

The accuracy values clearly indicate that affect state can accurately be identified from image segments.

The validation accuracy curves for ResNet-50 and SE-Resnet-50 are shown in Fig.



(a)



(b)

Fig. 9. (a) Validation accuracy curves using ResNet-50 on various segments of the face (b) Validation accuracy curves using SE-ResNet-50 on various segments of the face.

As seen in Figure 9, the validation accuracy of both the networks is very high. Resnet-50 achieves this accuracy in fewer number of epochsWhile the validation accuracy for both the networks is high, it is important to note that only Observing the accuracy can be misleading in certain situations as it does not give us information as to which affect states were classified correctly and which were not. Along with accuracy, we have also used Precision and Recall values as our performance matrices to enable us to quantify the results obtained. Precision gives us the false positives that are calculated by the network. Table-III gives us the Precision values obtained using ResNet-50.

Table- III: Precision values obtained using ResNet-50.

ResNet-50	Full Face	Right Segment	Left Segment	Lower Segment	Upper Segment
Angry	0.98	0.99	1.00	0.99	0.84
Contempt	1.00	1.00	1.00	1.00	1.00
Disgust	1.00	0.98	1.00	0.99	0.85
Fear	1.00	1.00	1.00	1.00	0.78
Happy	1.00	1.00	1.00	1.00	0.98
Sad	1.00	1.00	1.00	1.00	1.00
Surprise	1.00	0.98	1.00	0.99	1.00
Average	0.99	0.99	1.00	0.99	0.93

Table-IV gives us the Precision values obtained using SE-ResNet-50

Table- IV: Precision values obtained using SE-Reset-50

SE-ResNet-50	Full Face	Right Segment	Left Segment	Lower Segment	Upper Segment
Angry	0.97	0.96	0.95	0.96	0.93
Contempt	0.97	0.97	0.78	1.00	0.93
Disgust	0.94	0.92	1.00	0.97	0.98
Fear	0.92	1.00	0.98	0.89	0.97
Happy	0.98	1.00	0.94	0.98	0.97
Sad	0.99	0.98	0.89	0.96	1.00
Surprise	0.99	1.00	0.99	0.99	0.98
Average	0.97	0.98	0.94	0.97	0.97

While Precision gives us the false positives, Recall gives us the false negatives. Recall values obtained using ResNet-50 are given in Table-V.

Table-V: Recall values obtained using ResNet-50

ResNet-50	Full Face	Right Segment	Left Segment	Lower Segment	Upper Segment
Angry	1.00	1.00	1.00	0.99	0.99
Contempt	1.00	0.97	1.00	1.00	0.90
Disgust	1.00	0.99	1.00	0.99	0.98
Fear	1.00	0.97	1.00	1.00	0.98
Happy	1.00	0.99	1.00	1.00	0.89
Sad	0.99	1.00	1.00	1.00	0.74
Surprise	0.99	1.00	1.00	0.99	0.97
Average	0.99	0.99	1.00	0.99	0.93

Recall values obtained using SE-ResNet-50 are given in Table-VI.

Table-VI: Recall values obtained using SE-ResNet-50

SE-ResNet-50	Full Face	Right Segment	Left Segment	Lower Segment	Upper Segment
Angry	0.97	0.96	0.92	0.94	0.98
Contempt	0.97	0.97	1.00	0.90	0.97
Disgust	0.98	0.99	0.81	0.94	0.89
Fear	0.93	0.98	0.90	0.98	0.97
Happy	0.99	0.99	0.99	0.99	0.99
Sad	0.96	0.96	0.98	0.98	0.96
Surprise	0.97	0.99	0.98	0.98	0.99
Average	0.98	0.98	0.94	0.96	0.97

V. DISCUSSION

From our experiments we observe that ResNet-50 and SE-ResNet50 give us validation accuracies of 99% and 98 %for full face respectively. From Fig. 8 we note that ResNet-50 gives us very high validation accuracies for Right segment, Left segment and Lower segments of the face between 99% and 100%. The SE-Resnet-50 also gives us high validation accuracies for the NRight, Left and lower segments of the face with values ranging between 94% and 98%. Hence Resnet-50 gives us marginally better results compared to the SE-resnet50 for these three segments of the face.The ResNet-50 however gives us a comparatively low validation accuracy of 91% for the Upper segment of the face while SE-ResNet-50 gives us a higher validation accuracy of 97% for the Upper segment of the face.This implies that that Resnet50 is not as accurate at detecting the affect states form the uppwer segment of the face as compared to SE-ResNet50. From Fig.9 it is evident that Resent-50 achieves this high validation accuracy in fewer epochs compared to the SE ResNet-50 and hence is faster of the two networks. Along with accuracy, we have also computed precision and recall values. Precision gives us the false negatives while recall gives us the false positives computed by the network.Tables III and IV give us the precision values that are computed using ResNet50 and SE-Resnet50 networks. From the last column of Table III we observe that ResNet50 gives us false positives for the affect states of angry (0.84), Disgust (0.85) and Fear(0.78) from the Upper segment of the face. Tables V and VI give us the recall values computed by ResNet50 and SE-ResNet-50 networks. From the last column of the Table V we observe that ResNet-50 gives us false negatives for the affect state of Sad(0.74) . Comparing the results obtained in table III and V with table IV and VI, we state that that baring the upper segment, ResNet-50 performs marginally better that SE-ResNet-50 in identifying all the affect states. For the Upper segment of the face however, SE-ResNet-50 is more consistent in identifying all affect states.

VI. CONCLUSION

In this paper we have tried to investigate if accurate affect detection can be performed only from certain sections of the face. We have used deep learning networks to perform our experiments. The two networks used here are ResNet-50 and SE-ResNet-50. Based on all the performance matrices mentioned here, we can infer that Right segment, Left segment, Lower segment and Upper segments of the face contain sufficient visual information required to classify the various affect states. The experimental results presented in this paper show that CNN using pre-defined networks gives us very high accuracies and hence can be used to classify affect states even when there are substantial occlusions present in facial images. We can also infer from our results that even when the entire face is available we only need to work with 50% of the image to obtain accurate affect classification .

ACKNOWLEDGMENT

This work is supported in part by NVIDIA GPU grant program. We thank NVIDIA for giving us Titan XP GPU as a grant to carry out our work in deep learning.

REFERENCES

1. Calvo, Rafael A., and Sidney D'Mello. "Affect detection: An interdisciplinary review of models, methods, and their applications." IEEE Transactions on affective computing 1.1 (2010): 18-37.
2. Cireşan, Dan, Ueli Meier, and Jürgen Schmidhuber. "Multi-column deep neural networks for image classification." arXiv preprint arXiv:1202.2745 (2012).
3. Dachapally, Prudhvi Raj. "Facial Emotion Detection Using Convolutional Neural Networks and Representational Autoencoder Units." arXiv preprint arXiv:1706.01509 (2017).

4. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.
5. Raghuvanshi, Arushi, and Vivek Choksi. "Facial Expression Recognition with Convolutional Neural Networks." CS231 Course Projects (2016).
6. Happy, S. L., and Aurobinda Routray. "Automatic facial expression recognition using features of salient facial patches." IEEE transactions on Affective Computing 6.1 (2015): 1-12.
7. Hoque, Mohammed Ehsan, Daniel J. McDuff, and Rosalind W. Picard. "Exploring temporal patterns in classifying frustrated and delighted smiles." IEEE Transactions on Affective Computing 3.3 (2012): 323-334.
8. Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.
9. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.
10. Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141. 2018.
11. M. Z. Uddin, J. J. Lee, and T.-S. Kim, "An enhanced independent component-based human facial expression recognition from video," IEEE Trans. Consum. Electron., vol. 55, no. 4, pp. 2216– 2224, Nov. 2009.
12. Meng, Zibo, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. "Identity-aware convolutional neural network for facial expression recognition." In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 558-565. IEEE, 2017.
13. Nwosu, Lucy, Hui Wang, Jiang Lu, Ishaq Unwala, Xiaokun Yang, and Ting Zhang. "Deep convolutional neural network for facial expression recognition using facial parts." In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), pp. 1318-1321. IEEE, 2017.
14. Zhang, Tong, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. "Spatial–temporal recurrent neural network for emotion recognition." IEEE transactions on cybernetics 49, no. 3 (2018): 839-847.
15. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010, June). The extended cohn kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (pp. 94-101). IEEE.
16. Viola, Paul, and Michael J. Jones. "Robust real-time face detection." International journal of computer vision 57(2) 137-154, 2004.
17. F. Chollet, keras. GitHub, 2015

AUTHORS PROFILE

Dhananjay Theckedath is an Assistant Professor, at Thadomal Shahani Engineering college, India. He is pursuing his PhD from Mumbai University. His research interests are Image processing, Digital Signal processing and Control systems.

R.R. Sedamkar is a Professor and Dean at Thakur college of Engineering, India. He has more than 25 years of teaching experience. His areas of interest are Networking and data Compression. He is a PhD. guide