© 2018 IJSRSET | Volume 4 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
National Conference on Advanced Research Trends in Information and Computing Technologies (NCARTICT-2018), Department of IT, L. D. College of Engineering, Ahmedabad, Gujarat, India In association with
International Journal of Scientific Research in Science, Engineering and Technology

Survey on Different Methods to Improve Accuracy of The Facial Expression Recognition Using Artificial Neural Networks

Chirag Ravat[1] , Shital A. Solanki[2]

[1]M.E., I.T. Department, L.D College Of Engineering, Ahmedabad, Gujarat, India
[2]Assist. Prof., I.T. Department, L.D College Of Engineering, Ahmedabad, Gujarat, India

ABSTRACT

Facial expression recognition by computer plays a key role in human computer interaction. FER has many applications such as Human-Robot interaction, surveillance, Driving-safety, Health-care, Intelligent tutorial system, music for mood, etc. Basically, Facial expression recognition can be done using Artificial Neural Network (ANN) and Support Vector Machine (SVM). So the accuracy of facial expression depends on these two phases, Feature extraction phase and classification phase. In this paper I'm going to survey different methods of FER and even face recognition methods.

Keywords: Facial Expression, Face recognition, CNN

I. INTRODUCTION

Artificial Neural Networks are computing systems inspired by the biological neural networks that constitute animal brains. Such systems learn to do tasks by considering examples, generally without task-specific programming. The human brain is composed of 86 billion nerve cells called neurons[1]**.** They are connected to other thousand cells by Axons**.** ANNs are composed of multiple nodes, which imitate biological neurons of human brain. The neurons are connected by links and they interact with each other.

The nodes can take input data and perform simple operations on the data. The result of these operations is passed to other neurons. The output at each node is called its activation or node value. Each link is associated with weight. ANNs are capable of learning, which takes place by altering weight values[1].
However Convolutional Neural Networks(CNNs) have an edge over conventional MLPs in terms of image recognition and classification. The brief introduction of CNN is given in chapter III.

Facial Expression Recognition System

Almost all prediction based systems consist of mainly two phases: Training phase and Testing phase. In either Supervised or Un-supervised learning method first feature vectors are trained based on given labelled data in case of Supervised learning or on given attributes in case of Un-supervised learning. These trained vectors are then used to test an unseen data and gives label accordingly.

In FER it is identified as Feature Extraction and Classification phases. In feature extraction phase certain features of an image is extracted based on which classifier, in classification phase, classifies image to one of the label of domain.

For fast and better training, first of all image is pre-processed to reduce noise. Pre-processing also includes converting image into grayscale and resizing it into predefined dimensions. For different FER methods different sizes can be considered.

Then next step is to detect face from the pre-processed image. There are lots of techniques available for detecting face from an image like Haar classifier, Ada boost technique by Viola-jones, Adaptive skin colour, etc., which gives output as face image or non-face image[4].

After face is detected another step is to extract features from face image. It can be done directly from given image or also from video frames. Features such as eyes, nose, mouth, eyebrows, ears, etc. are detected. Feature extraction process works significantly well if face is already detected. There are two types of methods for feature extraction : Appearance based and Geometric based. Geometric based methods are more suitable for real time applications as Appearance based method consumes more power, time and memory but also highly discriminative. There are many feature extraction methods are there but Gabor feature and LBP[2] gives optimal results. There are also pre-defined Neural Net structures available such as AlexNet[6], which is a Convolutional Neural Net, useful in feature extraction.

In last classifier is used to train model using extracted features and classify them into label. Classifier learns the mapping between the given image and given label, so when we give a new image after sufficient training it can classify input image into certain label.

Also testing phase includes image pre-processing and feature extraction, based on these features classifier directly classifies that image into some label.

II. LITERATURE REVIEW

A. Facial Expression Recognition Using General Regression Neural Network[2]

In this paper authors proposed two phases for FER, namely Training phase and Testing phase. Both phases includes Image Pre-processing and Feature extraction. In training phase these extracted features are trained and mapping is learned between image and label for that image. And in testing phase extracted image can be directly classified into some label.

Here, in image pre-processing section they intent to reduce noise in image, converting image into grayscale and resizing image into various block sizes. They trained the model using various sized blocks : 256×256, 128×128, 64×64 and 32×32. Their experiment showed that 64×64 block size gives optimal result in terms in accuracy, Where 256×256 gives lowest one.

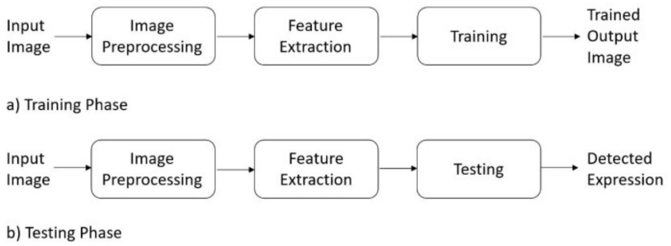


Figure 1. FER system[2]

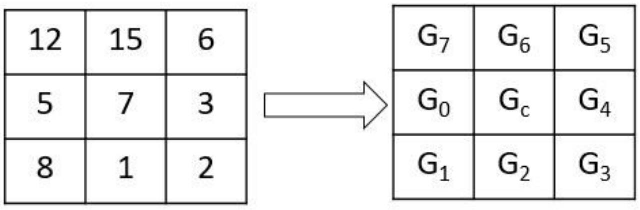
Then face detection module has been applied to pre-processed image. It targets the face and crops it. The cropped face is normalized using histogram equalization. Here Ada boost technique is used in Viola-Jones face detection algorithm.

Now comes the most important part of any image based classification method, Feature extraction. In this paper they used Local Binary Pattern(LBP) to extract feature from input image. LBP is used to determine feature vector which reduces the entire image data to a single feature vector ready for classification. Main purpose of feature extraction technique is to reduce large amount of pixel data into small most meaningful feature vector to ease classification problem.

The last step is classification. Extracted features are classified into six already known expression classes : Happy, Unhappy, Surprise, Disgust, Fear and Angry. For classification of these feature Unsupervised approach General Regression Neural Net(GRNN) is used. It trains network faster and does not require iterative training procedure.

Local Binary Pattern

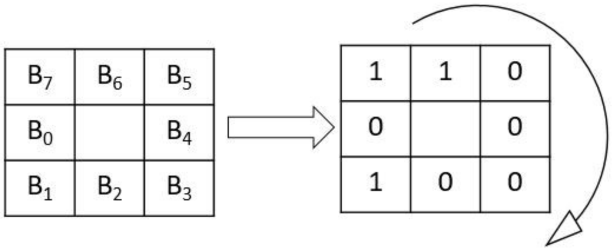
LBP was first introduced in 1996 by Ojala et al as a basic binary operator. It works as powerful texture classifier. LBP is simple tool for detection of feature. It is widely used because of its robustness and simplicity.



12	15	6
5	7	3
8	1	2

G_7	G_6	G_5
G_0	G_c	G_4
G_1	G_2	G_3

(a)



(b)

Figure 2. LBP Coding[2]

The pixels of the image are labelled by the binary operator by comparing the center pixel value with the 3×3 neighbourhood pixels. If value of pixel is greater than central pixel then it is assigned 1, if lower then assigned 0. Then these pixels traced circularly and formed LBP code.

This 8 bit code (Byte) is converted into decimal number. Advantage of LBP is, it calculates code relatively to the central pixel, so even if image is lighten or darken, relatively code will be same. So it is robust against illumination conditions.

General Regression Neural Network

GRNN consists of four layers: input layer, pattern layer, summation layer and output layer. Input nodes are fully connected to the pattern layer. GRNN estimates output by calculating the Euclidean distance between the train and test data.

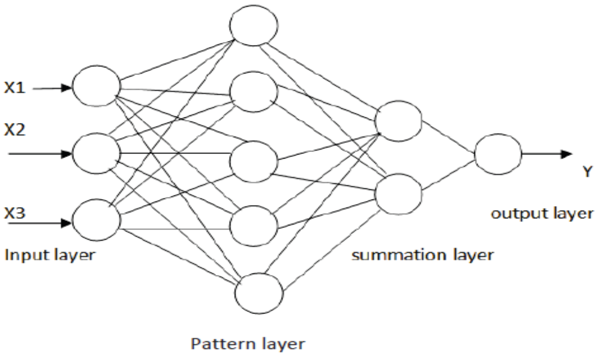


Figure 3. GRNN Architecture[2]

It uses free parameter named spread constant reduces Mean Square Error (MSE) and improve efficiency. Overall accuracy achieved is 94.25-95.48%, Which is really good. But there are contradictions in various expression classes. Neutral, happy, surprise, disgust classes are detected almost for 100% of the time, though unhappy, fear and angry classes give high rate of confusion and have come down to 62.5% for some tests.

B. Spontaneous facial micro-expression detection based on deep learning[3]

In this paper, method for automatic detection of facial micro-expressions detection is proposed. Dataset used in this method is a video dataset. Temporal Interpolation Method is used to normalize the video length.

Here, Dlib machine learning toolkit is used to detect face in each frame. Rest of the process follows as below.
After detecting face, very first and crucial part is facial landmark localization. It is done by Deep CNNs model. CNNs can transform raw pixels of an image to facial landmark positions and other related attributes. Here CNN architecture consists of convolutional layer, pooling layers, fully connected layer and loss layers.

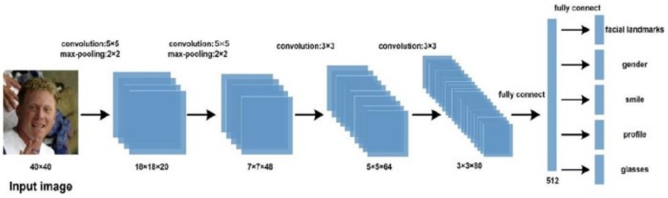


Figure 4. Architecture of CNN[3]

Convolutional layers filters the input image with 20 kernels of 5*5*3 with stride of 1 pixels. ReLU function is used as activation function. Then max pooling layers are applied. And in last fully connected layers are used to summarize all the learning. Here Euclidean loss function is used for facial landmarks and softmax loss function for other auxiliary tasks.
Concept of transfer learning is implemented here. Facial landmark localization on sparse facial landmarks, the trained model is then used to detect dense facial landmarks (68 points) , which really improves the training speed and capacity.

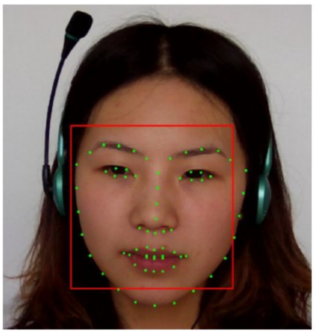


Figure 5. Detecting 68 landmarks[3]

As pipeline goes, facial region is split into 12 sub region according to 68 facial landmarks and Facial Action Coding System (FACS). These 12 regions are known as 12 ROIs (Region Of Interest).

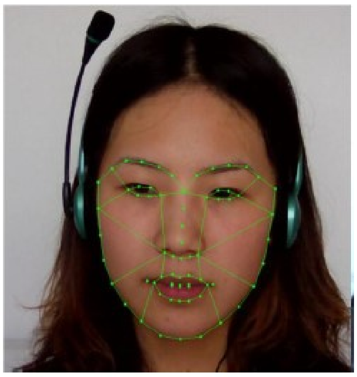


Figure 6. Detecting 12 ROIs[3]

Every ROI is then converted to HOOF (Histogram of Oriented Optical Flow) feature and normalized.
Finally Support Vector Machine (SVM) is used to classify micro-expressions.
Accuracy achieved is 80%, which is human level when it comes to detect micro-expression instead of just expressions.

C. A Robust Method for Face Recognition and Face Emotion Detection System using Support Vector Machine[4]

The proposed methodology of this paper is divided into two part : Face recognition and Emotion recognition.

Face Recognition

Author designed algorithm in such a way that if the person is recognizing for the first time then the system considers him as a new user and performs each step of operation, otherwise it is considered as Registered user and recognizes him.

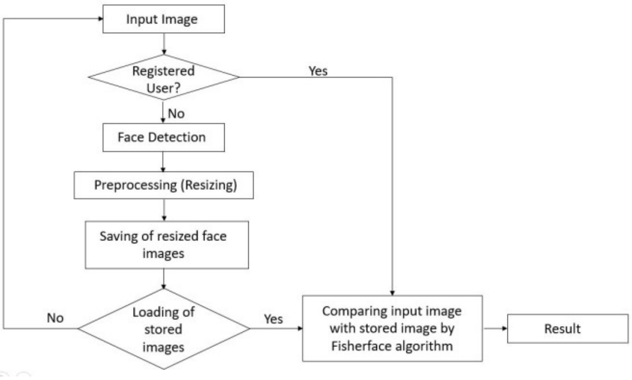


Figure 7. Face Recognition System[4]

Here OpenCV (Open Computer Vision) is used. It contains cascade classifiers in which Viola-jones algorithm is implemented. Haar-cascades is used to detect faces. It outputs image as positive or negative. Negative image is ignored for further process.

Here Fisherface algorithm is used to detect different users, so that false images can be discarded. Fisherface algorithm performs “Leave-one-out” cross validation for user identification.

Emotion Recognition

For detection of emotion, Face detection is must. So face detection part is done same as above. Then feature extraction is done using Dlib Machine Learning toolkit.

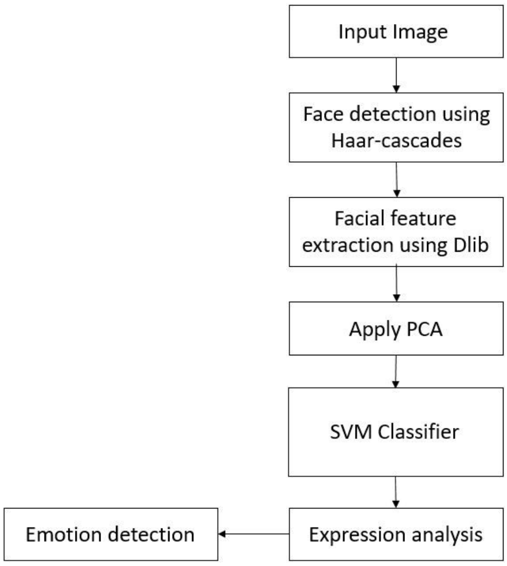

Input Image
Face detection using Haar-cascades
Facial feature extraction using Dlib
Apply PCA
SVM Classifier
Expression analysis
Emotion detection

Figure 8. Emotion detection system[4]

Principal Component Analysis(PCA) is applied to training images to reduce the dimensionality. Higher dimensions images take more time to train, so considering the sufficient quality for feature extraction 273×273 size is used. Facial features such as nose, eyes, lips, face contour are considered as keys. These keys are applied to SVM classifier. It analyses the features and labels according to that. The trained model is then used to test new images.

The method is robust as seems, but training phase is not that strong, as it is trained on CK+ database, which has very small number of labelled images. Overall running time of system is significantly less. And achieves accuracy over 90%.

D. Smile Detection Using Pair-wise Distance Vector and Extreme Learning Machine[5]

In this paper, as a feature vector for smile detection pair-wise distance vector is used. It is extracted only from points around mouth, because they convey the most detail whether face is a smile face or not. Here 68 facial points an 7 face regions are considered. From MUG Facial Expression Database, which contains 401 images of 26 subjects, has each smile-face with its corresponding neutral image. These two image of same subject is treated as a pair of images. The movements are calculated and these variances are added together for each corresponding component and portions are shown in fig. Of course, the movements around the mouth area is highest.

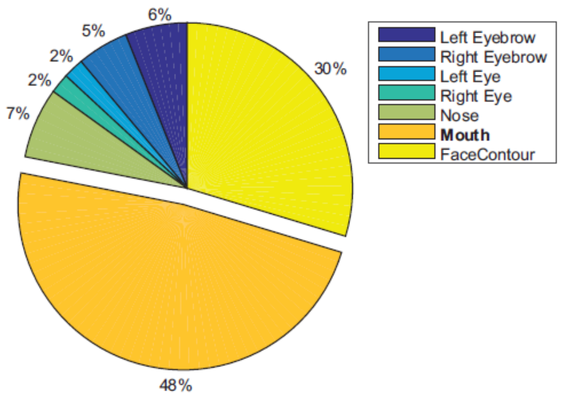


Figure 9. Impact factor of smiled face compared to neutral face[5]

Whole process is divided into 3 parts : Pre-processing, Feature extraction and Classification.

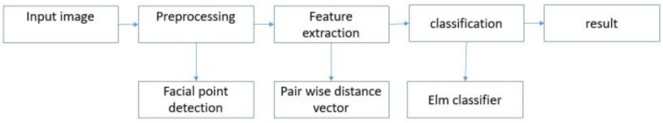


Figure 10. Smile detection system[5]

Pre-processing includes facial landmark detection, which should be accurate. State-of-the-art method named Coarse-to-Fine Auto-encoder Networks(CFAN), which is extremely fast and has achieved highest accuracy at present in terms of facial landmarking. It consists of a global Stacked Auto-encoder Network (SAN) and several local SANs. Global SAN can quickly reach the approximated location of facial landmarks and local SANs refine these locations step by step.

After pre-processing 68 points are obtained, but in training only 20 points are considered, which were around the mouth.

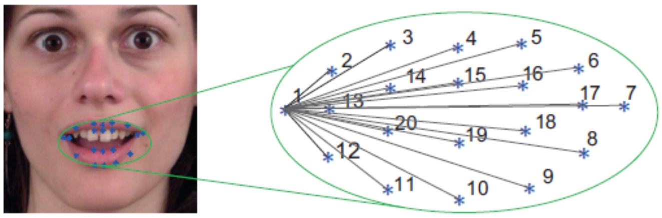


Figure 11. Detecting 20 points around the mouth[5]

Next step is feature extraction, in this part pair-wise distance vector is used. Pair-wise distance vector is an upgradation to the existing Euclidean distance matrix. This feature vector has important properties like Translation Invariance, Rotation Invariance and Scaling Invariance.

These extracted features are classified by Extreme Learning Machine, which is a single-hidden-layer feed-forward neural network. ELM assigns the parameters of the hidden nodes randomly without any iterative tuning, which makes it really faster than traditional net.

Accuracy achieved is above 90% for most of the time. Highest achieved is 94.96%, which is almost 95%. But algorithm running speed is relatively low and causes more sufficient hardware for real time implementation. Use of CFAN gives a boost to higher accuracy, CFAN is fast and most accurate facial landmark detection method as of now. By which this method achieves accurate landmarks and hence better post processing.

E. Gender and Age Classification of Human Faces for Automatic Detection of Anomalous Human Behaviour[6]

In this paper, authors aim to develop a system for automatic detection of anomalous Human behaviour. Main focus of paper is to detect Gender and Age of Human which will eventually support system to detect anomaly activity.

Convolutional Neural Nets achieves excellent performance in image feature extraction, but at the same time training a deep CNN from scratch requires vast amount of time, resources, datasets, high computational

power. In some cases it may go over days to train a deep CNN. So, the concept of Transfer Learning is used.
CNNs trained on some dataset can be tuned for a new task, even in a different domain.

As there are very few datasets available with labels of Gender and Age, Transfer learning approach is used.
Here, AlexNet – pre trained CNN is used, which has achieved excellent classification results on ImageNet competition. Input of ALexNet is 227×227 RGB image. Network has 5 convolutional layers and 3 fully connected layers.

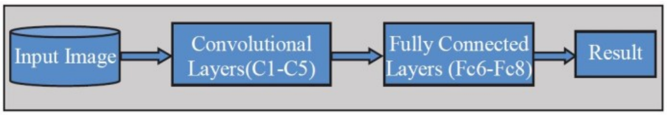


Figure 12. AlexNet architecture[6]

Feature extracted from AlexNet is classified using general classifier SVM.

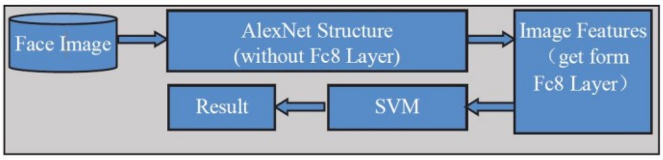

Face Image
AlexNet Structure (without Fc8 Layer)
Image Features （get form Fc8 Layer）
Result
SVM

Figure 13. AlexNet with SVM[6]

Before entering into ALexNet images are resized into fixed size 227×227, so net can perform well.
Experiments also showed that if we use Haar-like features to extracts features from image and classify them using SVM, the accuracy tents to go really low, whereas AlexNet performs outstanding.

Gender classification gives 90.33% accuracy in average and Age class gives 80.17%. Use of transfer learning concept with pre-trained CNN architecture AlexNet makes it easy to train efficiently on datasets having less images.

III. CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Network (CNN) architecture is inspired by mammalian visual cortex. Visual cortex processes images in hierarchical manner, first low level features and then high level features. CNN also works same as visual cortex, it first processes low level features of an image, such as curves, edges, then bit higher features like small part of an image and this hierarchy is continued layer by layer and in last whole image is processed.

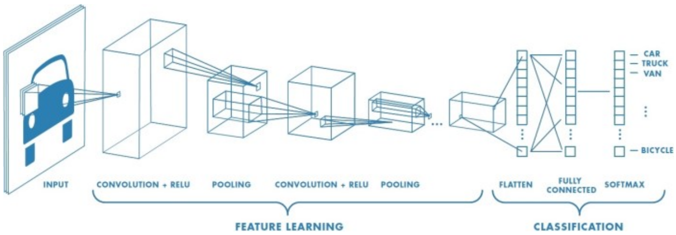


Figure 14. Basic CNN architecture

Basic layer of CNNs are convolutional layer, pooling layer, ReLU layer, Fully connected layer, loss layer, softmax layer, etc.

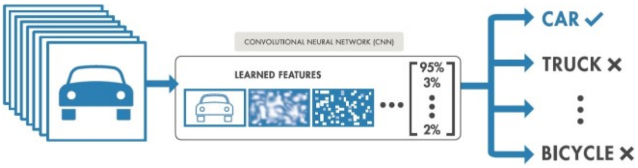


Figure 15. Result scheme of CNNs

Convolution layer takes an image as input and convolute it with feature vector or weight matrix and output more meaningful image. This image imported to other layer.
Pooling layer of CNN is used to extract most meaningful feature from each section of an image.

ReLU rectifies the image, it simply applied to check whether image gives some information or not. If image pixels are 0, the ReLU discards it.

Fully connected layer is used to sum up all learned features by connecting all neurons of previous layer to the next layer. It is usually used in later part of CNN architecture.

Loss layer in CNN, is used to apply different loss functions. For each attribute, a loss function can be applied. For example, softmax loss function is useful in multi-class classification and gives output a label with probability.

CNN is combination of these layers, not necessarily in same order.

IV. COMPARISON

Table 1

Method used	Advantages	Disadvantages	Accuracy
Adaboost by Viola-Jones+ LBP + GRNN [2]	Accuracy Especially in Neutral, Happy, Surprise, classes (mostly 100%).	Confusion rates are higher in Unhappy, fear and angry classes. 62.5 % is the lowest achieved in angry class. Gives optimal solution only on 64×64 size block.	94.25 - 95.48 %
Dlib toolkit+ CNN + Face split in 12 ROIs + HOOF Feature calculation + SVM classifier[3]	Uses transfer learning, which reduces the training time. Uses more effective features, so accuracy is good.	Eye blinking in clip is ignored as noise. Efficient only for short duration clips.	80 %
Face Recognition : Haar-cascades + Fisherface EmotiOn Detection : Haar-cascades + Dlib +PCA + SVM classifier[4]	Flow of method is good. Overall method takes significantly less time to classify image, Especially face detection part.	Trained under a dataset (CK+) having very less number of images. Noise removal is not efficient	≥ 90 %
CFAN + Pair-wise distance vector + ELM[5]	ELM outperforms SVM, Adaboost, which gives better accuracy comparatively.	Running speed of algorithm is relatively slow, so real time implementation is difficult and cause more efficient hardware.	93.42 ± 1.46 %
Pre-trained CNN – AlexNet + SVM [6]	Use of transfer learning through efficient pre-trained CNN – AlexNet gives very good results even dataset is small.	Even by using efficient feature extraction method, accuracy tents to low.	Gender : 90.33 % Age : 80.17 %

V. CONCLUSION

Based on above literature review, Comparison shows that highest accuracy for Facial Expression is achieved by GRNN classifier[2], which is an Un-supervised learning method. But it also gives higher confusion rates in some of emotion classes. 2nd highest is achieved in smile detection method using pair-wise distance vector and ELM[5], it also uses efficient facial landmark detection method CFAN[5]. On the basis of review, it can be seen that Feature extraction and classification plays major role for achieving higher accuracy. It can also be seen that parameters such as time complexity of algorithm for training, size of algorithm, etc are ignored and focused solely on accuracy rate.

VI. REFERENCES

[1] Nikhil Buduma, 2015. Fundamentals of Deep Learning.2th Edn., Sebastopol, CA., ISBN: 978-1-491-92561-4

[2] Kiran Talele, Archana Shirsat, Tejal Uplenchwar, Kushal Tuckley, “Facial Expression Recognition Using General Regression Neural Network”, IEEE Bombay Section Symposium(IBSS), 2016.

[3] Xiaohong Li, Jun Yu, Shu Zhan, “Spontaneous facial micro-expression detection based on deep learning” in IEEE, 2016.

[4] Rajesh K M, Naveenkumar M, “A Robust Method for Face Recognition and Face Emotion Detection using Support Vector Machines”, IEEE, 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques(ICEECCOT), 2016.

[5] Dongshun Cui, Guang-Bin Huang, Tianchi Liu, “Smile Detection Using Pair-wise Distance Vector and Extreme Learning Machine” in IEEE, 2016

[6] Xiaofeng Wang, Azliza Mohd Ali, Plamen Angelov, “Gender and Age Classification of Human Faces for Automatic Detection of Anomalous Human Behavior” in IEEE, 2017

[7] Jiaxing Li, Dexiang Zhang, Jingjing Zhang, Jun Zhang, “Facial Expression Recognition with Faster R-CNN”, Science Direct, Procedia Computer Science 107 (2017) 135 – 140, 2017.

[8] Nazima kauser, Jitendra Sharma, “Automatic Facial Expression Recognition: A Survey Based on Feature Extraction and Classification Techniques”, IEEE, 2016.

[9] Shubhada Deshmukh, Manasi Patwardhan, Anjali Mahajan, “Survey on real-time facial expression recognition techniques”, IET (The Institute of Engineering and Technology) Journals, pp. 1-9, 2016.

[10] Kewen Yan, Shaohui Huang, Yaoxian Song, Wei Liu, Neng Fan, “Face Recognition Based on Convolutional Neural Network”, IEEE, Proceedings of the 36th Chinese Control Conference, 2017.