

Modular Facial Expression Recognition using Double Channel DNNs

Sujata^a and Suman K. Mitra

Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India

Keywords: CNN, DNN, VGG16, HOG, HSOG, SVM, KNN.

Abstract: Recognizing human expressions is an important task for machines to understand emotional changes in humans. However, the accurate features that are closely linked to changes in expression are difficult to extract due to the influence of individual differences and variations in emotional intensity. The modular approach presented here imitates the human being's ability to identify a person with a limited facial part. In this article, we demonstrate experimentally that certain parts of the face, such as the eyes, nose, lips and forehead, contribute more to the recognition of expressions. A combination of two deep neural networks is also proposed to extract the characteristics of the facial images provided. Two preprocessing approaches are implemented, Histogram Equalization (to handle illumination) and Data Augmentation (increasing number of facial images), to restrict the regions used for recognition of the facial expression. Two-channel architecture used for implementation, one channel accepts input as a grayscale face image, processed by VGG16_ft (fine-tuned VGG16), and another channel accepts input as histograms face image. the second order gradients (HSOG), processed from the proposed CNN model and extracts the characteristics accordingly. Then concatenate the characteristics from the two channels. The final recognition result is calculated using the SVM and KNN classifiers. Experimental results indicate that the proposed algorithm is able to recognize six basic facial expressions (happiness, sadness, anger, disgust, fear and surprise) with great precision. Fine tuning is effective for FER activities with a well-trained model if there are not enough samples to collected.

1 INTRODUCTION


From the human facial images FER (Facial Expression Recognition) is trying to predict the basic face expressions like Happy, sad, Surprise, Fear and Disgust. Just by analyzing the face images the method helps the machine to understand the human emotions and intention. It has gained lot of attention because of its potential applications like, computer interfaces, health management, autonomous driving, detecting abnormal human behavior and other similar tasks.

Histogram Equalization (HE) and Data augmentation (DA) are pre-processing techniques, that are required for the facial images provided to make machine learn from images. HE is a simple but effective technique in image processing, which could build the distribution of gray values in numerous images more uniformly and reduce the interference caused by lighting. CNN needs huge data sets to generalize a particular problem. FER databases which are available publicly have not sufficient images to handle the problems. For creating synthetic images from the origi-

nal face image, one researcher Simard et al. (Simard et al., 2003) suggested the DA procedure to extend the datasets.

Despite recent rapid developments, FER remains challenging due to some factors such as lighting, head deflection and some occlusions in facial regions. These impedances can affect facial recognition performance and reduce the accuracy of FER. As demonstrated in the past, hand-craft features seems to be no longer appropriate for expression recognition activities with critical issues. Fortunately, the Deep Neural Network (DNN) is providing a satisfactory solution to these problems which have not been able to comply with hand craft techniques.

Humans have the capability to recognize a person with a limited facial regions. On account of acknowledgment of facial expressions, the utilization of full-face images can be repetitive since facial expression fundamentally misshapes certain specific zones of face images. There is one algorithm called as facial benchmark detection algorithm which is offered by Dlib helps to extricate the facial regions from the given face image. This Dlib is an open source machine learning library provided by King (King, 2009).

^a  <https://orcid.org/0000-0003-4166-1502>

That gives us 68 landmarks points on the face. Using those landmarks points, we are able to extract the regions of face like forehead, eyes, nose, and lips. Our proposed frame focuses only on these parts of the face, but to verify the efficiency of the proposed frame, we also performed experiments with the complete facial image.

Suggested framework is focuses on the double-channel architecture that simultaneously processes the grayscale face image, and the HSOG (histogram of second-order gradients) face image. Fig.2 shown, pre-processing steps such as histogram equalization (HE) (for handle illumination) and DA (Data Augmentation) (create synthetic images) are necessary for the input facial images provided. The detailed calculation of the HSOG is reported in the section. 3. HSOG is the variant of the Histogram of oriented gradient (HOG), as indicated in (Dalal and Triggs, 2005). HSOG extracts local information from the face image. DNNs are used for various channels grayscale and HSOG facial images. In one channel, a proposed VGG16_ft with original parameters acquired as in VGG16, which was trained in ImageNet, is created for grayscale facial images to extract features related to facial expression. On the other channel, HSOG facial images, a proposed two-layer CNN, which refers to the development of DeepID (Sun et al., 2015). The output of the two channels are concatenated and made an enormous feature vector. At long last, SVM and KNN with various separation estimations are utilized for classification to anticipate basic facial expressions (anger, happiness, sad, disgust, fear). To show its viability, FER databases JAFFE Database (Lyons et al., 1998), VIDEO Database (Shikkenawis and Mitra, 2016), CK+ Database (Lucey et al., 2010) and Oulu- Casia (Zhao et al., 2011) is used to test the framework in a modular way. This modular way is another significant commitment to current work. In summary, the commitments of this work are featured as follows:

- Firstly, Modular Approach (Where we extract the facial region automatically from the full face).
- Secondly, double channels of facial images, including grayscale images and their corresponding HSOG images are used for FER because of their complementary properties. As far as DNN is not trained with HSOG images.
- Thirdly, the fine-tuning methodology is used to make full utilization of a very much learned pre-trained VGG16 model (VGG16 model trained on ImageNet).

- At last, outputs of the two channels are combined to predict a vigorous outcome. Four benchmarking datasets and a few handy facial images are utilized to assess the successfulness of our work.

The rest of the article is organized as follows. The 2 section provides details of the proposed framework. The 3 section shows the results and analysis of the experiment. Section 4 Concludes the study.

2 OUR PROPOSED METHOD FOR FER

In this section, we mainly portray the premises of our procedure and propose the structure, which improves the effectiveness and precision of the recognition of facial expressions. As referenced above, we utilize the modular approach where we just take the forehead, eyes, nose and lips. So starting now and into the foreseeable future we work in these four facial regions. The proposed FER procedure utilized in this paper depends on a double channel design prepared to do viably recognizing expressions. The figure .2 shows the methodology of the proposed framework, which is divided into three stages: 1) Pre-processing, 2) Dual channel feature extraction technique 3) Classification by SVM and KNN.

2.1 Pre-processing

Before facial recognition, the image must be preprocessed. Our preprocessing begins with the transformation of the input image in grayscale. This process minimized the variation in facial images. This pre-processing is a necessary step because the CNN illustrated below provides that the image of the 3-channel input face and the image of the received grayscale face can be represented within the 3 channels. Next, we perform two procedures, which are Histogram Equalization (lighting management) and Data Augmentation (increasing the number of the face image in the database). The next section describes each of these steps in detail.

2.1.1 Histogram Equalization

In facial images, some problems should also be considered. Due to the different lighting conditions, when taking images, the segments of the face will be displayed with various brightness, which can cause enormous interference in the results of facial recognition. Therefore, we tend to perform histogram equalization (HE) before recognition. HE is a simple but effective technique in image processing, which

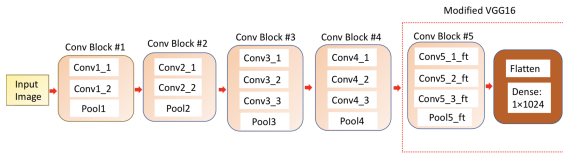


Figure 1: Framework for the modified VGG16_ft network, utilized for extraction of the facial expression features from the given face image.

could build the distribution of gray values in numerous images more uniformly and reduce the interference caused by illumination.

2.1.2 Data Augmentation

CNN needs immense data sets to generalize a specific issue. In any case, freely accessible FER datasets need more images to deal with the issue. Data augmentation strategy to expand the data sets by making manufactured facial images for every unique facial image. Enlivened by this technique, the accompanying activities were utilized as data augmentation: 1) flip the image vertically and horizontally 2) Rotate each image in the dataset, rotate it right angle if the face image is square and turn it by 180^0 if it is an image it is rectangular 3) Add arbitrary noise to the landmarks so as to present little deformations on the faces. Thusly, the subsequent face image is unique in relation to the original face the one utilized with CNN for pre-training. This distinction among processed and non-processes data could influence results. This might be because of the way that the network has learned the features of the original face image and will most likely be unable to extract the features from the processed face images. Consequently, we likewise give results with a network trained with original facial images.

2.2 Feature Extraction from Grayscale Facial Images

The absence of satisfactory raining samples restricts the execution of the CNN FER approach. Expanding data can partly deal with the issue of over-fitting. Hence, fine-tuning is utilized to extract the expressions features from the input face through the deep neural network (DNN) which has made over the top progress in comparable errands.

Our proposed system utilizes DNN for the extraction of expressions features for FER dependent on the VGG neywok introduced by Simonyan and Zisserman (Simonyan and Zisserman, 2014). They accompany two versions of VGG: VGG-16 and VGG-19(i.e. sixteen and nineteen layers separately). VGG16 (Sujata and Mitra, 2020) was picked for its successful

execution in visual recognition and fast convergence. It has 138 million parameters and contains 13 convolutional layers, followed by 3 fully connected layers (FC). The initial two fully connected (FC) layers have 4,096 outputs and the last layer has 2,622 outputs. Since the VGG network isn't intended for FER tasks, we adjust the structure as per our necessities. Figure. 1 shows the fundamental module of the network. Contrasted with the original VGG16, our VGG16_ft (where "ft" implies fine-tuning) is disentangled by eliminating two dense layers. The input size for the forehead is 54×48 , for the eyes it is 39×117 , for the nose it is 50×55 and for the lips it is 48×74 .

Now, we fix the structures of the four initial CONV (convolution) blocks of VGG16_ft. However, we change the structure of the fifth CONV block of VGG16_ft and furthermore change the names of each layer simply by adding "ft" to the end of the original layer name. So the name of the fifth CONV block layer resembles CONV_5_1_ft. The parameters of the layer are appeared in the Table. 1. In view of the experiments, the last dense layer was saved and set its size to 1×1024 . That dimension is actually the extracted feature of input image denoted as feature vector "fv_1_1" for the forehead, "fv_1_2" for the eyes, "fv_1_3" for the nose and "fv_1_4" for the lips. We have diminished the learning rate of the layers that have a spot with the fifth CONV block 10 times (the learning rate for the fifth CONV block is .001) of the other block learning rate(.01 utilized for other CONV blocks) to guarantee that we will learn more certain information. At last, the initial part of the system is initialized with the weights of the VGG16 model, which is trained in the Imagenet dataset. ReLu (Rectified linear unit) is applied after each convolutional layer.

Table 1: Parameters set for fifth block.

	CONV_5_1_ft	CONV_5_2_ft	CONV_5_3_ft	POOL_5_ft
Filters	256	256	512	
size	7×7	3×3	3×3	2×2
stride	1	1	1	2
pad	3	0	0	0

2.3 Feature Extraction from HSOG (Histograms of the Second Order Gradients) Facial Images

To the best of our knowledge, there is no model trained on the HSOG images. So here first compute the HSOG facial images.

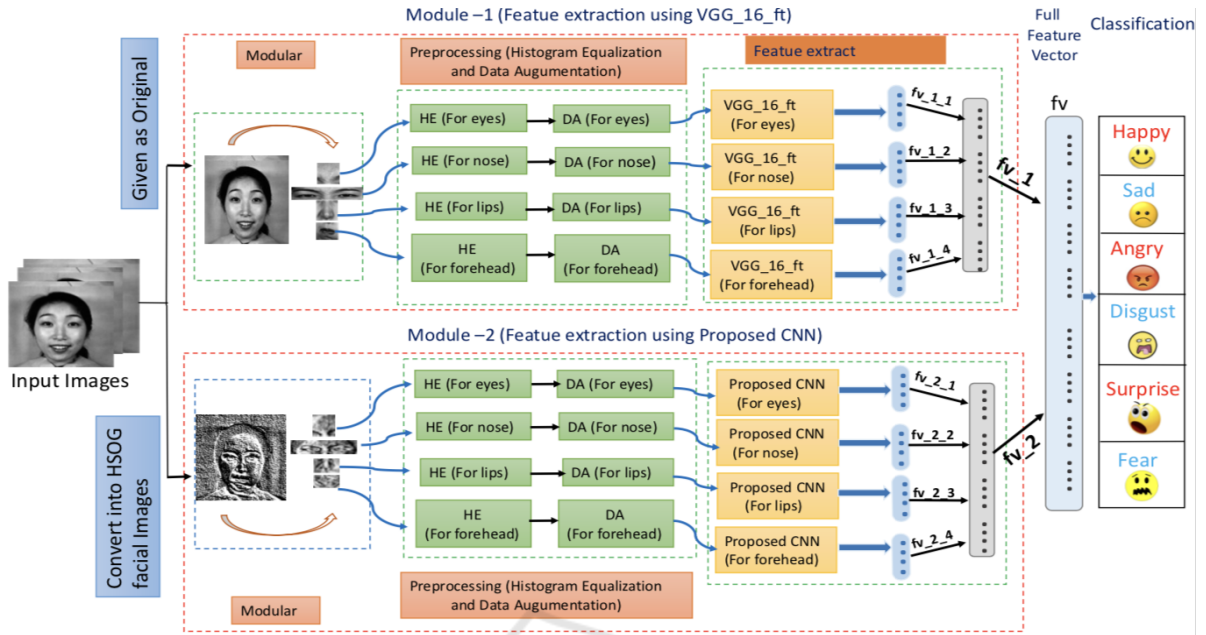


Figure 2: Illustration of the proposed Framework.

2.3.1 Computation of First Order Oriented Gradient Maps (OGMs)

Image descriptor begins from computing the 1st order oriented gradient map (OGM). The initial step of estimation in many feature detectors in image preprocessing is to guarantee normalized color and gamma values. Image preprocessing gives little effect on execution. The initial step of count is the calculation of the gradient values. The most widely recognized strategy is to apply the 1-D centered, point discrete derivative masks in either of the horizontal and vertical directions. In particular, this strategy requires filtering the color or intensity information of the image with the following filter kernels:

$[-1, 0, 1]$ and $[-1, 0, 1]^T$. Resulting image is denoted as “ G_1 ”. After that apply the convolution of these gradient maps G with the Gaussian kernel “ G_2 ”. Defined as

$$G = G_1 \times G_2 \quad (1)$$

Change in the image contrast in which the intensity values are multiplied by the constant will result in the multiplication of the gradient computation. These properties will be important for actualizing the image descriptor for outward appearance acknowledgment. Utilizing those 1st OGM “ G ”. Figure the 2nd Order Gradient in the subsequent stage.

2.3.2 Computation of Second Order Gradient (OGMs)

Once 1st OGM is computed, they are use as inputs to the 2nd order gradient calculation over the image region I . In each pixel location OGM compute the gradient magnitude Mag and gradient orientation Φ as

$$Mag(x,y) = \sqrt{\left(\frac{\partial G(x,y)}{\partial x}\right)^2 + \left(\frac{\partial G(x,y)}{\partial y}\right)^2} \quad (2)$$

$$\Phi(x,y) = \arctan\left(\frac{\frac{\partial J_G(x,y)}{\partial y}}{\frac{\partial G(x,y)}{\partial x}}\right) \quad (3)$$

$$\frac{\partial G(x,y)}{\partial x} = G(x+1,y) - G(x-1,y) \quad (4)$$

$$\frac{\partial G(x,y)}{\partial y} = G(x,y+1) - G(x,y-1) \quad (5)$$

Orientation Φ exists in scope of $[-\Pi/2, \Pi/2]$. Map orientation from $[-\Pi/2, \Pi/2]$ to $[0, 2\Pi]$. A pivotal issue to be managed when processing the second order gradients is the affect-ability of the resultant local image descriptor as for noise. The truth of utilizing the Gaussian kernel to simulate human simple cells and smooth first order gradients by gives descriptor a desirable robustness to noise.

As indicated by our insight into up until this point, no current model is trained on the HSOG images.

Thus, we build two layer CNN model that automatically extracts the features from the HSOG facial images.

Fig. 3 illustrates the proposed CNN structure, which consists input layer, two convolution layers C_1 and C_2 and two sub-sampling layers S_1 and S_2. All parameters utilized in the proposed CNN are listed in Table 2.

Table 2: parameter set for the proposed CNN.

	C_1	S_1	C_2	S_2
Filters	64		256	
size	7×7	2×2	3×3	2×2
stride	1	2	1	2
pad	3	0	0	0

Now, the output is given to the fully connected layer (Dense layer) with the 1024 neurons. From here extract the feature vector (fv_2) of size 1×1024 . To deal with the nonlinear data, include the "Relu" activations after the S_1 and S_2 layers. At that point we extricate the feature vector "fv_2.1" for the forehead, "fv_2.2" for the eyes, "fv_2.3" for the nose and "fv_2.4" for the lips.

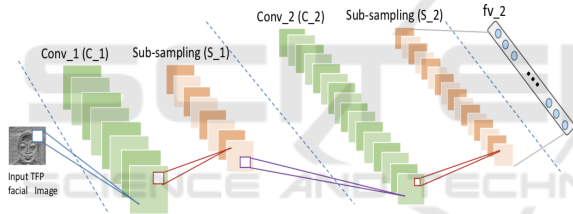


Figure 3: Framework for the proposed CNN used for extraction of the expression features from the HSOG facial images.

2.4 Concatenation of Different Outputs and Classification

Feature vector that came from the forehead (fv_1.1), eyes (fv_1.2), nose ((fv_1.3) and lips (fv_1.4) are concatenated and make a long feature vector for facial expression. These are the features that originate from grayscale images utilizing VGG16.ft with fine-tuning method. Likewise feature vector fv_2 is a concatenation of the feature vector came from the forehead (fv_2.1), eyes (fv_2.2), nose (fv_2.3) and lips (fv_2.4). These fv_2 features came from the HSOG facial images utilizing proposed CNN architecture. At long last, we get full feature vector "fv" that is the blend of the fv_1 and fv_2. Will go in Next advance for classification.

In the classification cycle, the comparability between extracted features of the display setting and the test set is assessed by the SVM and K nearest neighbor

(K=1,2,3) classifier with different separation measures.

The Support Vector Machine (SVM) algorithm is applied to classification. At the point when all local image descriptors are changed to a fixed length feature vector, similarity is processed to measure the similitude between each pair of the feature vectors. At last, each image for the test is classified into an object class with the greatest SVM output decision value. We tune the parameters of the classifier on the training set, and acquire the recognition accuracy on the test set.

Other classifier is K nearest-neighbor (K=1,2,3) classifier with various distance measures. Euclidean distance, Chi-square distance, as well as histogram intersection (HI) are utilized in our experiments. Mean Square error (MSE) is used as for the computation of loss for the SVM and KNN.

3 EXPERIMENTAL ANALYSIS AND RESULTS

To support the theoretical finish of the proposed framework, tests have been conducted on some genuine datasets, as archived in this section.

Experiments of FER have been conducted on the four FER datasets. Facial images for the most part incorporate extremely huge dimensions. Managing such broad information ends up being very hard for machines. Hence, the modular methodology is applied when just certain data areas of the face are thought of. Facial expression give signals of the individual's emotional state, even without verbal correspondence. The eyes are the most open aspect of an individual's face and reveal a lot about their sentiments. Not with standing the eyes, lips, forehead, nose, and so forth they are also information regions. During the expression analysis task, we saw that, in addition to the eyes, nose and lips/mouth, the forehead additionally assumes a significant role regarding expressions. Most FER strategies are presently applied to full face images. This article focuses in just on some information area of the face, as talked about. To make a correlation, we did the holistic experiments (where the full face image was utilized) as well as modular.

To see the effectiveness of our method, our FER methodology works under Keras on the macOS Mojave system platform. To make the assessments correct and effective, 4 benchmark datasets were used, consisting of facial images. Representations of the datasets used listed below.

3.0.1 JAFFE

JAFFE (Lyons et al., 1998) database having 213 facial images of 10 Japanese female models of 7 facial expressions (6 basic facial expressions + 1 neutral). Out of 213 images, random 140 images were chosen for training and the remaining 73 used for testing. Fig. 4 shows the trends of accuracy and loss during the training and testing with the increase in iterations. Table 3 shows reported the average accuracy in the Holistic as well as modular approach both. In the JAFFE dataset average accuracy is 95.67%.

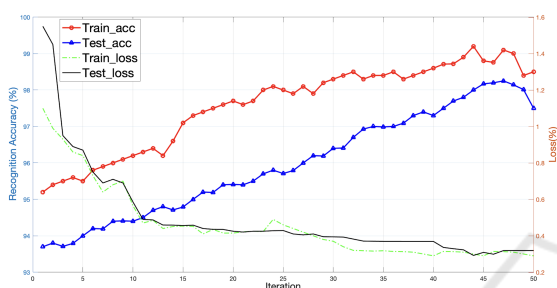


Figure 4: Curves of Accuracy and Loss during training and testing phases for JAFFE dataset.

3.0.2 Video

The Video (Shikkenawis and Mitra, 2016) database has videos of 11 persons. The single video contains four different facial expressions: Smiling, Angry, Open mouth, and Normal. Out of 6668 images, randomly 70% images were chosen for training and remaining 30% images used as a testing. Out of 6668 images, randomly 70% images were chosen for training and remaining 30% images used as testing. Fig. 5 shows the trends of accuracy and loss during the training and testing with the increase in iterations. Table 3 shows reported the average accuracy in the Holistic as well as modular approach both. In the VIDEO dataset average accuracy is 97.77%.

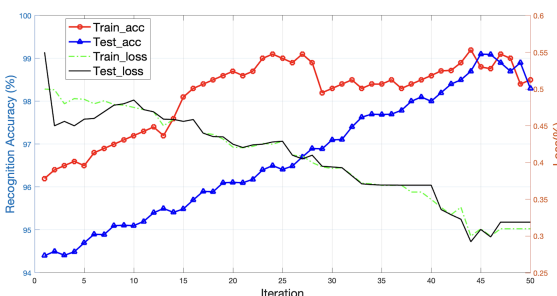


Figure 5: Curves of Accuracy and Loss during training and testing phases for VIDEO dataset.

3.0.3 CK+

In CK+ (Lucey et al., 2010) there are 593 sequences across 123 subjects giving 8 facial expressions. All sequences are captured from the neutral face to the peak expression. Participants were 18 to 50 years of age, 69% female, 81%, Euro-American, 13% Afro-American, and 6% other groups. This paper uses image sequences of 99 subjects with 7 facial expressions. Fig. 6 shows the trends of accuracy and loss during the training and testing with the increase in iterations. Table 3 shows reported the average accuracy in the Holistic as well as modular approach both. In the VIDEO dataset average accuracy is 94.78%.

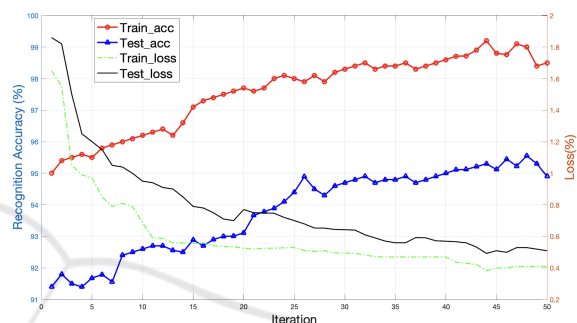


Figure 6: Curves of Accuracy and Loss during training and testing phases for CK+ dataset.

3.0.4 Oulu-Casia

Oulu-Casia (Zhao et al., 2011) has 6 facial expressions (anger, happiness, surprise, fear, disgust and sad) from 80 different subjects between 23 to 58 years of age. 73.8% of the persons are males. Out of 3360 images randomly 70% images were chosen for training and remaining 30% images used as testing. Fig. 6 shows the trends of accuracy and loss during the training and testing with the increase in iterations. Table 3 shows reported the average accuracy in the Holistic as well as modular approach both. In the Oulu-Casia dataset average accuracy is 95.86%.

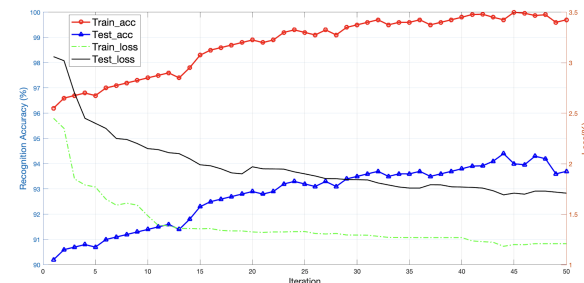


Figure 7: Curves of Accuracy and Loss during training and testing phases for Oulu-Casia dataset.

Table 3: Comparison between the Holistic and Modular approach in our proposed framework in the light of SVM and KNN as the classifier for all datasets (In terms of average accuracy (%) reported for 50 iterations).

Datasets	SVM	Holistic			Modular			
		KNN			SVM	KNN		
		Euclidean	Chi Square	Histogram Intersection		Euclidean	Chi Square	Histogram Intersection
JAFFE	90.02	87.32	81.42	80.02	95.67	92.14	88.32	86.97
VIDEO	91.55	87.43	79.89	77.12	97.77	91.54	88.12	86.91
CK+	90.15	87.71	86.41	83.54	94.78	91.24	86.79	85.64
OULU-CASIA	89.40	86.30	79.79	75.20	95.86	88.76	84.61	82.20

3.1 Expressions at Different Intensity Rates

There are several models regarding the character of emotion and define how it is represented in the body and brain. The goal is to decide the distinctive level of emotions. With this approach, we find the predominant emotion as well as the emotion rates presented on the face. Here we introduce the new strategy to analyze the level of emotion as you move from one phase of emotion to the next higher state. Some of the emotions that have been influenced by the changes in the time interval, as shown in the graphs below in Fig. 8, 9, 10 and 11. Some basic variations in the proportion of emotions with a different time interval are also represented. Our methodology is incredibly useful for exploring micro expressions.

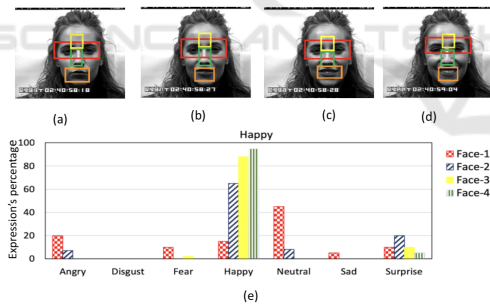


Figure 8: (a) (b) (c) (d) Shows the expression of the Happy face from the lower level to the extreme level, and (e) Graphical representation of expression percentages and how other expressions influence while the expression level changes from low to high.

4 CONCLUSION

In this study, we investigate the FER technique primarily based on double channel architecture that processes the grayscale facial image and HSOG facial image at the same time. Both image channels which are utilized are complementary, and capture local and global information from the given grayscale and

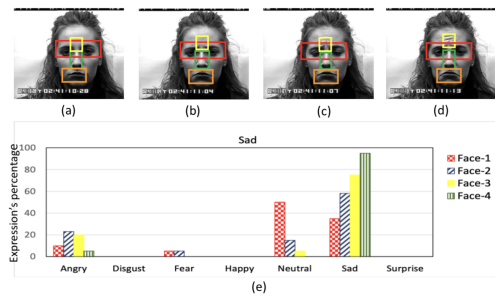


Figure 9: (a) (b) (c) (d) Shows the expression of the Sad face from the lower level to the extreme level, and (e) Graphical representation of expression percentages and how other expressions influence while the expression level changes from low to high.

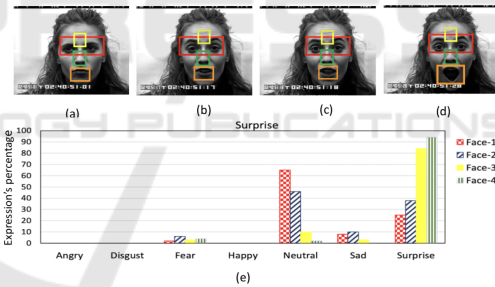


Figure 10: (a) (b) (c) (d) Shows the expression of the Surprise face from the lower level to the extreme level, and (e) Graphical representation of expression percentages and how other expressions influence while the expression level changes from low to high.

HSOG facial image. It can enhance the recognition capacity. Concatenation strategy is proposed to completely use the features that are extracted from both image channels (VGG16_ft and proposed CNN). VGG16_ft has automatically extracted the features from the given grayscale face images. A proposed CNN is built to automatically extracts the features from the HSOG facial images as of the pre-trained model is not trained on HSOG facial images. Furthermore, concatenated features have been classified using SVM and KNN classifier with different distance measures.

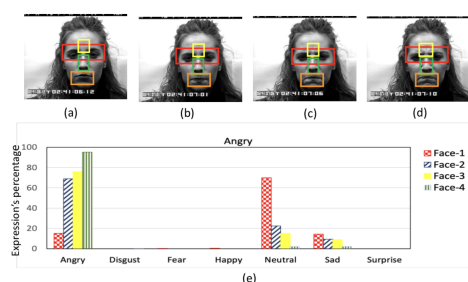


Figure 11: (a) (b) (c) (d) Shows the expression of the angry face from the lower level to the extreme level, and (e) Graphical representation of expression percentages and how other expressions influence while the expression level changes from low to high.

Capability of our proposed method is to recognize a facial expression of a person using partial information from the given whole face image. The proposed method is applied to the most informative regions of the face, i.e., forehead, eyes, nose, and lips. It is observed that a combination of these regions is useful enough to distinguish facial expressions of different persons or the same persons in most of the cases. The result obtained by the proposed method is comparable with the most of the state of the art methods.

REFERENCES

- Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.
- King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758.
- Liu, C. and Wechsler, H. (2002). Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image processing*, 11(4):467–476.
- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). 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*, pages 94–101. IEEE.
- Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. (1998). Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE.
- Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE TRANSACTIONS on Information and Systems*, 74(10):3474–3483.
- Shikkenawis, G. and Mitra, S. K. (2016). On some variants of locality preserving projection. *Neurocomputing*, 173:196–211.
- Simard, P. Y., Steinkraus, D., and Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *null*, page 958. IEEE.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sujata and Mitra, S. K. (2020). Dnnfg: Dnn based on fourier transform followed by gabor filtering for the modular fer. In *ICPRAM*, pages 212–219.
- Sun, Y., Liang, D., Wang, X., and Tang, X. (2015). Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*.
- Valstar, M., Pantic, M., and Patras, I. (2004). Motion history for facial action detection in video. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 1, pages 635–640. IEEE.
- Walecki, R., Rudovic, O., Pavlovic, V., and Pantic, M. (2015). Variable-state latent conditional random fields for facial expression recognition and action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE.
- Zhang, T. (2017). Facial expression recognition based on deep learning: A survey. In *International Conference on Intelligent and Interactive Systems and Applications*, pages 345–352. Springer.
- Zhang, W., Zhang, Y., Ma, L., Guan, J., and Gong, S. (2015). Multimodal learning for facial expression recognition. *Pattern Recognition*, 48(10):3191–3202.
- Zhao, G., Huang, X., Taini, M., Li, S. Z., and Pietikäläinen, M. (2011). Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619.
- Zhong, L., Liu, Q., Yang, P., Huang, J., and Metaxas, D. N. (2014). Learning multiscale active facial patches for expression analysis. *IEEE transactions on cybernetics*, 45(8):1499–1510.