

DNNFG: DNN based on Fourier Transform Followed by Gabor Filtering for the Modular FER

Sujata^a and Suman K. Mitra

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

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

Abstract: The modular approach mimics the capability of the human brain to identify a person with a limited facial part. In this article, we experimentally show that some facial parts like eyes, nose, lips, and forehead contribute more in the expression recognition task. Deep neural network, VGG16_ft, is proposed to automatically extricate features from the given facial images. Fine-tuning is very fruitful to the FER (Facial Expression Recognition) with pre-trained models, if sufficient facial images are not collected. Two preprocessing approaches, Fourier transform followed by Gabor filters and Data Augmentation (DA), are implemented to restrain the regions used for Facial expression recognition (FER). The features from four facial regions are concatenated and classification is done using SVM and KNN (with different distance measure). The experimental result shows that the proposed framework can recognize the facial expressions like happy, anger, sad, surprise, disgust and fear with high accuracy for the benchmark datasets like “JAFFE”, “VIDEO”, “CK+” and “Oulu-Casia”.

1 INTRODUCTION

Many of the existing techniques conduct the facial expression recognition supported by the image/image sequences, while not considering temporal data due to the convenience of data handling and the easy accessibility of training and testing material. Training the deep neural networks with small FER datasets leads to overfitting. To moderate this issue, several studies use further task-oriented information to pre-train their networks from fine-tuned or scratch on existing pre-trained models like VGG (Simonyan and Zisserman, 2014), GoogleNet (Szegedy et al., 2015) and AlexNet (Krizhevsky et al., 2012). Kahou et al. (Kahou et al., 2013), (Kaneko et al., 2016) demonstrated that the utilization of extra information can enable models with high capacity without overfitting, accordingly improves the FER performance.

Based on the traditional CNN architecture, several studies have proposed the addition of well-designed auxiliary layers or blocks to improve the potential of the features learned from the expression. HoloNet (Yao et al., 2016) was destined to FER, which is based on the new architecture of CNN, where the residual structure is combined with CReLU (Shang et al., 2016) to extend the depth of the network without de-

creasing the competition. Another network of CNN SSE (Supervised Scoring Ensemble) (Hu et al., 2017) extends the degree of supervision of the FER. And an FSN (feature selection network) introduced the incorporation of a feature option into AlexNet, which automatically filters irrelevant features and focuses on related functionality from maps of facial features learned. Previous analyzes had indicated that more network assemblies would defeat a single network. Many existing FER networks have specialized in one task and learned expressive-sensitive characteristics without considering interactions with various factors. In any case, in reality, FER is intertwined with different variables, for example the posture of the head, the identity of the subject and illumination. Reed et al. (Reed et al., 2014) has developed a Boltzmann machine with different coordinates for factors relevant to facial expressions.

The works of (Devries et al., 2014) (Pons and Masip, 2018) have suggested that at the same time performs FER with different tasks, locate facial milestones and detect units of facial action (AU) (Ekman and Rosenberg, 1997), can mutually enhance the execution of FER. In (Zhao et al., 2015), deep belief networks (DBNs) were first trained to detect faces and then initially identify areas related to facial expression. At that point, these analyzed face segments were grouped by a stacked automatic encoder. In (Ri-

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

fai et al., 2012), it was proposed that CCNET would acquire LTI representations (Local translation invariant).

For the recognition of the invariant expression of posture, Lai et al (Lai and Lai, 2018) proposed the GAN (Generative adversarial networks)-based frontalization system, whenever the generator frontal the input face images whereas conserving the identity and expression attributes and the discriminator recognizes the original face images from the produced face images. Also, Zhang et al.(Zhang et al., 2018) proposed the GAN model that produces images with various facial expressions under discretionary postures for multi-view FER.

The objective of this article is more basic, but also more general, namely: can recurring connectivity from associative areas to perceptive areas be useful for classifying expressive events? Our hypothesis is that the deep connectivity of the neural network offers an advantage in recognizing and anticipating more ambiguous expressions. For example, at the beginning of a sequence composed of expressions of neutral with higher intensity. To validate this very general hypothesis using computational models, we compare the simplest and comparable types of deep neural networks to test the importance of recurrent connections, with everything as similar as possible (ie identical learning rate, synaptic weight correction, procedure of training / test, etc.).

Human's have the capability to identify a person with a limited facial part. To extract these facial parts from the face we have used the Facial Landmark Detection algorithm offered by Dlib which is an open source machine learning library. The facial landmark detection algorithm offered by Dlib is an implementation of the Ensemble of Regression Trees. It utilizes the technique of pixel intensity difference to directly estimate the landmark positions. The algorithm has a very fast response rate and detects a set of 68 landmarks on a given face. The landmarks (key points) of our interest are those that describe the forehead, eyes, nose and lips. Using the landmarks of the eyes, eyebrows and nose we find the upper patch between two eyes which we have called the forehead. Using the landmarks of the eyes we have extracted the eye region and similarly the nose and lips. These patches are cropped out for each face image and saved. These regions are selected as they give most of the information about the expressions as proved in (Taheri et al., 2014).

Our proposed framework is primarily based on the architecture that processes the gray-scale facial regions of the input face image as shown in Fig. 1, some preprocessing steps such as Fourier transform fol-

lowed by Gabor filters and Data Augmentation (AU) (for increasing the number of training samples) are essential for given facial images. Proposed VGG16_ft uses the original parameters acquired from the pre-trained VGG16 which is trained on ImageNet dataset of grayscale facial images to extract the facial expression related features. Outputs from all the regions are concatenated into a large feature vector. Finally, SVM and KNN with different distance measures are used for the classification, to predict the basic facial expressions (happiness, anger, sadness, disgust, surprise, and fear).

To show its efficiency, proposed framework is tested on well-known facial expression databases like 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) in modular way. That is similar to the Ekman FACS (Facial Action Coding System)(Ekman and Friesen, 1976). This modular approach is another main contribution of present work. Instead of taking the full face some significant portion (forehead, eyes, nose, and lips) of the face image is used.

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

2 PROPOSED ARCHITECTURE

In this segment, we discuss the premise of our technique and proposed framework which improves the efficiency and accuracy of the facial expression recognition. As mentioned earlier in the modular approach, now onward we only consider forehead, eyes, nose, and lips regions. Fig. 1 demonstrates the procedure of the proposed framework which is divided into three phases - 1) Preprocessing 2) Feature extraction 3) Classification using SVM and KNN.

2.1 Preprocessing

We used the little similar preprocessing as in the EMPATH model given by Dailey et al. (Dailey et al., 2002). Before the facial recognition, some image preprocessing need to be done first. Our preprocessing starts with the transformation of the input facial image to grayscale. This process minimized the variation of face images. This is a necessary step because CNN depicted later expects 3 channel input facial image, this grayscale facial image is depicted within the 3 channel. Subsequently, we run two procedures, Fourier transforms followed by Gabor filters

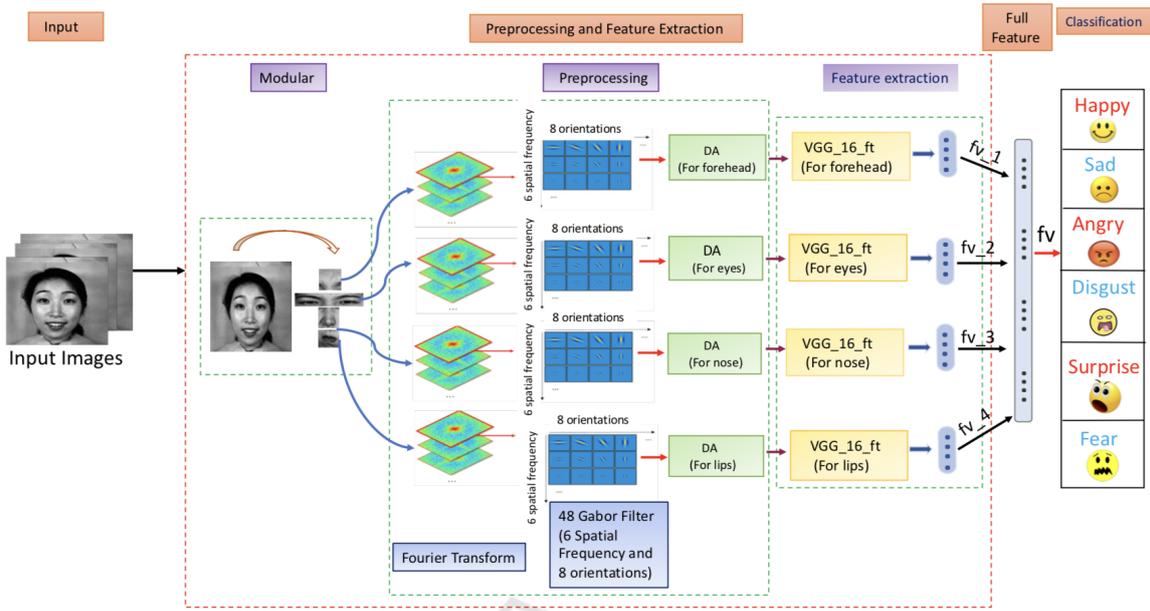


Figure 1: Illustration of the proposed Framework.

(improve the speed and encodes the edges) and Data Augmentation (increase number of face images in the database). The subsequent section describes each of those steps in details.

2.1.1 Fast Fourier Transform (FFT) and Gabor Filtering

Fast Fourier transform can speed up our procedure very smoothly. Computation of the 1 Dimensional (1D) Fourier transformation of N points specifically requires the order of N^2 addition/multiplication operations. Whereas Fast Fourier Transform (FFT) fulfills the same task in $N \log N$ operations. 2D Fourier transform is computed by the given equation-

$$F(p, q) = \frac{1}{MN} \sum_{r=0}^{M-1} \sum_{s=0}^{N-1} f(r, s) e^{-j2\pi(\frac{pr}{M} + \frac{qs}{N})} \quad (1)$$

The face images are transformed in the Fourier domain and filtered by 48 Gabor filters (GFs) corresponding to 6 spatial frequencies, with one octave between the focuses of two continuous spatial frequency channels that are $f_i = 5.41; 10.77; 21.60; 43.20; 86.40; 172.8$ cycles per face image and eight exclusive orientations that are $\theta = 0, \frac{\pi}{8}, \frac{2\pi}{8}, \frac{3\pi}{8}, \frac{4\pi}{8}, \frac{5\pi}{8}, \frac{6\pi}{8}, \frac{7\pi}{8}$ in radians.

GF can effectively express the characteristics of the texture. It captures the most exceptional visual properties and has very positive results in facial recognition. GF cores that contain the real part and the imaginary part. GF kernels are similar to the profiles

of the receptive field in simple cortical cells, characterized by localization, selective orientation and frequency selectivity. An image is processed by the kernel element and, then, to produce its corresponding frequency images, which are further employed to compute to obtain Gabor features for the image.

Different experiments have demonstrated that the use of GFs impacts in a pinnacle estimation of the responsive fields of the primary cells of the imperative visible cortex (Hubel and Wiesel, 1968)), given that the applied math analysis of the residual error between the distinction within the response profiles of V1 easy cells and Gabor filters aren't distinguishable from probability (Jones and Palmer, 1987).

The face images transferred in the Fourier space to boost the speed and ease the mathematical processes and GFs were applied to every thumbnail by means that of multiplication within the spectral domain (which is resembling a convolution of the Gabor receptive fields within the spatial domain) is:

$$G(p, q) = \exp\left[-\left(\frac{(u_\theta - f_i)^2}{2\sigma_v^2} + \frac{v_\theta^2}{2\sigma_u^2}\right)\right] \quad (2)$$

where $p_\theta = p \cos \theta + q \sin \theta$ and $q_\theta = q \cos \theta - p \sin \theta$. σ_u and σ_v are standard deviations (SD's) of the Gaussian enfold in the p_θ and q_θ (for example orthogonal to θ). The yields of Gabor channels were the provincial vitality spectra that are multiplied by the kernel of the GF. The GF were applied to the images acquired from the Fourier domain. So now we getting 48 images of each given image from the 6 spatial frequency and 8 Gabor channels.

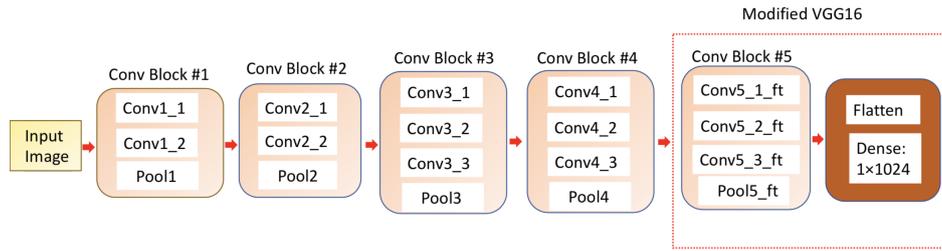


Figure 2: Framework for the modified VGG16.ft network, used for extraction of the expression features from the given facial images

2.1.2 Data Augmentation

CNN needs massive data so to have the option, to sum up to a given issue. However, publically available FER databases do not have sufficient images to handle the problem. Simard et al. (Simard et al., 2003) suggested data augmentation (DA) procedure extend the databases through the creation of synthetic face images for every original face image. Inspired by this procedure, the following activities had been utilized as the data augmentation: 1) flipping image vertically and horizontally 2) Rotate each database image, rotate it at right angles if image is square and rotate it as 180^0 if image is rectangular 3) Add the random noise to the landmarks so as to introduce little deformations to faces.

2.2 Feature Extraction From Given Facial Images

Our proposed framework utilizes DNN (deep neural network) for feature extraction for FER is relies on VGG network of Simonyan and Zisserman (Simonyan and Zisserman, 2014). They come up with two versions of VGG: VGG-16 and VGG-19 (i.e. sixteen and nineteen layers, respectively). VGG16 is chosen due to the fact of its effective performance in visible detection and speedy convergence. It's concerning 138 million parameters and contains 13 convolutional layers, followed by 3 fully-connected layers (FCs). The initial two fully connected layers (FCs) have 4,096 outputs and the last layer has 2,622 outputs. Since the VGG framework not designed for the FER tasks so we modified the framework according to our requirements. Fig. 2 demonstrates the essential module of the framework. Compared with the original VGG16, our VGG16.ft (where "ft" means fine-tuning) is simplified by doing away with two dense layers.

The dimension of the input data for forehead is 54×48 , for eyes is 39×117 , for nose is 50×55 and for lips is 48×74 . At that point, we fix the struc-

Table 1: Parameters set for fifth block.

	conv5_1_ft	conv5_2_ft	conv5_3_ft	Maxpool5_ft
Filters	512	512	1024	
size	7×7	5×5	3×3	2×2
stride	1	1	1	2
pad	3	0	0	0

tures of the initial four conv (convolution) blocks of the VGG16.ft. But we change the structure of fifth conv block of VGG16.ft and also change the names of each layer just by adding "ft" at the end of the original layer name. So now layer name of fifth conv block is like conv5_1_ft. The parameters whose change the structure of the layer is shown in Table. ???. Based on experiments last dense layer preserved and set its dimension to 1×1024 . That dimension is actually the extracted feature of input image denoted as feature vector "fv_1" for the forehead, "fv_2" for the eyes, "fv_3" for the nose and "fv_4" for the lips. We decline the learning rates of layers that have a place with the fifth conv block by 10 times (learning rate for fifth conv block is .001) of other block learning rate (.01 used for other conv blocks) to ensure that that they'll learn more positive information. At last, the initial portion of the system is initialized with the VGG16 model weights which are trained on the Imagenet dataset. ReLu (Rectified Linear Unit) is applied after every convolutional layer.

2.3 Concatenation of Different Outputs and Classification

Fig. 1 shows our proposed framework. Expression features fv is the concatenation of the feature vector came from forehead (fv_1), eyes (fv_2), nose (fv_3) and lips (fv_4). After getting the feature vector next step to do the classification. In the classification process, the similarity between extracted features of the display set and the probe set is evaluated by the SVM and 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. Which are defined as in Eq. 3, 4 and 5

$$d(x_1, y_1) = \sqrt{\sum_{i=0}^n (x_{1i} - y_{1i})^2} \quad (3)$$

$$\chi^2 = \sum_{i,j} \frac{(x_{1i,j} - y_{1i,j})^2}{(x_{1i,j} + y_{1i,j})} \quad (4)$$

$$D_{HI}(x_1, y_1) = - \sum_{i,j} \min(x_{1i,j}, y_{1i,j}) \quad (5)$$

For the computation loss, we used the MSE (mean square error) till now it is best for the SVM and KNN classification, Which is defined as

$$Loss = \frac{1}{N} \sum_1^N \| O_i - O'_i \|^2 \quad (6)$$

Where N is the total numbers of input images, Y and Y' the true and predicted outputs, respectively.

3 EXPERIMENTAL RESULTS AND ANALYSIS

To approve the hypothetical conclusion of the proposed framework, experiments were performed on the four facial datasets. 1) JAFFE database having 213 facial images of 10 Japanese female models of 7 facial expressions (6 basic facial expressions + 1 neutral) All the face images are of size 256 × 256 which are cut as discussed in four regions. Out of 213 images, random 140 images were chosen for the training and the remaining 73 were used for testing.

2) The Video database has videos of 11 persons. Each video contains four different expressions: Normal, Smiling, Angry, and Open mouth. Out of 6668 images, randomly 70% images were chosen for training and remaining 30% images used as testing.

3) In CK+ there are 593 sequences across 123 persons giving 8 facial expressions. This paper uses image sequences of 99 subjects with 7 facial expressions. The face images of CK+ are cut into four informative regions.

4) Oulu-Casia has 6 facial expressions (anger, happiness, surprise, fear, disgust and sad) form 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. The face images of Oulu-Casia are cut into four informative regions.

The convergences of the proposed methodology are assessed in four benchmark datasets, and the outcomes are delineated in Figs. 3, 4, 5 and 6. Each sub-figure demonstrates the trends of accuracy and loss with the rise in iterations. Table 2 shows the Comparison between the Holistic and Modular approach in our proposed framework in the light of SVM and KNN as the classifier for all datasets.

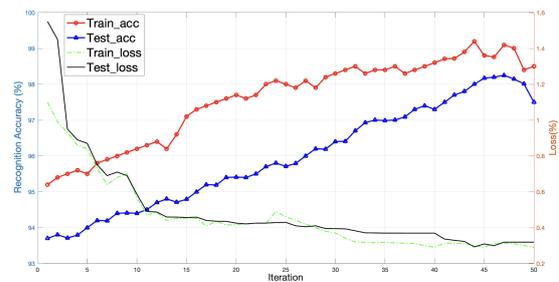


Figure 3: Curves of Accuracy and Loss during training and testing phases for JAFFE dataset.

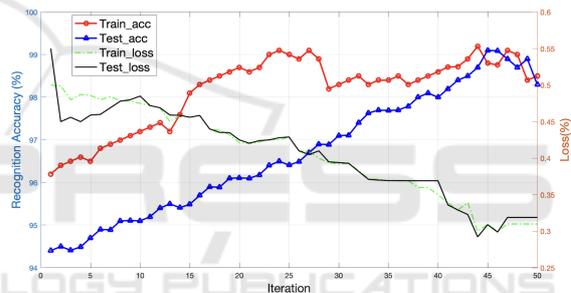


Figure 4: Curves of Accuracy and Loss during training and testing phases for VIDEO dataset.

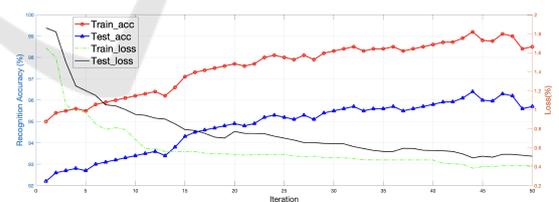


Figure 5: Curves of Accuracy and Loss during training and testing phases for CK+ dataset.

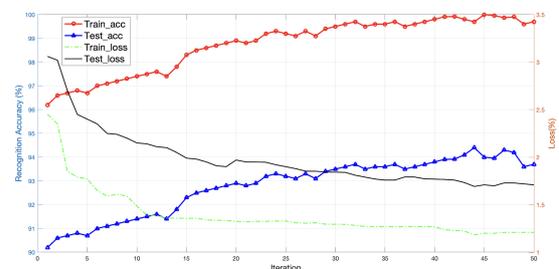


Figure 6: Curves of Accuracy and Loss during training and testing phases for Oulu-Casia dataset.

Table 2: 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			SVM	Modular		
		KNN				KNN		
		Euclidean	Chi Square	Histogram Intersection		Euclidean	Chi Square	Histogram Intersection
JAFFE	93.02	90.32	82.42	80.02	95.87	93.42	90.32	87.27
VIDEO	92.47	88.23	80.98	78.02	96.67	92.50	89.41	85.49
CK+	91.45	88.71	86.41	83.54	96.78	91.24	87.79	89.64
OULU-CASIA	91.40	87.30	79.89	76.20	96.08	89.56	85.64	83.20

Table 3: Comparison with Recognition Accuracy reported in some State-of-the-Art facial expression methods.

DataBase	Methods	Network Type	Additional Classifiers	Accuracy
CK+	Ouellet (Ouellet, 2014)	CNN (AlexNet)	SVM	94.40%
	Li et al. (Li and Lam, 2015)	RBM	-	95.04 %
	Liu et al. (Liu et al., 2014)	DBN	Adaboost	96.07%
	Liu et al. (Liu et al., 2013)	CNN, RBM	SVM	92.05%
	Khorrani et al. (Khorrani et al., 2015)	zero-bias CNN	-	95.01%
	Ding et al. (Ding et al., 2017)	CNN with fine-tune	-	96.08%
	Zeng et al. (Zeng et al., 2018)	DAE	-	93.78%
	Cai et al. (Cai et al., 2018)	CNN+loss layer	-	90.66%
	Liu et al. (Liu et al., 2017)	CNN+loss layer	-	96.10%
	Yang et al. (Yang et al., 2018)	GAN	-	96.00%
	Ours	DNNFG	-	96.78%
JAFFE	Hamester et al. (Hamester et al., 2015)	CNN, CAE	-	95.8%
	shan et al. (Shan et al., 2017)	CNN	-	76.74 %
	Liu et al. (Liu et al., 2014)	DBN	Adaboost	91.8%
	Yang et al. (Yang et al., 2018)	WMDNN	-	92.89%
	Su et al. (Su et al., 2017)	zero-bias CNN	-	95.01%
	Ours	DNNFG	SVM	95.87%
OULU-CASIA	Yang et al. (Yang et al., 2018)	WMDNN	-	92.89%
	Salman et al. (Salmam et al., 2018)	FDP	NN	84.70 %
	Aly et. al. (Aly et al., 2016)	HOG	DKDA	84.21%
	Lopes et. al. (Lopes et al., 2017)	CNN	-	96.42%
		Ours	DNNFG	SVM

To assess the qualitative execution of the proposed framework, facial images are gathered from the Internet for evaluation. Fig.7 interpret the successful expression recognition, Whereas Fig. 8 failed recognition of expression. Table 3 compares recognition results of the proposed technique with that of few State-of-the-art Neural Network-based facial expression recognition techniques.

4 CONCLUSION

This study investigates the FER technique primarily based on the architecture that processes facial regions of the given grayscale facial image and captures the

local information of the face. VGG16.ft has automatically extracted the features from the given facial regions and concatenated the features from all the regions of the face.

Fine-tuning utilized to train the system with the original parameters achieved from the Imagenet dataset.

Furthermore, classifiers like SVM and KNN (with different distance measures) are used to classify the concatenated features. The proposed technique to recognize an individual expression using partial facts from the whole face image is explored during this work. The proposed method is applied to 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 expres-

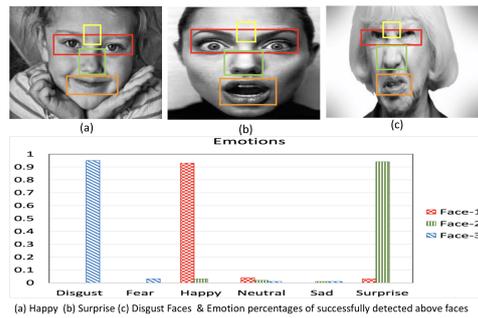


Figure 7: Successful recognition of face images taken from the Internet.

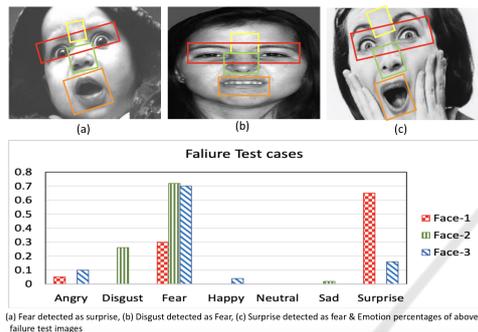


Figure 8: Unsuccessful recognition of face images taken from the Internet.

sions of different persons or the same persons in most of the cases. The evaluation was done on four datasets (JAFFE, VIDEO, CK+ and Oulu-casia) to prove the effectiveness of our framework by recognizing the basic expressions.

In the future, we will focus on simplifying the network used to boost up the algorithm. Furthermore, we intend to add channels of facial images that can be utilized to improve the framework.

REFERENCES

Aly, S., Abbott, A. L., and Torki, M. (2016). A multi-modal feature fusion framework for kinect-based facial expression recognition using dual kernel discriminant analysis (dkda). In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE.

Cai, J., Meng, Z., Khan, A. S., Li, Z., O’Reilly, J., and Tong, Y. (2018). Island loss for learning discriminative features in facial expression recognition. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 302–309. IEEE.

Dailey, M. N., Cottrell, G. W., Padgett, C., and Adolphs, R. (2002). Empath: A neural network that categorizes

facial expressions. *Journal of cognitive neuroscience*, 14(8):1158–1173.

Devries, T., Biswaranjan, K., and Taylor, G. W. (2014). Multi-task learning of facial landmarks and expression. In *2014 Canadian Conference on Computer and Robot Vision (CRV)*, pages 98–103. IEEE.

Ding, H., Zhou, S. K., and Chellappa, R. (2017). Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 118–126. IEEE.

Ekman, P. and Friesen, W. V. (1976). Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75.

Ekman, P. and Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Hamester, D., Barros, P., and Wermter, S. (2015). Face expression recognition with a 2-channel convolutional neural network. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–8. IEEE.

Hu, P., Cai, D., Wang, S., Yao, A., and Chen, Y. (2017). Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 553–560. ACM.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243.

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1233–1258.

Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R. C., et al. (2013). Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM.

Kaneko, T., Hiramatsu, K., and Kashino, K. (2016). Adaptive visual feedback generation for facial expression improvement with multi-task deep neural networks. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 327–331. ACM.

Khorrani, P., Paine, T., and Huang, T. (2015). Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Lai, Y.-H. and Lai, S.-H. (2018). Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. In *Automatic Face & Gesture Recognition (FG 2018), 2018*

- 13th IEEE International Conference on, pages 263–270. IEEE.
- Li, J. and Lam, E. Y. (2015). Facial expression recognition using deep neural networks. In *Imaging Systems and Techniques (IST), 2015 IEEE International Conference on*, pages 1–6. IEEE.
- Liu, M., Li, S., Shan, S., and Chen, X. (2013). Au-aware deep networks for facial expression recognition. In *FG*, pages 1–6.
- Liu, P., Han, S., Meng, Z., and Tong, Y. (2014). Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812.
- Liu, X., Kumar, B. V., You, J., and Jia, P. (2017). Adaptive deep metric learning for identity-aware facial expression recognition. In *CVPR Workshops*, pages 522–531.
- Lopes, A. T., de Aguiar, E., De Souza, A. F., and Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, 61:610–628.
- Lucy, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohnkanade 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.
- Ouellet, S. (2014). Real-time emotion recognition for gaming using deep convolutional network features. *arXiv preprint arXiv:1408.3750*.
- Pons, G. and Masip, D. (2018). Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. *arXiv preprint arXiv:1802.06664*.
- Reed, S., Sohn, K., Zhang, Y., and Lee, H. (2014). Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pages 1431–1439.
- Rifai, S., Bengio, Y., Courville, A., Vincent, P., and Mirza, M. (2012). Disentangling factors of variation for facial expression recognition. In *Computer Vision–ECCV 2012*, pages 808–822. Springer.
- Salmam, F. Z., Madani, A., and Kissi, M. (2018). Emotion recognition from facial expression based on fiducial points detection and using neural network. *International Journal of Electrical and Computer Engineering*, 8(1):52.
- Shan, K., Guo, J., You, W., Lu, D., and Bie, R. (2017). Automatic facial expression recognition based on a deep convolutional-neural-network structure. In *Software Engineering Research, Management and Applications (SERA), 2017 IEEE 15th International Conference on*, pages 123–128. IEEE.
- Shang, W., Sohn, K., Almeida, D., and Lee, H. (2016). Understanding and improving convolutional neural networks via concatenated rectified linear units. In *International Conference on Machine Learning*, pages 2217–2225.
- 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*.
- Su, W., Chen, L., Wu, M., Zhou, M., Liu, Z., and Cao, W. (2017). Nesterov accelerated gradient descent-based convolution neural network with dropout for facial expression recognition. In *Control Conference (ASCC), 2017 11th Asian*, pages 1063–1068. IEEE.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.
- Taheri, S., Qiu, Q., and Chellappa, R. (2014). Structure-preserving sparse decomposition for facial expression analysis. *IEEE Transactions on Image Processing*, 23(8):3590–3603.
- Yang, H., Ciftci, U., and Yin, L. (2018). Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2168–2177.
- Yao, A., Cai, D., Hu, P., Wang, S., Sha, L., and Chen, Y. (2016). Holonet: towards robust emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 472–478. ACM.
- Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., and Dobaie, A. M. (2018). Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, 273:643–649.
- Zhang, F., Zhang, T., Mao, Q., and Xu, C. (2018). Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3359–3368.
- Zhao, G., Huang, X., Taini, M., Li, S. Z., and Pietikäinen, M. (2011). Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619.
- Zhao, X., Shi, X., and Zhang, S. (2015). Facial expression recognition via deep learning. *IETE technical review*, 32(5):347–355.