Fusion of Deep-CNN and Texture Features for Emotion Recognition using Support Vector Machines

J. Sujanaa
Research Scholar, Dept. of Computer Science and Engineering Annamalai University, India

S. Palanivel
Professor, Dept. of Computer Science and Engineering Annamalai University, India

***Abstract* - Emotion recognition has always been growing attention since identifying the internal notion or ideas of human remains critical due to its complex patterns. The fundamental idea is to identify the facial expression by extracting more relevant features that represent deeper insights into the input patterns. In this paper, a fusion technique for emotion recognition is made by employing the deep-learning-based classifier is proposed. The various texture descriptors like Histogram of oriented gradients (HOG), Local binary patterns (LBP), and VGG19 pre-trained model deep features are extracted and fused to form a combine feature vector. The feature selection is done using the principal component analysis (PCA). The support vector machine (SVM) classifier is integrated to perform the classification task. The CK+ facial expression dataset is used where the proposed method gave an accuracy of 84.73%.**

Keywords* - *Feature extraction, Emotion recognition, Histogram of Oriented Gradients, Local Binary Pattern, VGG19.

I. INTRODUCTION

Emotion recognition is a growing field since it has many interesting applications like feedback-based e-learning systems, mood-based music recommendation systems and much more. The study of emotion dates back from the Darwin theory of species evolution[1]. The emotions are broadly classified into six major categories such as happy, sad, surprised, fear, anger and disgust. The inner feeling of a human can be easily predicted through emotions[2].

Facial emotions can contribute more important features since identifying the real inner feeling during human communications. These emotions fluctuate from one person to another person due to the complex nature of facial muscle structure. Facial features include eyebrows, eyes, nose, mouth, cheek and jawline. During the communication, eyebrows will lift for surprising emotion, mouth region will open in wide; jawlines are uplifted, cheeks are uplifted during happy emotion and so on. Emotion recognition can be identified through verbal and non-verbal methods. Speech-based emotions is a form of verbal emotion where the audio signals like electroencephalography (EEG)[3], [4] is utilized to extract the features and identify the emotions in clinical studies. The facial emotions are the form of non-verbal communication method which paved way for many interesting applications. In recent studies, facial expressions gave a very important visual representation of the human mind. These visual representation-based study contributed remarkable achievements in the deep learning-based application.

The deep learning is an advanced method of machine learning, which uses samples in large scale for its powerful feature learning. The computer vision study uses images or videos for analysis like real-time object detection, real-time monitoring, hand-written recognition, image classification, etc., Over the decades, deep convolutional neural networks have shown improved performance in emotion recognition. These deep features represent the emotions in the most possible way. The transfer learning methods help in extracting features from the image net classification models such as Inception, Xception[5], GoogleNet[6], ResNet[7], VGG16, VGG19[8], etc. These complex features can be utilized to detect and recognize emotions.

II. RELATED WORKS

There exist a few investigations dependent on the FER approaches. This zone is as yet an arising field to the high-level capacities of Artificial Intelligence (AI) advances and instinct to distinguish the human thoughts with these mind-boggling machine learning capabilities. Several studies focus on developing an emotion recognition application using texture-based features such as histogram of oriented gradients hog, local binary pattern LBP, SURF, SIFT, HAAR[9], Gabor[10], etc. These textures are called as patterns or templates that are matched with the regions in the entire image. These patterns are useful for extracting the most relevant information from the image area. The HOG and LBP features are fused using a genetic programming-based method and this fusion function is more robust due to rotation and illumination invariant[11]. The authors[12] identified that HOG and LBP features can contribute the best accuracy to detect humans with the method of cross over and mutation functions. The authors identified that the fusion of CNN, HOG and LBP features in spoof detection[13]. The fusion of HOG and LBP features is also used in gesture recognition,

pedestrian detection, facial expression detection etc. The authors also used to the fusion of HOG and LBP method to identify the facial expressions here the JAFFE dataset is used and 68 landmark points or extracted for fog feature extraction the facial pages like eyes nose and mouth regions or extract and individually for HOG and LBP feature extraction then the PCA is used for dimensional reduction in the Matlab framework produced an accuracy of 96.20%[14]. From the literature review, it is evident that the fusion of features is more robust in predicting emotions.

III. PROPOSED METHODOLOGY

The objective of this work is to develop an emotion recognition system using the fusion of deep features, HOG and LBP texture features. The proposed system uses the transfer learning method to extract the deep features from the input images. The system detects seven facial expressions in the CK+ dataset. The main steps involved in this work is as follows: Dataset collection, Feature extraction, Feature fusion, Dimensionality reduction, Training and Testing the SVM models. The block diagram of the proposed system is given in Fig. 1.

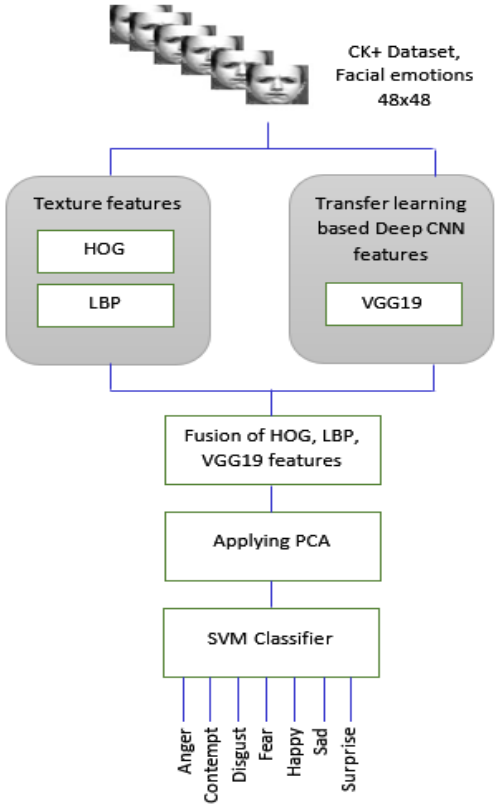

Fig. 1. Block of the proposed emotion recognition system.

The input to the system is a 48x48 facial grayscale image. At first, HOG and LBP texture features are extracted. Transfer learning is used to extract deep VGG19 features. All these features are combined to form as the final feature vector. This array is fed as input for training and testing the SVM classifier. Also, the deep feature vector is reduced using the PCA method.

A. Feature extraction

Feature extraction is a process of eliminating the redundant and unwanted piece of information and grouping useful information into a feature vector. The various feature descriptors such as HOG[15], LBP[16], [17], SURF[18], and SIFT[19] are used in classification tasks but HOG and LBP are utilized because they can capture efficient features than the other descriptors and 'OpenCV-Contrib-python' package is used to implement the same [20].

B. Histogram of Oriented Gradients (HOG)

The HOG is initially proposed by Dalal and Briggs in the year 2005 for human detection. The special characteristics of HOG help in identifying objects and other computer vision applications. The following are the steps to compute the HOG features. The entire image region is divided into cells and blocks where each cell has n x n pixels and each block has m x m cells is shown in Fig. 2.

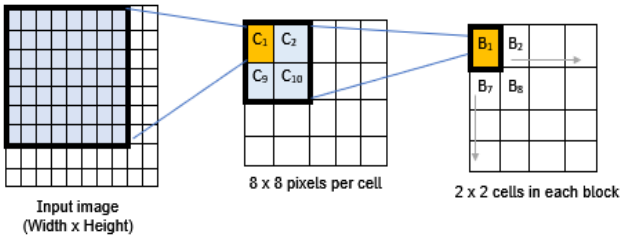

Fig. 2. Cells and Blocks in HOG.

The 'x' and 'y' gradient for the image is computed by applying a derivative mask (Sobel filter) along the row-wise and column-wise over the input images.

$$g(x) = [-1\ 0\ 1] \tag{1}$$

$$g(y) = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \tag{2}$$

The gradient images contain the regions of important highlights such as edges and corners in the images. The histogram is computed for every block. After computing the gradients, the total magnitude and orientation are computed using (3) and (4).

$$mag = \sqrt{\left[(g(x))^2 + (g(y))^2\right]} \tag{3}$$

$$\tan(\emptyset) = g(x)/g(y) \tag{4}$$

The histogram bins are equally divided between the range 0 to 180 degree for singed gradient and between 0 to 360 degree for unsigned gradient bins. Every pixel in the image array will vote a histogram bin based on the magnitude value. Since the histogram should be invariant to illumination and lightening changes, an L2-block normalization is performed to remove the scale. The final histogram has the n-

dimensional HOG feature array of size no. of samples x no. of features.

C. Local Binary Pattern (LBP)

The LBP was proposed by Ojala et. al[21] for identifying the robust features which are invariant for illumination, grayscale changes in every 3x3 pixels. The LBP is used in several applications due to its faster and reliable computation. The following are the steps to compute the LBP features. The entire image array is divided into rows and columns. The radius of size 'r' and 'p' points are chosen. The 3x3 neighbourhood with varying r and p is shown in Fig. 3.

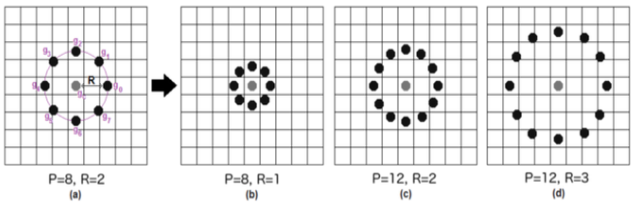

Fig. 3. The radius for the computation of LBP descriptor at different scales, [22].

Basically, for a 3x3 neighbourhood, the centre pixel acts as a threshold value. The surrounding 8 pixels in the 3x3 array value is updated based on the centre pixel. If the surrounding pixel value is greater than the centre pixel, binary '1' is assigned otherwise '0' is assigned to the pixel. From the top-most corner in the clockwise direction, is taken for forming a binary value. The decimal equivalent of the binary value is replaced in the pixel. For extracting the uniform binary patterns, the utmost 2 transitions are used. E.g., 0-1 transition and 1-0 transition, in the binary number 00000010 has two transitions whereas in 00101010 has 6 transitions which are said to be a uniform pattern. Similarly, the 3x3 neighbourhood pixel move throughout the image and the normalized histogram is concatenated to form the LBP feature array.

D. Transfer Learning VGG_19 features

Transfer learning is the process of imparting knowledge from one domain to the other domain. Through this kind of learning[23], the time consumed to design and train a new architecture from the scratch is reduced. The VGG19 architecture was the runner up ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[8] conducted in the year 2014. This architecture is a 19-layer deep, consisting of different blocks. Each block has multiple convolution layers followed by a max-pooling layer. The input for the VGG19 is a 224x224 RGB image. The first two blocks have two convolutional layers, block 3, block 4 and block 5 has three convolutional layers, having max-pooling layer at the end of each block. The flattening layer has 25088 features. The Fully Connected layers (FC1, FC2) has 4096 neurons. The last layer is the prediction layer which has 1000 neurons to classify 1000 objects in the ImageNet database. The VGG19 architecture is shown in Fig. 4.

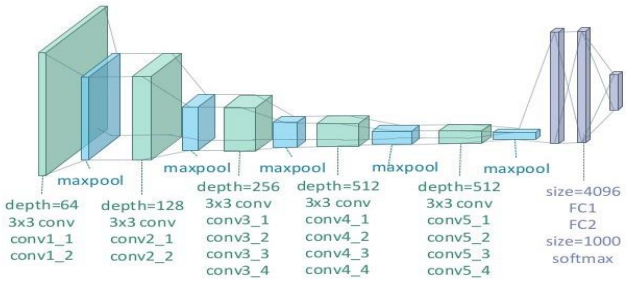

Fig. 4. The architecture of VGG19, [24].

In transfer learning, the top is set to FALSE i.e., the fully connected layers are removed. The block 5 max-pooling layer has 512 features which are extracted for our input emotion images.

E. Principal Component Analysis

The Principal Component Analysis (PCA) is a unique technique to reduce the dimension of the feature array and to increase its capacity by minimizing the loss of information from the data[25]–[27]. The large dataset is converted into a smaller one containing only the accurate piece of information. This technique helps the machine learning algorithms to predict the output much easier without the need to compute unnecessary variables. In PCA, the standardisation is quite critical such that the variance of the initial variables is transformed into smaller ranges. The following are the steps to compute the PCA. The mean is computed for the data and means subtracted data is calculated. The covariance matrix is computed for the mean subtracted data to identify the relationship between them. Highly correlated data indicates that there may be chances for redundant information in the data. The covariance matrix is a symmetric matrix where the main diagonal consists of the variance of the initial variables. The eigenvector and eigenvalue are computed from the covariance matrix. This eigenvector and eigenvalue determine the principal components. The first principal component contains the foremost information about the initial variables, the PC2 contains the next set of important information of the input variables, and so on. The principal component represents the maximum variance such that the direction of the data, has the most important information about the data.

F. Support Vector Machine

It is a standard classifier algorithm for classifying binary or multiclass problems. The set of hyperplanes that divides the input emotions into different categories is drawn. The optimal hyperplanes form the list of hyperplanes is chosen in such a way it has the maximum margin between the data points is shown in Fig. 5.

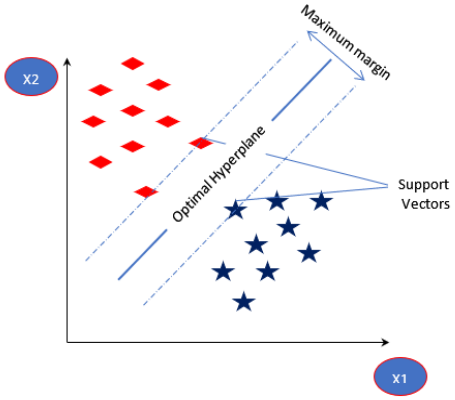

Fig. 5. SVM optimal hyperplane.

The SVM classification is used when the training input and their corresponding labels are well-known. SVM can be applied for classification and regression problems. The learning of the hyperplane is either linear, sigmoid, or polynomial kernals.

IV. EXPERIMENTAL RESULTS

A. Dataset

The dataset is collected from the CK+ database where it consists of 593 sequences with 123 subjects ranging from neutral expressions to peak expressions. In our work, seven facial expressions namely anger, contempt, fear, disgust, happy, sad and surprised are taken. The CK+ dataset containing emotions is divided into training and test sets. The samples of the dataset are used for the proposed emotion recognition system is shown in Table 1. The dataset is divided into 784 training samples and 197 testing samples.

TABLE I. NO. OF EMOTION IMAGES TAKEN FOR THE PROPOSED EMOTION RECOGNITION SYSTEM IN THE CK+ DATASET.

Emotions	No. of samples
Anger	135
Contempt	54
Disgust	177
Fear	75
Happy	207
Sad	84
Surprised	249

B. Feature fusion

The feature fusion is a worthy technique since different technique captures different kinds of important highlights from the image regions. In this fusion technique, the hand-crafted features HOG and LBP are fused along with the transfer learning-based VGG19 model features. These ImageNet models can capture effective features with a diverse range of filters and kernels. So, these features are fused along with texture features to identify the emotions. To extract the HOG features, the cell size of 8x8 pixels and block size of 2x2 is used. The 9-bin orientation with L2-block normalization is used and 900 HOG features are extracted of the shape, (981,900). To extract the uniform LBP patterns, the radius of size 2 and 8 points is taken. The 59 uniform LBP features are extracted for the emotion images in the shape (981,59). Using the transfer learning method, in VGG19 model, the 'top' is removed from the model where the block 5 max-pooling layer gives 512 deep features, which are extracted. The shape of the feature array is (981,512). All these features are fused into a single feature vector of size (981,1471) i.e., no. of samples x no. of total features. This vector is fed for dimensionality reduction using PCA.

C. Training SVM

The features from the HOG, LBP and VGG19 which are fused is fed as input to the SVM model to detect the emotion where the model captures the internal structure. The multi-SVM used for classification learns from the internal data. The SVM model gave an accuracy of 71.06%. After applying PCA, the results of the SVM classifier are shown in Table 2.

TABLE 2. ACCURACY FOR THE SVM MODEL AFTER APPLYING PCA.

'n' Principal Components	Training time (mm: ss)	Recall (R)	Precision (P)	Accuracy (A)
1	0.0230	0.5431	0.5067	0.5431
2	0.0221	0.6091	0.5459	0.6091
5	0.0204	0.7614	0.7656	0.7614
7	0.2429	0.8477	0.8473	0.8479
10	0.0223	0.8345	0.8274	0.8274

The model gave better results when applying after applying PCA to the training data, where 7 principal components gave an accuracy of 84.79% beyond which the accuracy deteriorates.

D. Testing SVM

Testing is a crucial part of identifying the real performance of any classification algorithm. The trained models are loaded using Keras[28] 'load model' function and the test samples are predicted using the loaded model. The SVM model with 1471 features and 7 principal components gave an accuracy of 84.79%.

E. Performance Evaluation

The performance measures like Precision (P), Recall (R), and F1-score (F) are the most commonly used metrics to study the performance of the machine learning classifiers. In the confusion matrix, the True Positive (TP) is the total number of emotions correctly identified as the labelled emotion. The False Positive (FP) is the number of emotions mistakenly classified as the labelled emotion. True Negative (TN) is the samples correctly identified as labelled emotion in other

classes. Whereas, False Negative (FN) is the incorrectly classified emotion to other classes. The accuracy(A) is defined as the total number of corrected identified emotions to the total number of samples and it is given by,

$$A = \frac{[(TP+TN)]}{[(TP+TN+FP+FN)]} \quad (5)$$

The precision is the ratio of TP with TP and FP and it is calculated as,

$$P = \frac{TP}{TP+FP} \quad (6)$$

The recall is the ratio of TP with TP and FN and it is calculated as,

$$R = \frac{TP}{TP+FN} \quad (7)$$

The precision and recall are combined to form the measure called F1-score, which is the harmonic mean of P and R.

$$F = 2 * \left[\frac{precision * recall}{precision + recall}\right] \quad (8)$$

The precision, recall and f1-score for the SVM model are shown in Fig. 6 for n=7 when applying PCA.

CLASSIFICATON REPORT:

	precision	recall	f1-score	support
Anger	0.70	0.84	0.76	25
Contempt	0.50	0.55	0.52	11
Disgust	0.83	0.77	0.80	44
Fear	0.60	0.30	0.40	10
Happy	0.93	1.00	0.96	37
Sadness	1.00	0.56	0.71	9
Surprise	0.94	0.98	0.96	61
accuracy			0.84	197
macro avg	0.78	0.71	0.73	197
weighted avg	0.84	0.84	0.84	197

Fig. 6. Performance evaluation of emotion recognition using the proposed method for CK+ dataset.

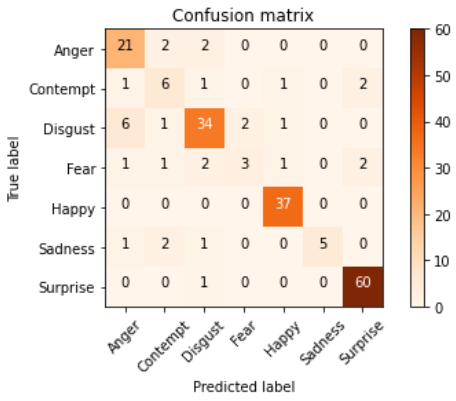

Fig. 7. Confusion matrix of emotion recognition using the proposed method for CK+ dataset.

The confusion matrix is used to analyse the classifier performance with true positive (TP), true negative (TN), false positive (FP) and false-negative (FN) values. In Fig. 7, the 21 samples are correctly classified as anger emotion and 4 samples are misclassified to other classes. Similarly, for Surprise class, 60 emotions are correctly classified as surprise emotion and 1 sample is incorrectly classified as disgust emotion. The system, thus correctly classifies 166 samples and incorrectly classifies 31 samples.

V. CONCLUSION

We present an efficient feature fusion methodology to recognize the emotion. Our method applies the PCA for dimensionality reduction. With this implementation, the proposed method recognizes the seven facial expressions with an accuracy of 84.79%. Our system fails to recognize the emotion when some person shows the wrong expressions or when they cannot fully express their feelings. It will increase the false-positive results. In future, we can improve the recognition rate by dividing the entire face image into various salient patches such as eyes, nose, and mouth regions and applying PCA on these individual patches.

CONFLICT OF INTEREST

The authors acknowledge that there is no conflict of interest.

ACKNOWLEDGMENT

The authors are very grateful to Annamalai University, Annamalainagar, for offering a suitable and appropriate environment with well-equipped Vision and speech laboratory in the Department of Computer Science and Engineering to elevate the research projects effectively.

REFERENCES

[1] M. Ghiselin, P. Ekman, and H. Gruber, “Darwin and Facial Expression: A Century of Research in Review.,” *Syst. Zool.*, 1974, doi: 10.2307/2412481.

[2] Y. Li, S. Wang, Y. Zhao, and Q. Ji, “Simultaneous facial feature tracking and facial expression recognition,” *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2559–2573, 2013, doi: 10.1109/TIP.2013.2253477.

[3] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, “Spatial-Temporal Recurrent Neural Network for Emotion Recognition,” *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 939–947, 2019, doi: 10.1109/TCYB.2017.2788081.

[4] D. W. Chen *et al.*, “A feature extraction method based on differential entropy and linear discriminant analysis for emotion recognition,” *Sensors (Switzerland)*, vol. 19, no. 7, pp. 1–17, 2019, doi: 10.3390/s19071631.

[5] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 1800–1807, 2017, doi: 10.1109/CVPR.2017.195.

[6] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07-12-June, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[7] D. Sokolov and M. Patkin, “Real-time emotion recognition on mobile devices,” *Proc. - 13th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2018*, p. 787, 2018, doi: 10.1109/FG.2018.00124.

[8] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” pp. 1–14, 2014.

[9] S. Lu and F. Evans, “Haar Wavelet Transform Based Facial Emotion Recognition,” *Adv. Comput. Sci. Res.*, vol. 61, no. ECMS 2017, pp.

342–346, 2017.
[10] A. B. Andres Hernandez-Matamoros and H. P.-M. Enrique Escamilla-Hernandez, Mariko Nakano-Miyatake, "A Facial Expression Recognition with Automatic Segmentation of Face Regions Andres," . *Commun. Comput. Inf. Sci. Springer*, vol. 532, pp. 529–540, 2015, doi: 10.1007/978-3-319-22689-7.
[11] M. Hazgui, H. Ghazouani, and W. Barhoumi, "Genetic programming-based fusion of HOG and LBP features for fully automated texture classification," *Vis. Comput.*, pp. 1–20, Jan. 2021, doi: 10.1007/s00371-020-02028-8.
[12] P. Verma, "Human Detection using Feature Fusion Set of LBP and HOG," no. September, pp. 261–265, 2017.
[13] G. G. Das, "Fusion of CNN and LBP-HOG features for Face Detection," *Int. J. Res. Sci. Innov.*, vol. 7, no. 6, pp. 58–59, 2020, doi: 10.1109/ACCESS.2018.2812208.
[14] Y. Liu, Y. Li, X. Ma, and R. Song, "Facial expression recognition with fusion features extracted from salient facial areas," *Sensors (Switzerland)*, vol. 17, no. 4, pp. 1–18, 2017, doi: 10.3390/s17040712.
[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, 2005, vol. I, pp. 886–893, doi: 10.1109/CVPR.2005.177.
[16] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face Recognition with Local Binary Patterns," in *European Conference on Computer Vision*, 2004, doi: 10.1007/978-3-540-24670-1_36.
[17] Z. Borui, G. Liu, and G. Xie, "Facial expression recognition using LBP and LPQ based on Gabor wavelet transform," *2016 2nd IEEE Int. Conf. Comput. Commun. ICCC 2016 - Proc.*, pp. 365–369, 2017, doi: 10.1109/CompComm.2016.7924724.
[18] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, 2008, doi: 10.1016/j.cviu.2007.09.014.
[19] A. Majumdar and R. K. Ward, "Discriminative sift features for face recognition," in *Canadian Conference on Electrical and Computer Engineering*, 2009, pp. 27–30, doi: 10.1109/CCECE.2009.5090085.
[20] A. Kaehler and G. Bradski, *Learning OpenCV 3*. 2016.
[21] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: 10.1109/TPAMI.2002.1017623.
[22] ArIES and IIT Roorkee, "Texture analysis using LBP. Texture classification is an important… | by ArIES, IIT Roorkee | Medium," 2018. [Online]. Available: https://medium.com/@ariesiitr/texture-analysis-using-lbp-e61e87a9056d. [Accessed: 15-Jan-2021].
[23] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image Vis. Comput.*, vol. 65, pp. 66–75, 2017, doi: 10.1016/j.imavis.2017.01.012.
[24] Park Chansung, "Transfer Learning in Tensorflow (VGG19 on CIFAR-10): Part 1 | by Park Chansung | Towards Data Science." [Online]. Available: https://towardsdatascience.com/transfer-learning-in-tensorflow-9e4f7eae3bb4. [Accessed: 15-Jan-2021].
[25] D. Dagar, A. Hudait, H. K. Tripathy, and M. N. Das, "Automatic emotion detection model from facial expression," in *Proceedings of 2016 International Conference on Advanced Communication Control and Computing Technologies, ICACCCT 2016*, 2017, pp. 77–85, doi: 10.1109/ICACCCT.2016.7831605.
[26] H. Ali, M. Hariharan, S. Yaacob, and A. H. Adom, "Facial emotion recognition using empirical mode decomposition," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1261–1277, 2015, doi: 10.1016/j.eswa.2014.08.049.
[27] M. Khan, S. Chakraborty, R. Astya, and S. Khepra, "Face Detection and Recognition Using OpenCV," in *Proceedings - 2019 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2019*, 2019, doi: 10.1109/ICCCIS48478.2019.8974493.
[28] "keras." [Online]. Available: https://keras.io/.