Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020)
IEEE Xplore Part Number: CFP20P17-ART; ISBN: 978-1-7281-5821-1

Real-Time Facial Emotion Recognition System With Improved Preprocessing and Feature Extraction

Dr Ansamma John*, Abhishek MC†, Ananthu S Ajayan‡, Sanoop S§ and Vishnu R Kumar¶
Department of Computer Science and Engineering, TKM College of Engineering
APJ Abdul Kalam Technological University
Kerala, India
Email: *ansamma.john@gmail.com, †mcabhishek@yahoo.com, ‡ananthusajayan3747@gmail.com,
§sanoopatholi707@gmail.com, ¶rkumar.vishnu28@gmail.com

***Abstract*—Human emotion recognition plays a vital role in interpersonal communication and human-machine interaction domain. Emotions are expressed through speech, hand gestures and by the movements of other body parts and through facial expression. Facial emotions are one of the most important factors in human communication that help us to understand, what the other person is trying to communicate. People understand only one-third of the message verbally, and two-third of it is through non-verbal means. There are many face emotion recognition (FER) systems present right now, but in real-life scenarios, they do not perform efficiently. Though there are many which claim to be a near-perfect system and to achieve the results in favourable and optimal conditions. The wide variety of expressions shown by people and the diversity in facial features of different people will not aid in the process of coming up with a system that is definite in nature. Hence developing a reliable system without any flaws showed by the existing systems is a challenging task. This paper aims to build an enhanced system that can analyse the exact facial expression of a user at that particular time and generate the corresponding emotion. Datasets like JAFFE and FER2013 were used for performance analysis. Pre-processing methods like facial landmark and HOG were incorporated into a convolutional neural network (CNN), and this has achieved good accuracy when compared with the already existing models.**

***Index Terms*—Convolutional neural network (CNN); emotion recognition; facial expressions; pre-processing; feature extraction; facial emotion**

I. INTRODUCTION

Humans have always been intrigued by the idea of creating a replica of the human mind. Moreover, the most essential characteristic which these systems need is the ability to capture the emotions of a human being and react accordingly to a particular situation. The real goal is to bridge the gap between a machine-like robot and a human being, thereby making it more reliable. The mind controls everything and thus, it plays a massive role in deciding the behavioural patterns of a person and his ability to communicate. The mind translates the thoughts and feelings in a human brain by facial emotion, speech, body movements and many other involuntary actions, of which the first two are considered as the primary aspects in communication. So decoding the emotions of a person is the easiest way by which a machine can interact with a person. The development in this field first began with the basic principles which were related to biological nature and then went onto more sophisticated things like machine learning with the help of modern technology. Though a significant progress has happened in this area, it still lacks the ability to give a definite answer due to the complexity and wide variety of expressions possessed by the people. Most of the new methods and the already existing ones cannot still guarantee that the results are entirely accurate.

Automatic recognition of facial expression plays a vital role in artificial intelligence and robotics, and thus it is a need of the generation. Some application related to this includes Personal identification and Access control, Videophone and Teleconferencing, Forensic application, Human-Computer Interaction, Automated Surveillance, Cosmetology etc. Facial emotion recognition helps to decode a message and add clarity to what a person, is trying to say. It is said that while communicating with a person, we convey two-third of the message through non-verbal components and one-third through verbal components. Emotions play a huge role in these non-verbal components. So by decoding the emotion of a person we can find the essence of a message hidden by facial emotions.

The developments in FER system is now extensively used in the medical sector also. It was used to find people who were genetically vulnerable and had a chance of having bipolar disorder [1]. FER system was used to identify the emotions of people who were suffering from myotonic dystrophy [2], a medical condition in which people gradually lose muscle power. The system was also tested in helping autistic children and people who were in post-traumatic stress order [3]. Nowadays, it is also used in understanding tourist satisfaction after visiting a place. So the application of this technology is limitless and is going to solve many problems we face today.

Facial recognition has been used nowadays from security to identifying a person, but it is still far from achieving the goal of facial emotion recognition. Ekman and Friesen [4] addressed this by adopting the system developed by CarlHerman Hjortsjo called facial action coding system (FACS).¨ FACS is a system based on facial muscle changes and can characterise facial actions to express individual human emotions as defined by Ekman and Friesen in 1978. FACS encodes the movements of specific facial muscles called action units (AUs), which reflect distinct momentary changes in facial appearance. Ekman and Friesen had earlier

defined the six basic emotions as fear, happy, disgust, surprise, sad and anger. This was later on used as the most fundamental emotions to be identified added with the neutral emotion.

Initially, emotions were detected using the semantic and syntactic properties of a language. However, this faced many problems like, messages were misinterpreted and the usage of a language differed among different groups of people. So the original emotions were not always present in a text, and it featured the classic problem in sentiment analysis. When negations, ironies and words with change meaning according to the context of the text were present in a text it made things more difficult. Conventional FER approaches were later used and became very popular. In conventional FER approaches, the face was first detected, then the features from the face were extracted and then were classified with the help of classifiers. This traditional approach then gave way to deep learning-based FER techniques such as CNN.

The proposed method in this paper was able to produce comparable results to the methods which are already present, and variation in accuracy after the inclusion of preprocessing techniques in the CNN model is also being discussed in the paper. The rest of the paper discusses things related to the proposed system as follows. Section II describes the related works, usage of CNN based techniques, and issues in current FER. Section III explains the proposed methodology and the architecture of the system. Section IV discusses the experiments and the inferences from it. Section V presents the conclusion.

II. Related works

Works which are related to the proposed system is discussed in this section. All these methods have chosen different approaches in order to tackle different problems and improve the works already done in the field of emotion recognition. Though this paper is focusing solely on CNN model, different feature extraction and preprocessing methods are included in this section because we are comparing the change in accuracy when these are incorporated with the CNN model.

Ozdemir et al. [5] proposed a method where they combined three datasets (JAFFE, KDEF and their custom dataset) and used Haar cascade library for removing the unimportant pixels which were outside the facial region. This CNN model was able to achieve an accuracy of 91.81% for the classification of seven different emotions. This model was more accurate at predicting surprise, fear and neutral emotion states and was less accurate at predicting the sad emotional state. Jaiswal and Nandi [6] proposed a system which smoothly worked in a real-time speed and had an accuracy of 74% over eight different datasets making it a more robust system. The different variations like occlusion, illumination, age, race and gender differences were tested since 8 datasets were used. They improved the accuracy by varying filter size, network convolution layers and were able to reduce the time complexity of the Vanilla CNN model by half.

Mollahosseini et al. [7] introduced a new deep neural network architecture for emotion recognition, where it had two convolutional layers which were followed by max-pooling and four inception layers. The experiment was done over seven different datasets and achieved better accuracy and training time when compared to traditional CNN models. Bo-Kyeong Kim et al. [8] did six different experiment scenarios in order to compare the performance change in CNN. In this method, the inclusion of alignment mapping networks were found out to have improved the performance of FER when non-alignable faces were detected which is common in real-world scenarios. Pramerdorfer and Kampel [9] proposed a method based on the CNN model and was exclusively tested on the FER2013 dataset. An ensemble of modern deep CNN was used in this method and they achieved an accuracy of 75.2% on FER2013 dataset. They utilisd modern architectures to improve the facial emotion recognition performance and were able to outperform the shallow and basic CNN architectures. Pitaloka et al. [10] proposed a system working on the CNN model, which could classify six emotions and they found out the influence of different preprocessing techniques. They used preprocessing techniques like resizing, face detection, cropping, adding noises, and data normalisation consists of local normalisation, global contrast normalisation and histogram equalisation. From their result, it is said that cropping the face to the region of interest boosted the accuracy which they achieved using the CNN model. Though they were able to improve the accuracy they were not able to detect all the seven emotions. It is understood from many experiments that once the neutral expression was added in the classification, the accuracy reduces.

Lopes et al. [11] proposed an approach based on CNN for emotion recognition and achieved improved results when changes were applied in preprocessing techniques. Their experiment results show that the combination of normalisation technique made significant improvements to the accuracy. This method uses the location of each of the eyes while doing the preprocessing techniques, and it was found out that it could easily be included without affecting the real time nature of the system. The accuracy of the model is said to be 96%, but for certain expressions like sad, they were only able to achieve an accuracy of 84%.

The overall architecture and methodology followed is discussed in the next section.

III. Methodology

This paper proposes a novel method for improving the real time emotion recognition; the proposed method adopts some additional feature extraction methods for increasing the training accuracy along with the CNN model. Figure 1 shows the underlying architecture of the proposed system. It contains four modules. The first module captures real-time video through webcam and detects the face using Local binary patterns (LBP) [12] cascade classifier. The process of

detection was carried out by considering the image as a composition of micro patterns. Next module deals with pre-processing of the image. Cropping, resizing and intensity normalisation are

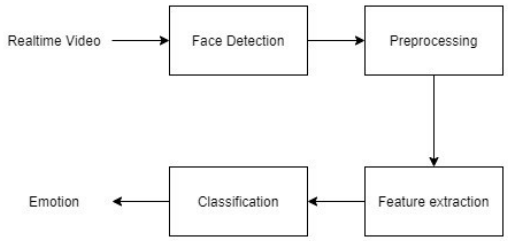

Fig. 1. Basic steps involved in the system

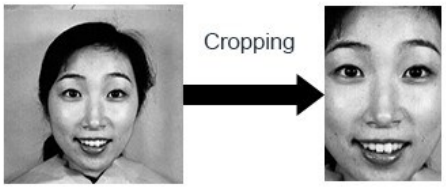

Fig. 2. Cropping

the method incorporated. Feature extraction methods were implemented to extract features from the raw image. Then detection is done based on these features selected.

Pre-processing is the most important step in this system. Next section describes the pre-processing techniques incorporated in the proposed system.

A. Pre-processing

Image captured through the webcam contains portions which were not required for detecting facial expression. For example, parts of the neck, hair etc. are not needed. So these unwanted information were removed. Else the detection method will have to deal with more data and thus making it complicated and less efficient. Pre-processing the raw image includes the removal of this unwanted information from the image. Various steps in pre-processing are cropping, resizing and intensity normalisation. Cropping the raw image ensures that image parts that do not have expression specific information are removed. The facial regions around the mouth, eye are most important for detection of emotion. The cropped image is further resized to ensure the data size of the pixel file to match with the input size of CNN.

The image brightness and contrast vary with illumination and lighting condition of the object. Such variations cause increased complexity of feature sets and the detection method. In order to reduce these issues an intensity normalisation [13] was applied. MinMax normalisation is used in the proposed system, where linear transformation was performed on the original image.

Next section describes the feature extraction methods deployed in the proposed system.

B. Feature Extraction

Properly pre-processed data is given as input into the next module where feature extraction is done. Feature extraction

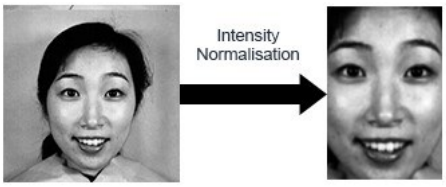

Fig. 3. Intensity normalisation

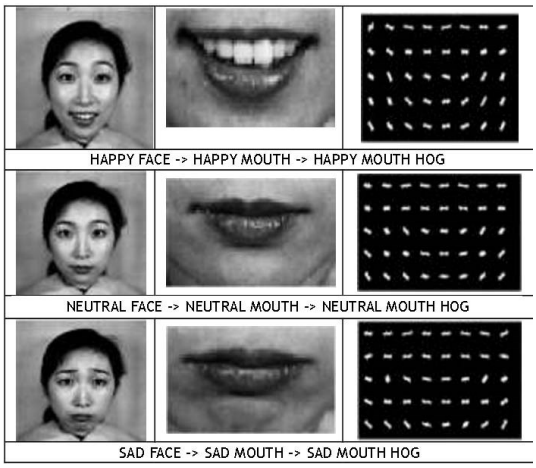

Fig. 4. Output after HOG

selects the relevant information contained in the image so that the detecting the emotion is made easy [14]. This work implements a hybrid method of feature extraction by combining the features extracted by CNN with Histogram oriented gradients (HOG) [15] and facial landmarks [16]. HOG describes local object appearance and shape within an image by the distribution of intensity gradients or edge directions [15]. HOG operates on local cells; thus, it is invariant to geometric transformations. These HOG features vary for each expression, so distinguishing them is more comfortable. This is why we selected HOG as the feature selector for this system. Inorder to extract HOG features, skimage package in python is used. After HOG, next we have to extract the features by facial landmark detection. Facial landmark detection is a method which detects key landmarks on the face. In order to detect facial landmarks, Dlib function from OpenCV is used. This function can take an image region which contains an object as input and then output a set of locations which defines a pose of that object. Using this method, it can identify 68 facial landmarks. As described earlier, only the portions around the mouth and eye were needed for detecting emotion since only the muscles in those areas changes with facial expressions. So the subset of facial landmarks is considered. Output after applying HOG and

facial landmark descriptors are shown in Figure 4 and Figure 5.

This system implements a Convolutional Neural Network (CNN) [17] for classification of emotion using the features

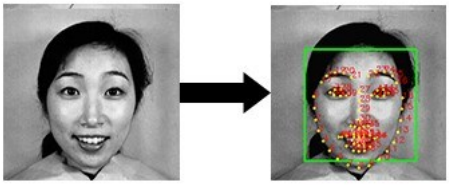

Fig. 5. Output after DLib. Facial landmarks are marked in the image.

extracted using capabilities of CNN and the additional features extracted in the previous module. Proposed CNN has an input layer, four layers of convolution, two layers of pooling, and two fully connected layers for classification. Next section describes the experimental setup and discusses the results obtained.

IV. Experiments and Discussion

All the experiments were conducted using two publicly available databases in facial emotion recognition field: The Japanese Female Facial Expressions (JAFFE) [18] database and FER2013 dataset [19]. JAFFE dataset contains 213 images of 7 facial expressions (angry, disgust, fear, happy, sad, surprise and neutral) posed by 10 Japanese female models. Images are 256x256 grey scaled. These images were resized since our CNN model accepts 48x48 images as input. FER2013 dataset contains 35887 images, size of 48 x 48 pixels of 7 facial expressions. The number of images in each emotion category and sample of images are depicted in Table 1 Figure
2, respectively.

In order to analyse the performance of the system using the datasets, we implemented the system with the help of OpenCV, Keras and Python. Pre-processing is done with OpenCV, and CNN is built using Keras. The architecture of our proposed CNN is shown in Figure 2.

The first layer of convolution use 64 5x5 convolution kernels, while the second convolution layer, the third convolution layer, and the fourth convolution layer uses 128, 512, and 512 3X3 convolution kernels respectively.

Also, in all the layers, Rectified Linear unit (ReLU) is used as the activation function. Max-pooling layer of 2x2 is embedded into each layer after applying ReLU activation function. After four convolution layers, the network is led to two fully connected layers. In the first fully connected layer, we had a hidden layer with 256 neurons while the second fully connected layer is having a hidden layer with 512 neurons. The image matrix, which is obtained as the output of four convolution layers, is converted into a vector and fed into first fully connected layer then output of first fully connected layer is fed into the second fully connected layer. The last layer comprises the sigmoid function. The proposed system combines HOG features and facial landmarks with those features extracted by convolutional layers by utilising the same CNN architecture, but the HOG features and facial landmarks are added to those exiting the last convolutional

TABLE I
NUMBER OF IMAGES IN EACH EMOTION OF THE TWO DATABASES.

Emotion	No of images	
	JAFFE	FER2013
Angry	30	4593
Disgust	29	547
Fear	32	5121
Happy	32	8989
Sad	30	6077
Surprise	30	4002
Neutral	30	6198

TABLE II

Method	Accuracy on FER2013
Proposed Method	74.4%
Work by A. Mollahosseini et al. [21]	66.4%
Work by B. Kim et al. [22]	71.86%
Work by Z. Yu et al. [20]	61.29%
Work by Y. Tang [23]	68.9%

layer. The hybrid feature set then enters the fully connected layers for further processing. This combined features and used for classification of emotions.

Recent works [20] [21] tested and evaluated their method separately with each dataset. So we also evaluated the system separately with JAFFE and FER2013 datasets. For evaluation, both datasets are split in the ratio of 80:20 for training and testing respectively. Performance of the system was evaluated using the metric accuracy and its formula is shown in Equation
1.

$$Accuracy = Num.\frac{correctlypredictedsamples}{TotalNum.samples} \quad (1)$$

Performance analysis is performed with and without pre-processing. First, we evaluated using FER2013 dataset. From Table 3, it is clear that without including pre-processing steps, the accuracy obtained was 54.2% and when cropping is applied, it resulted in significant increase in the accuracy to 74%. When both intensity normalisation and cropping is done, then the accuracy increased to 74.4%. Then evaluation was performed using JAFFE dataset. Without including pre-processing steps the accuracy obtained was 90.698%. When cropping is applied, it resulted in a slight increase in the accuracy to 91.2%. No change in accuracy was there when intensity normalisation is applied, since the dataset is small compared to FER2013. It is clear that there is a significant increase in accuracy when pre-processing was done on the raw image. Cropping increases the accuracy since it removes unwanted portions of the image, which are not relevant for classifying facial emotions. When intensity normalisation is done then also there is a slight increase in accuracy. FER2013 dataset contains more images compared to JAFFE and some of these images are aligned and distorted, that is why accuracy

on FER2013 is less compared to JAFFE. Results obtained on this system is compared with other works in Table 2. It is clear that our proposed work is comparable to the existing methods. As described above, proposed methods combine feature

Intensity normalisation	53.2%	55.2%	54.8%
Both cropping and intensity normalisation	73.5%	74.4%	73.2%

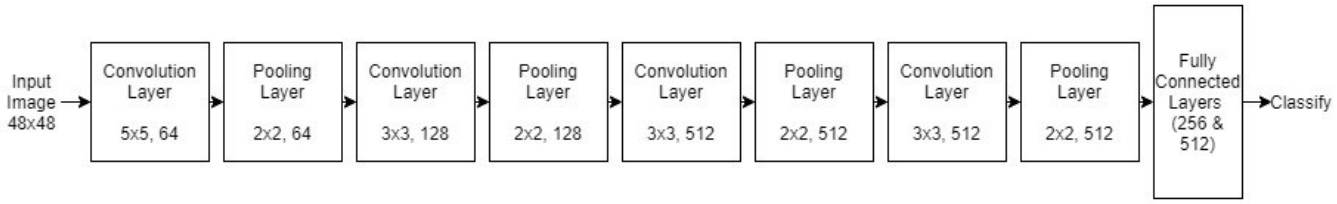

Fig. 6. Architecture of proposed CNN

descriptors with CNN. The system is tested by implementing HoG and Facial landmarks. From Table 4, it is clear that there is a significant increase in accuracy when HoG is applied. When both HoG and facial landmarks are applied, there is a slight decrease in accuracy, which is due to over-fitting.

TABLE III
ACCURACY WHEN FEATURE DESCRIPTORS ARE INCLUDED

Experiments	Accuracy	
	JAFFE	FER2013
CNN (on raw data)	90.698%	73.5%
CNN + Facial landmarks	91.2%	74.4%
CNN + Facial landmarks + HOG	90.698%	73.2%

V. CONCLUSION

This paper proposed a real-time facial emotion recognition system. In the proposed method, seven different facial expressions of different people from two datasets, JAFFE and FER2013 have been analysed. The facial images were pre-processed after it was captured, from which the features were extracted and the emotion was detected by the CNN model based on the training. To measure the performance of the proposed algorithm and check the results, the system was evaluated using the metric accuracy. The same datasets were used for both training and testing by dividing the datasets into training samples and testing samples in the ratio of 80:20 for both JAFFE and FER2013 datasets. Experiment results on two databases, JAFFE and the FER2013 dataset, show that the proposed method can achieve an excellent performance. An accuracy of 91.2% and 74.4% was obtained on JAFFE and FER2013 database respectively. The inclusion of pre-processing methods like cropping and intensity normalisation and feature extraction methods like HoG and facial landmarks improved the accuracy. Facial emotion recognition is still a very challenging problem. More efforts can be made to improve the classification performance for important applications. Future work on the proposed method can be improving the performance of the system and deriving more appropriate classifications with additional pre-processing methods, and combining other feature extraction methods.

TABLE IV
ACCURACY WHEN INCORPORATING PREPROCESSING METHODS.

Experiment	CNN (raw data)	CNN + Landmarks	CNN + Landmarks + HOG
No pre-processing	53.1%	54.8%	54.2%
Cropping	73.5%	74%	73.1%

REFERENCES

[1] S. Is¸ık-Ulusoy, S. Gulseren¨ , N. Ozkan¨ , and C. Bilen, "Facial emotion recognition deficits in patients with bipolar disorder and their healthy parents," *General Hospital Psychiatry*, vol. 65, 04 2020.

[2] S. Lenzoni, V. Bozzoni, F. Burgio, Beatrice, A. Wennberg, A. Botta, E. Pegoraro, and C. Semenza, "Recognition of emotions conveyed by facial expression and body postures in myotonic dystrophy (dm)," *Cortex*, vol. 127, pp. 58–66, Jun. 2020.

[3] S. Passardi, P. Peyk, M. Rufer, T. S. H. Wingenbach, and M. C. Pfaltz, "Facial mimicry, facial emotion recognition and alexithymia in post-traumatic stress disorder." *Behaviour research and therapy*, vol. 122, p. 103436, 2019.

[4] P. Ekman and Friesen, "Facial action coding system: Investigator's guide," *; Consulting Psychologists Press: Palo Alto, CA, USA*, p. 9993626619., 1978.

[5] M. Ozdemir¨ , B. Elagoz¨ , A. Alaybeyoglul , R. Sadighzadeh , and A. Akan, "Real time emotion recognition from facial expressions using cnn architecture," 10 2019.

[6] S. Jaiswal and G. Nandi, "Robust real-time emotion detection system using cnn architecture," *Neural Computing and Applications*, 10 2019.

[7] A. Mollahosseini, D. Chan, and M. Mahoor, "Going deeper in facial expression recognition using deep neural networks," 11 2015.

[8] Bo-Kyeong Kim , Suh-Yeon Dong , Jihyeon Roh, Geonmin Kim , and Soo-Young Lee , "Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach," 07 2016.

[9] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," 12 2016.

[10] D.A. Pitaloka, A. Wulandari, T. Basaruddin, and Dewi Yanti Liliana, "Enhancing cnn with preprocessing stage in automatic emotion recognition," *Procedia Computer Science*, vol. 116, pp. 523–529, 12 2017.

[11] A. Lopes, E. Aguiar, A. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognition*, vol. 61, 07 2016.

[12] M-W. Huang, Z-W. Wang, and Z-L. Ying, "A new method for facial expression recognition based on sparse representation plus lbp," *Proceedings - 2010 3rd International Congress on Image and Signal Processing, CISP 2010*, vol. 4, pp. 1750 – 1754, 11 2010.

[13] S.G. Patro and K. K. Sahu, "Normalization: A preprocessing stage," *IARJSET*, 03 2015.

[14] G. Kumar and P. K. Bhatia, "A detailed review of feature extraction in image processing systems," in *2014 Fourth International Conference on Advanced Computing Communication Technologies*, 2014, pp. 5–12.

[15] A. Nandi, P. Dutta, and Md. Nasir, "Automatic facial expression recognition using histogram oriented gradients (hog) of shape information matrix," in *Intelligent Computing and Communication*, V. Bhateja, S. C. Satapathy, Y.-D. Zhang, and V. N. M. Aradhya, Eds. Singapore: Springer Singapore, 2020, pp. 343–351.

[16] M. Bodini, "A review of facial landmark extraction in 2d images and videos using deep learning," *Big Data and Cognitive Computing*, vol. 3, p. 14, 02 2019.

[17] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," 12 2016.

[18] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "The japanese female facial expression (jaffe) database."

[19] Wolfram Research, *FER-2013*. Wolfram Data Repository, 2018.

[20] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," November 2015.

[21] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," *CoRR*, vol. abs/1511.04110, 2015. [Online]. Available: http://arxiv.org/abs/1511.04110

[22] B. Kim, S. Dong, J. Roh, G. Kim, and S. Lee, "Fusing aligned and nonaligned face information for automatic affect recognition in the wild: A deep learning approach," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 1499–1508.

[23] Y. Tang, "Deep learning using support vector machines," *CoRR*, vol. abs/1306.0239, 2013. [Online]. Available: http://arxiv.org/abs/1306.0239

[24] S. Manoharan, "Image detection, classification and recognition for leak detection in automobiles," *Journal of Innovative Image Processing*, vol. 1, pp. 61–70, 12 2019.