International Journal of All Research Education and Scientific Methods (IJARESM), ISSN: 2455-6211
Volume 10, Issue 11, November-2022, Impact Factor: 7.429, Available online at: www.ijaresm.com

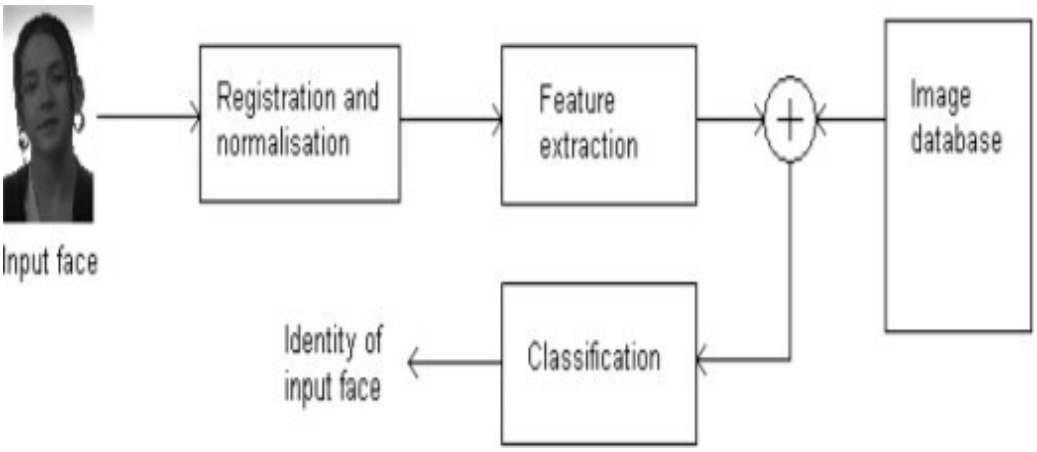
Passenger Security Using Facial Emotion Recognition System

A.D. Bhanu Prakash[1], Nikunj Arvindbhai Gothadiya[2], Atharva Hatekar[3], Sahil Das[4], Tejaswini Prashant Vibhute[5], Abhishek Garimella[6], Revanth Shahukaru[7]

----------**********----------

1. ABSTRACT

In this project we are going to make a "PASSENGER SECURITY USING FACIAL EMOTION RECOGNITION" system using DCNN (Deep convolutional neural network). It has proved to work with much greater accuracy than CNN (convolutional neural network). The facial expression of we humans' changes and we only have the intelligence of understanding the meaning or classify the facial expression as - Happy, Sad, Angry, Fear, Surprise, Disgust and Neutral. So, our aim in this project is to show the mood of the person in front of the camera. And this needs that we provide our computer with the intelligence required to do so like our brains. Our brains have neural networks which are responsible for all kinds of thinking (decision making, understanding) that we do and we try to develop these neuron capabilities artificially called artificial neural network. Moreover, we will be using the concept of the deep convolutional neural network to build the application which can tell what the expression on the face of the person is.



2. OBJECTIVE

This paper objective is to introduce the needs and applications of facial expression recognition. Between Verbal & Non-Verbal form of communication facial expression is a form of non-verbal communication but it plays a pivotal role. It expresses the human perspective or filling & his or her mental situation. We will be identifying suitable image processing techniques for facial expression recognition. The Objective of the project is also to create a product suitable for Taxi Companies, to recognize any kind of fear detection among the passengers travelling, in order to promote Safety and comfort of passengers at a good level.

3. LITERATURE REVIEW

This project concerns about recognition system through identifying the facial expression of human being. It has often been said that the eyes are the "window to the soul". This statement may be carried to a logical assumption that not only the eyes but the entire face may reflect the "hidden" emotions of the individual. The human face is the most complex and versatile of all species. For humans, the face is a rich and versatile instrument serving many different functions, it serves as a window to display one's own motivational state. This makes one's behaviour more predictable and understandable to others and improves communication. A quick facial display can reveal the speaker's attitude about the information being conveyed. As Facial expressions can indicate emotion and pain, regulate social behaviour, and reveal brain function, Research in psychology has indicated that at least six emotions are universally associated with distinct facial expressions. The six principal emotions are happiness, sadness, surprise, fear, anger, and disgust. Several other emotions and many combinations of emotions have been studied but remain unconfirmed as universally distinguishable.

4. LITERATURE SURVEY

Sr. No.	Title	Methodology	Issues	Dataset	Metrics Used
1.	Facial emotion Reconization using multi model information.	Machine recognization of facial expression by correctly integrating audio and video data.	Sometimes emotion Misclassification happens.		Weighting matrix
2.	Emotion recognition system using short-term monitoring of physiological signals	The system was developed to operate as a user-independent system, based on physiological signal databases obtained from multiple subjects. vector machine was adopted as a pattern classifier	Could not employ non-linear or chaotic analyses, as they usually require long-term monitoring of signals.	physiological signal databases obtained from multiple subjects.	Vector
3.	Multilinear Image Analysis for Facial Recognition.	recognition method based on multilinear analysis analogous to the conventional one for linear PCA analysis.		A facial image database.	conventional matrix singular value decomposition .
4.	Low dimensional	Principal component		A facial image	Eigenvector matrix.

	procedure for the characterization of human faces.	analysis.		database.	
5.	Facial Expression Recognition Using 3D Facial Feature Distances	Distance measures extracted from 3D face vectors	Almost all of the methods developed to use the 2D distribution of facial features as inputs into a classification system, and the outcome is one of the facial expression classes. They differ mainly in the facial features selected and in the classifiers used to distinguish between the different facial expressions.	BU-3DFE database	
6.	Facial Expression Recognition Using 3D Facial Feature Distances	Distance measures extracted from 3D face vectors Neural network Classifier	Sometimes emotion Misclassification happens.		Weighting matrix
7.	Aggregating Local Image	Image indexing scheme for		INRIA Holidays	Mean Average

	Descriptors into Compact Codes	very large databases. Bag-of-words, Fisher Kernel and VLAD representations		Oxford5K Building Exalead100 M	precision recall@R
8.	Facial Expression Recognition using Neural Network.	Feature extraction using principal component analysis and feed forward back propagation neural network method.	Sometimes emotion Misclassification happens.		Weighting matrix
9.	A Human Facial Expression Recognition Model based on Eigen Face Approach	Face Detection is done using Hue-Saturation Value			Euclidean Distance Measure
10	A Real-Time Facial Expression Classification System Using Local Binary Patterns	Haar Classifier based method for face detection LBP based feature extraction method Dimensionality is reduced using Principal Component Analysis	The proposed methodology is limited to classify frontal image only. However, rotation of face or occlusions degrades the performance of the system.		
11	A Comprehensive Survey on Techniques	1. Feed-Forward Neural Network	Problems faced in homogenous faces.	Yale and JAFFE Databases	Support Vector Machine (SVM)
	for PASSENGER SECURITY USING FACIAL EMOTION RECOGNITION	2. Multiple Deep Convolutional Neural Networks (CNN)			

5. REQUIREMENTS

As the goal of the project is highly targeted towards research, the entire system was developed in Python, a high-level language and scientific environment. Its capabilities are enhanced by using the integration with OpenCV, a library of functions mainly aimed towards Computer Vision usage.

We will use Cohn-Kanade dataset, a free and publicly available dataset on the web, for training and testing the face recognition system.

6) INVENTION DETAILS

6.1. Objects of the Invention

1. Field of the Invention

The present invention relates to systems and methods for automatic facial recognition, including one-to-one and one-to-many correlations. More particularly, it relates to a system for correcting pose or lighting prior to determining recognition.

2. Discussion of Related Art

Automated facial recognition systems, used to identify individuals from images of faces, have existed for some time. There are several different types of facial recognition systems. In all such systems, a newly acquired image is compared to one or more stored images. Generally, facial recognition systems can be separated into two categories: authentication systems and identification system.

In both authentication systems and identification systems, the images are processed in various ways to determine matches. In some systems, the images are analyzed to locate specific facial features. In other systems, the entire image is analyzed to determine relationships between light and dark areas in the image. In any type of facial identification system, variations in the conditions under which an image is acquired can affect the characteristics of an image and the ability of the system to determine matches. For example, differences in lighting change the shadowing on the face and the associated light and dark areas. For best results, a face should be illuminated by a uniform light from the front. Furthermore, differences ina pose can affect the characteristics of the image.

In both authentication systems and identification systems, the images are processed in various ways to determine matches. In some systems, the images are analyzed to locate specific facial features. The relative locations of the facial features are compared to determine a match. In other systems, the entire image is analyzed to determine relationships between light and dark areas in the image. In any type of facial identification system, variations in the conditions under which an image is acquired can affect the characteristics of an image and the ability of the system to determine matches. For example, differences in lighting change the shadowing on the face and the associated light and dark areas. For best results, a face should be illuminated by a uniform light from the front. Furthermore, differences in a pose can affect the characteristics of the image. For best results with matching, the individual should look directly at the camera. If an individual is looking in a different direction, the distances between facial

features change due to differences in perspective. Generally, authentication systems provide more uniform images than for identification systems. The individuals are cooperative and can be directed to look directly at a camera with proper illumination for acquiring the images, both for the database and for the comparison image. Often, in an identification system, the subject is not cooperative and lighting conditions are poor. Therefore, a need exists for a system which allows modification of images to correct for differences in lighting or pose, in either identification systems or authentication systems.

6.2. Summary of Invention

The present invention substantially overcomes the deficiencies of the prior art by providing a facial recognition system which processes images to correct for lighting and poses prior to comparison. According to another aspect of the invention, the images are corrected for lighting and pose by using shape information. The system processes two-dimensional image of a face to create a three-dimensional image of the face. The three-dimensional image is manipulated to change the pose and lighting characteristics. Finally, the modified three-dimensional image is converted back to a two-dimensional image prior to processing for recognition. According to another aspect of the invention, the three-dimensional image is manipulated to be facing forward and with diffuse light from the front.

According to another aspect of the invention, a facial recognition system compares a newly acquired image of a face to images of faces in a database to determine a match. The newly acquired image includes one or more two-dimensional images. The system processes the one or more two-dimensional images to create a three-dimensional image of the face. The three-dimensional image is manipulated to change the pose and lighting characteristics. According to an aspect of the invention, the three-dimensional image is manipulated to be facing forward and with diffuse light from the front. The three-dimensional image is processed to create a second two-dimensional image. The three-dimensional images are stored in a database for later comparison with another image. According to another aspect of the invention, the three-dimensional images are converted into two-dimensional images before being stored in a database. According to another aspect of the invention, the three-dimensional images are manipulated to be facing forward and with diffuse light from the front. The second two-dimensional image is compared with images in the database to determine whether a match exists. According to another aspect of the invention, an iterative process is used to create a three-dimensional image from the original two-dimensional image. An initial shape is used with data from the two-dimensional image to create a three-dimensional shape. The three-dimensional shape is iteratively adjusted to match the original image. According to an aspect of the invention, at each iteration, a two-dimensional image is rendered from the three-dimensional image.

6.3. PROPOSED SYSTEM

This project aims to predict the emotions of the person by his/her facial expression and swap the appropriate emoticon in place of the face.

There are seven types of human emotions shown, that are universally recognised across different cultures namely anger, happiness, disgust, fear, sadness, surprise and contempt.

STEP 1: Preparing the dataset
It involves the preparation of dataset upon which the learning algorithm will work. We will be using different datasets of faces available on internet example CK+ dataset and one of our own and applying it to convolution neural networks. We will be training the data to recognize different emotions.

STEP 2: Detecting Real-time passengers' faces
This is done by HAAR cascade function in OpenCV After detecting the faces the image is converted to greyscale and is resized to the same size as the images in a dataset.

STEP 3: Learning to Recognise Emotions
This step involves training the program to recognise and differentiate between the emotions.
Training is done by Convolution Neural Network, a deep learning algorithm and will be on the layers on the dataset of faces

STEP 4: Swapping emoticons in place of faces
This is the final step and it involves placing the appropriate emoticon over the face of the person according to his/her emotion.

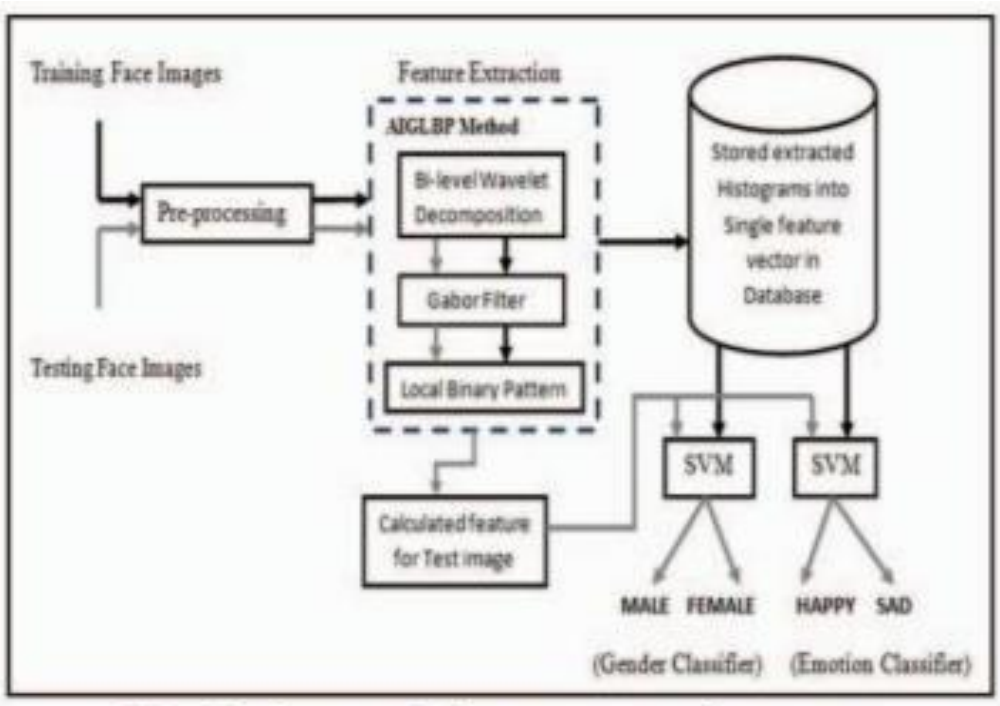
The HAAR cascade function returns the coordinates of the faces detected and these coordinates can be used to place the emoticon at the exact place.

- **STEP 5: Notifying Taxi Company in case of an emergency**

When the software detects Fear continuously for 10 seconds, an auto- mail is generated and sent to the company notifying them, the passenger(s) are scared or are in trouble.

6.3.1 RELATED WORK:

A number of works exist in literature in which most of the researchers working on expression detection have focused on the application of a variety of feature extraction and classification techniques on face images. Some work had been done in which different approaches are used for emotion detection. They also introduced the effectiveness of face alignment on the accuracy of facial emotion classification:



Work flow of the proposed system

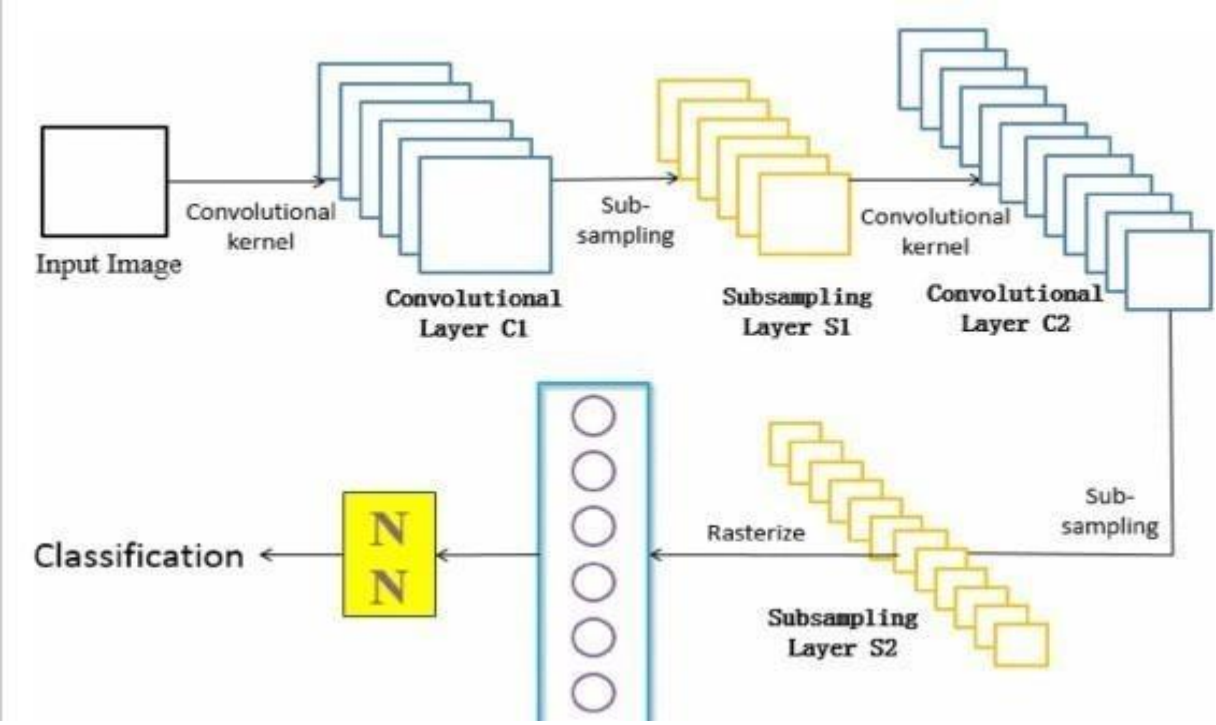
Recent work on facial expression analysis and recognition has used these "basic expressions" or a subset of them. Some used optical flow (OF) to recognize facial expressions. Some used a flexible shape and appearance model for image10 coding, person identification, pose recovery, and facial expression recognition.

6.4 Architecture of project

In this project, we will be making the expression recognition system using the DCNN (deep convolution neural network). This approach has been proven to be better than CNN.

Convolutional Neural Networks (CNN) is composed of two basic layers,

respectively called convolutional layer (C layer) and subsampling layer (Slayer). Different from general deep learning models, CNN can directly accept 2D images as the input data, so that it has a unique advantage in the field of image recognition. A classic CNN model is shown as in the below figure. 2D images are directly inputted into the network, and then convoluted with several adjustable convolutional kernels to generate corresponding feature maps to form layer C1. Feature maps in layer C1 will be subsampled to reduce their size and form layer S1. Normally, the pooling size is 2×2. This procedure also repeats in layer C2 and layer S2. After extracting enough features, the two-dimensional pixels are rasterized into 1D data and inputted to the traditional neural network classifier.



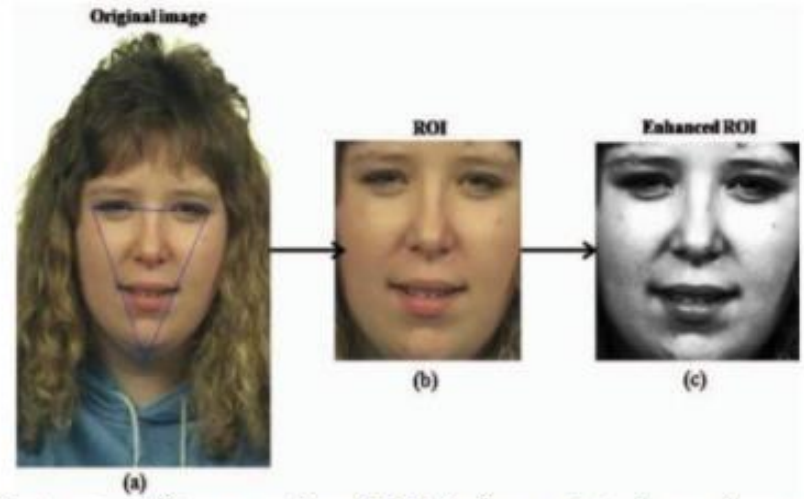
6.5 MODULES:

In this project, we will be using 4 modules-

(i) Input module
(ii) Pre-processing module
(iii) Recognition module
(iv) Output module.

6.5.1 PRE-PROCESSING:

Mostly, the images available in face databases contain some irrelevant details from a classification point of view (e.g. background, hair etc.). The region comprising of only face information is termed as the relevant area of interest and it should be cropped before feature extraction. To find the relevant area, the centre of the chin right and left corners of the eyes, three-point corners are selected on the face. To make triangle these three points are linked with each other. Then, the computational process is being performed on the centroid of the triangle (say, C) and the distance of a corner from the centroid (say, P). From centroid C, a square of length (say, 2P) is drawn. Within this square, the captured area of the image is the desired face image. Now, grayscale images are generated from the cropped face images. The relevant cropped area i.e. face image are of different size, and hence images are further resized to 230X230 pixels by using bicubic interpolation technique.



Extracted face region (ROI) from database image

6.5.2. Algorithms to be used:

1. Face detectionalgorithms

For static images, Viola-Jones, which achieves fast and reliable detection for frontal faces will be used.

2. FEATURE EXTRACTION TECHNIQUE:

After the pre-processing step, the feature extraction technique is applied to the face image to extract the features from the image. In this work approximation image, Gabor local binary pattern features method is applied on the face image for extracting the features as discussed below.

1. AIG BP (Approximation image Gabor Local Binary Pattern)

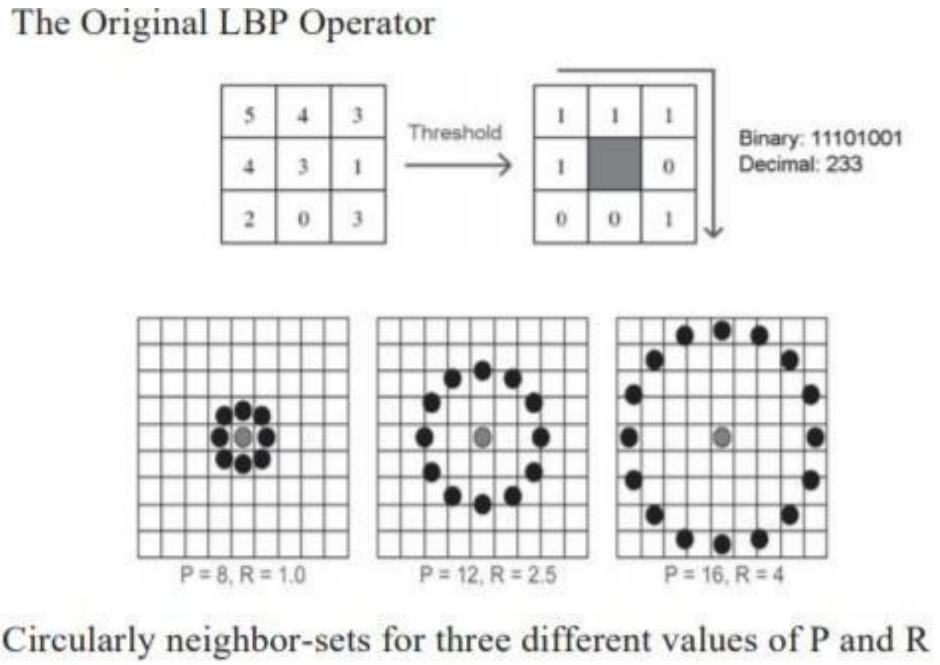
Initially, on the face images, bi-level wavelet decomposition method has been applied, which transforms face images into approximation images. Then, on approximation images, Gabor filter has been applied, which restricts distortion and noise present at a distinct location in the image up to a certain extent and also provides robustness against brightness and contrast of images. Further, local binary patterns (LBP) have been used to extract local features of the face images. AIGLBP method is a combination of wavelets decomposition and Gabor filter along with LBP method which is effective in terms of accuracy and it reduces time computation.

2. Bi-level wavelet decomposition

Wavelet decomposition method is a time-frequency signal analysis method. It can be used to decompose a face image into many sub-band images with different spatial resolution, frequency characteristic and directional features. In this work, the approximation coefficient is being computed by decomposing the face image up to two levels. Approximation coefficients contain the lowest frequency components and details coefficients contain the highest frequency component of an image. Instead of considering the entire coefficients, only approximation coefficients are taken for further procedure. Approximation coefficients contain low-frequency components of the face image, which carry whole information of the face. The change of expression and small scale obstruct does not affect the low-frequency part but the high-frequency part of the image only. More than two level decomposition of face image result in information loss and hence, not considered in this work.

3. Gabor filters

In the fields of computer vision, pattern recognition and image processing, Gabor filter have a large number of applications. 2D Gabor filter is a selective filter in terms of frequency and orientation. Gabor filter response hasn't been disturbed by noise and distortion exists at different locations due to accuracy in time-frequency localization. Hence, the performance of the Gabor filter is to mark for noisy images [29]. As modulated by Gaussian envelope, for particular frequency and orientation, Gabor filter is being considered as a sinusoidal plane. Gabor filter is applied at different orientations to get face image information. Gabor function has cosine (real part) and sine (imaginary part) functions of the sinusoidal plane. In the proposed work, for further processing, only the real parts of the Gabor function are to be considered. The experimental analysis of the imaginary part also proves the same.



Circularly neighbor-sets for three different values of P and R

4. Principles of local binary pattern

The outcome of the Gabor filter has been applied to LBP (local binary pattern) in which face image is first split into small regions from which local binary patterns also known as histograms are being extracted. Different LBP histograms extracted from each of the face images are finally concatenated into single feature histogram, which forms the representation of the face image. Using AIG BP method for the feature extraction process, when we applied LBP technique along with wavelets and Gabor filters, then the computational time of overall process gets reduced, which increases the efficiency of the system and also tends to improve the overall performance of the system.

SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a supervised learning algorithm which has been based on the hyper plane's concept that aims to separate a set of objects with maximum distance. After extracting the features from the image, SVM finds support vectors from the feature area. These vectors help to determine the optimal hyperplane.

Classification algorithms

A set of faces with corresponding labels are fed into a classifier, which upon training, it learns and predicts the emotion class for a new face.

1. Keras Model-

Keras is an open source neural network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, or Theano. Designed to enable fast experimentation with deep neural networks, it focuses on being user- friendly, modular, and extensible. It was developed as part of the research effort of project ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System), and its primary author and maintainer are François Chollet, a Google engineer.

In 2017, Google's TensorFlow team decided to support Keras in TensorFlow's core library. Chollet explained that Keras was conceived to be an interface rather than a standalone machine-learning framework. It offers a higher-level, more intuitive set of abstractions that make it easy to develop deep learning models regardless of the computational backend used. Microsoft added a CNTK backend to Keras as well, available as of CNTK.

Features-

- Keras contains numerous implementations of commonly used neural network building blocks such as layers, objectives, activation functions, optimizers, and a host of tools to make working with image and text data easier.
- Keras allows users to productize deep models on smartphones (iOS and Android), on the web, or on the Java Virtual Machine. It also allows the use of distributed training of deep learning models on clusters of Graphics Processing Units (GPU) and Tensor processing units (TPU).

A tensor processing unit (TPU) is an AI accelerator application-specific integrated circuit (ASIC) developed by Google specifically for neural network machine learning.

A graphics processing unit (GPU) is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device. GPUs are used in embedded systems, mobile phones, personal computers, workstations, and game consoles. Modern GPUs are very efficient at manipulating computer graphics and image processing, and their highly parallel structure makes them more efficient than general-purpose CPUs for algorithms where the processing of large blocks of data is done in parallel. In a personal computer, a GPU can be present on a video card, or it can be embedded on the motherboard or—in certain CPUs—on the CPU die.

6.5. Detailed Description of the Invention

The present invention relates to an improved method and apparatus for facial recognition. Generally, various facial recognition systems are known. In such systems, one or more images are compared with one or more previously stored images to determine whether a match exists. Typically, facial recognition systems fall into two categories, authentication and identification. An authentication system is used to determine whether a person is who he claims to be. The authentication system includes a database having images and identification information regarding a set of people. A specific image is retrieved based upon the identification information. The retrieved image is compared to a newly acquired image of the person to determine whether a match exists. If the images are sufficiently similar, then the person is determined to correspond to the stored person. Authentication systems can be used to control access to secure locations or information. If the person matches the stored image, then access is allowed. Otherwise, access is denied. In identification systems, the person needs to be identified.

According to another embodiment of the present invention, a standard shape is used as the initial three-dimensional shape. The standard shape is created as a threshold when the error value does not change significantly between iterations or after a predetermined maximum number of iterations. Once the shape has been properly estimated, the three- dimensional model needs to be adjusted to correct for the lighting. Specifically, the intensity mask for the face is restored using the determined shape. Image 112 illustrates the recovered shape with adjusted lighting. With a given three dimensional shape, two- dimensional images can be created using any desired lighting conditions using a Lambertian surface assumption. When the surface of an object is assumed to have a Lambertian reflectance property (meaning that it irradiated illuminating light uniformly in all directions) an image generation can be mathematically described Image=A*max (B*s,0) (1) Where

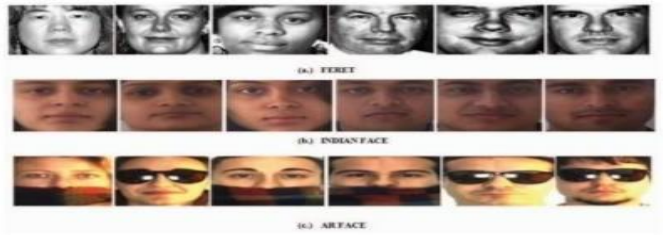
A—Intensity, color map of the surface, s=(s1,s2,s3)−Light Source,

B = (n 1 1 , n 2 1 , n 3 1 n 1 2 , n 2 2 , n 3 2 ⋯ n 1 k , n 2 k , n 3 k) the matrix of normal vectors to the facial surface
In the case of human flesh, a Lambertian assumption describes actual surface property almost perfectly. Using this assumption it is possible to generate artificial images, which would be hard to distinguish from naturally acquired ones. The final artificial image 113 is used in the facial recognition engine 40 for determining a matching image.

EXPERIMENTAL ANALYSIS

Python platform is used to perform all the simulations operations. The performance is measured by classification rate (CR). The classification rate is considered as the percentage of correctly classified face images with the overall face images in the testing set. Face databases used for experimentation The evaluation of the proposed system has been done by using a number of publicly available face databases having sufficient amount of male and female images with their facial expressions taken in conditional as well as unconditional Environments. Face databases used in this work are:

(a) FERET
(b) INDIAN FACE
(c) AR FACE



Standard face image databases

Performance Measures:

The proposed work used 2-fold cross-validation in which the overall dataset is split into two groupsets, let say (D1) and (D2) of equal size. The proposed system is trained with a group set (D1) and testing has been done on group set (D2). As both the training and testing group sets are large in size, the idea of 2 cross-validations helps the testers to get an accurate conclusion in short span of time. During the training phase, the proposed system uses 50% of total images and the rest of the images used in the testing phase and each sample image is used for both training and testing during each fold. With two-fold cross-validation, computation time gets reduced, which is an important necessity in real systems. The purpose of the proposed system is to do correct classification the face images as a male/female with their emotions as happy/sad. The system is developed with the aim of real-life deployment and hence the criterion suits the purpose, i.e., Classification rate (CR %), which is to be considered to evaluate system performance. Classification rate concludes the accuracy of the system by showing a number of images which is correctly classified. The classification rate gives result in the percentage of correctly classified face images from a total number of face images in the testing set.CR% of a system is defined by

$$CR\% = M1/M2 * 100$$

Where M1 denotes the number of correctly classified face images and M2 is the total number of face images in the testing set.

Experimental Results:

In this proposed system we have two classifications results i.e. gender classification and emotion detection. With the captured face image, different results are obtained with a number of publicly available databases as there are variations in expressions and poses of the face images. FERET is used by most of the researchers due to good quality images. Also, INDIAN FACE database contains images of a bright homogeneous background. AR FACE database contains occluded face images i.e. person images in this database wears scarf and sunglasses and it is used for cross-database validation.

I Testing the performance for normal faces images

Considering the classification of gender, the proposed system has achieves 94.98 % of classification rate (CR) with the FERET database. With the Indian face database, the system has obtained 90% of the classification rate. Whereas for the emotion classification the proposed system has achieves79.33 % of classification rate (CR) with the FERET database. With the Indian face database, the system has achieves 80 % of the classification rate.

Dataset	Classification Rate (CR %)	
	Gender classification	*Emotion classification*
FERET	94.98	79.33
INDIAN FACE	90	80

Classification rate of gender and emotion classification

II. Testing the performance for cross-database

Occlusion on face images usually occurs when a person wears sunglasses, suffers an injury on faces, covers his/her face with scarf or hand, or puts a mole on his/her face. To get cross- database performance, the system uses non-occluded face images for training (FERET and INDIAN FACE, in our case), but testing is done for gender classification on occluded images (i.e. AR Face database, in our case). Moreover, the system is independent of person, i.e. face images used for training and testing are of a different set of people. With the FERET database, the proposed work gives accuracy up to 58.52% and with INDIAN FACE database, the system gives an accuracy rate of up to 52.68%. TABLE 2 indicates the cross-database performance with standard databases.

Dataset	Classification Rate (CR %)
Training/Testing	AR FACE dataset
FERET	58.52
INDIAN FACE	52.68

Cross database classification rate for non occluded images

CHOSEN PARAMETERS:

- **Uniform Patterns**
- **LBP Operator**
- **Region Sizes**
- **Region Weights**

Assigned Region Weights

6.6. Working Scenario

6.6.1 Implementation – Screenshots

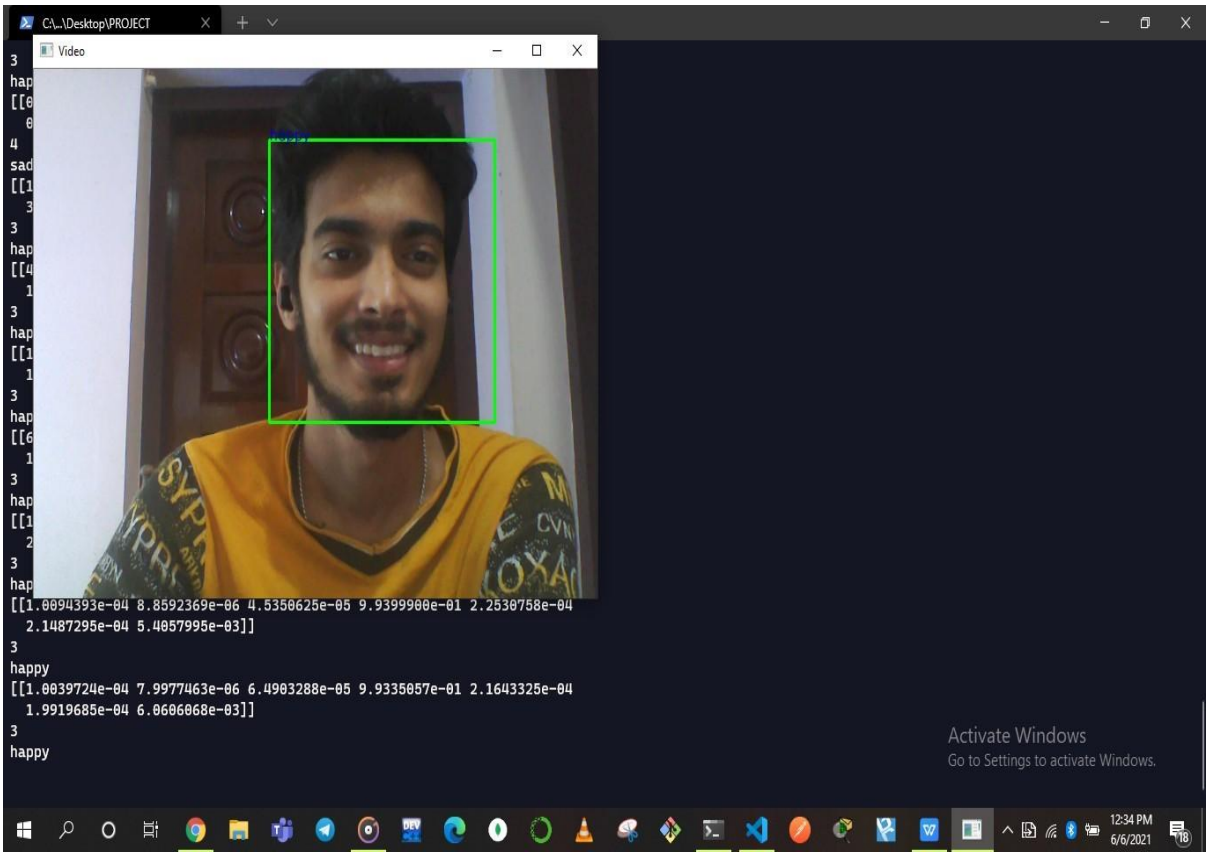


Fig 6.6.1.1. Detecting faces and emotions respectively

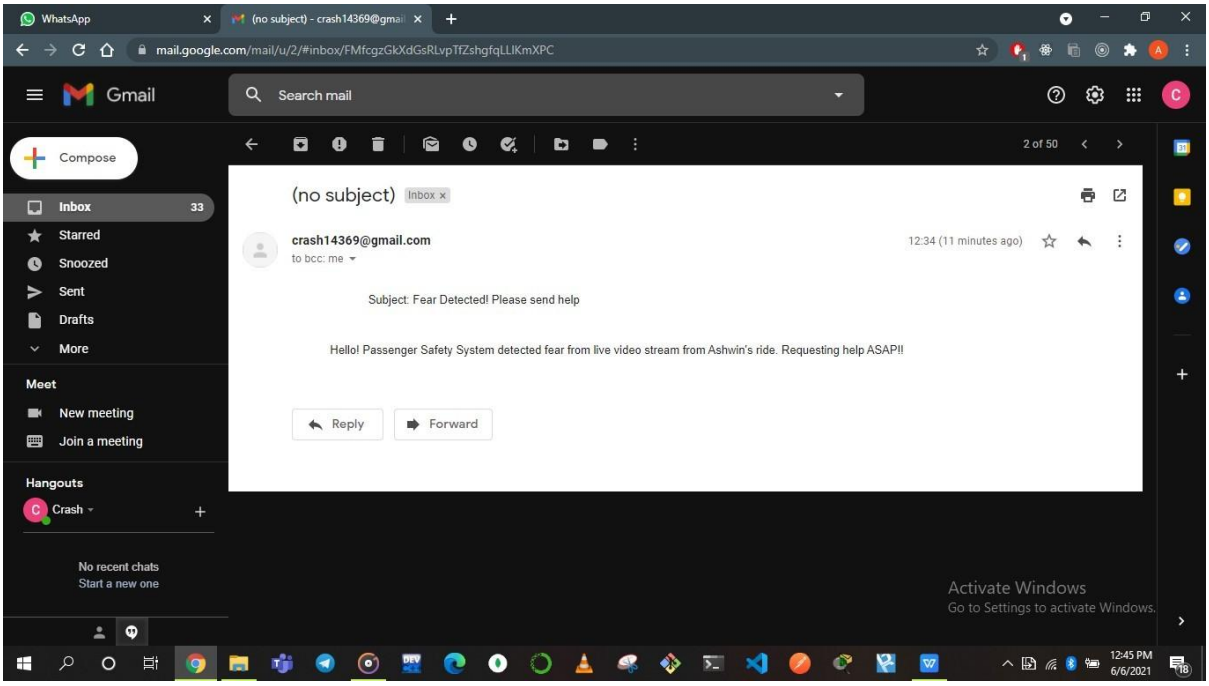


Fig 6.6.1.2. Mail Generated after detecting fear

CONCLUSION

Emotion recognition technology has come a long way in the last twenty years. Today, machines are able to automatically verify identity information for secure transactions, for surveillance and security tasks, and for access control to buildings etc. These applications usually work in controlled environments and recognition algorithms can take advantage of the environmental constraints to obtain high recognition accuracy. However, next generation face recognition systems are going to have widespread application in smart environments -- where computers and machines are more like helpful assistants.

To achieve these goal computers must be able to reliably identify nearby people in a manner that fits naturally within the pattern of normal human interactions. They must not require special interactions and must conform to human intuitions about when recognition is likely. This implies that future smart environments should use the same modalities as humans, and have approximately the same limitations. These goals now appear in reach -- however, substantial research remains to be done in making person recognition technology work reliably, in widely varying conditions using information from single or multiple modalities.

And hence, putting these goals to a better application and ensuring the security of passengers can create a better business scope of Taxi companies, building trust between passengers and drivers.

REFERENCES

[1] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: a survey," 2019, https://arxiv.org/abs/1905.05055.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.

[3] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8759–8768, Salt Lake City, UT, United States, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn.," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, Venice, Italy, 2017.

[6] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, Springer, 2016.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection,"

in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, Las Vegas, NV, United States, 2016.

[8] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, Honolulu, HI, United States, 2017.

[9] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[10] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.