ATHARVA EDUCATIONAL TRUST'S
ATHARVA COLLEGE OF ENGINEERING
(Approved by AICTE, Recognized by Government of Maharashtra &
Affiliated to University of Mumbai – Est. 1999-2000)
Department of Computer Engineering
Academic Year 2021-22

"Facial expression recognition through optical analysis and deep learning"

Submitted in partial fulfilment of the requirements of the degree
BACHELOR OF ENGINEERING IN COMPUTER ENGINEERING

PRESENTED BY –

Full Name Reg/Roll No: Saheel Chavan (14)
Full Name Reg/Roll No: Malhar Ravindra Gudekar (24)
Full Name Reg/Roll No: Swetha Chandrasekar Iyer(26)

Supervisor
Prof. Ranjita Asati

ATHARVA EDUCATIONAL TRUST'S
ATHARVA COLLEGE OF ENGINEERING
(Approved by AICTE, Recognized by Government of Maharashtra &
Affiliated to University of Mumbai – Est. 1999-2000)
Department of Computer Engineering
Academic Year 2021-22

CERTIFICATE

This is to certify that the Mini Project entitled "Facial expression recognition through optical analysis and deep learning" is a bonafide work of **Saheel Sanjay Chavan(14), Malhar Ravindra Gudekar (24) and Swetha Chandrasekar Iyer (26)**
submitted to the University of Mumbai in partial fulfilment of the requirement for the award of the degree of **"Bachelor of Engineering" in "Computer Engineering".**

(Prof.) Ranjita Asati
Supervisor

30/4/2022

(Dr. Suvarna Pansambal)
Head of Department

(Dr. Shrikant Kallurkar)
Principal

ATHARVA EDUCATIONAL TRUST'S
ATHARVA COLLEGE OF ENGINEERING
(Approved by AICTE, Recognized by Government of Maharashtra &
Affiliated to University of Mumbai – Est. 1999-2000)
Department of Computer Engineering
Academic Year 2021-22

Mini Project Approval

This Mini Project entitled "Facial expression recognition through optical analysis and deep learning" by **Saheel Sanjay Chavan(14), Malhar Ravindra Gudekar (24) and Swetha Chandrasekar Iyer (26)** is **approved** for the degree of **"Bachelor of Engineering" in "Computer Engineering".**

Examiners

Prof. Ranjita Asati

(Internal Examiner Name & Sign)

Prof. Yogesh Gite
30/04/2022

(External Examiner Name & Sign)
Date: 30/04/2022
Place: MUMBAI

Acknowledgement

We cannot express enough thanks to our project guide **Prof. Ranjita Asati** for their continued support and encouragement, whose valuable guidance has been the one that helped us in shaping this project. Their suggestions and instructions have served as a major contribution towards the successful completion of the project. We would like to thank our **HOD Dr. Suvarna Pansambal** who gave us this wonderful opportunity to do this project on the topic "Facial expression recognition through optical analysis and deep learning". Then we would also like to thank our college "**Atharva College of Engineering**" to provide us the all the basic resourses which we have used to communicate with our project guide. Finally, we would like to thank our classmates and friends who came out of their ways to help us sort the problems we faced during the completion of this project.

CONTENTS

Abstract

Human Face expression Recognition is one of the most powerful and challenging tasks in social communication. Generally, face expressions are natural and direct means for human beings to communicate their emotions and intentions. Face expressions are the key characteristics of non-verbal communication. The performance of various FER techniques is compared based on the number of expressions recognized and complexity of algorithms. Databases like JAFFE, CK, and some other variety of facial expression databases are discussed in this survey. The study on classifiers gather from recent papers reveals a more powerful and reliable understanding of the peculiar characteristics of classifiers for research fellows.

Emotion recognition is one of the trending research fields. It is involved in several applications. Its most interesting applications include robotic vision and interactive robotic communication.. Facial expressions can be considered as ideal means for detecting the persons' emotions. The proposed approach consists of four phases: pre-processing, key point generation, key point selection and angular encoding, and classification. The main idea is to generate key points using Media Pipe face mesh algorithm, which is based on real-time deep learning. In addition, the generated key points are encoded using a sequence of carefully designed mesh generator and angular encoding modules. Furthermore, feature decomposition is performed using Principal Component Analysis (PCA). This phase is deployed to enhance the accuracy of emotion detection. Finally, the decomposed features are enrolled into a Machine Learning (ML) technique that depends on a Support Vector Machine (SVM), k-Nearest Neighbour (KNN), Naïve Bayes (NB), Logistic Regression (LR), or Random Forest (RF) classifier. Moreover, we deploy a Multilayer Perceptron (MLP) as an efficient deep neural network technique. The presented techniques are evaluated on different datasets with different evaluation metrics. The simulation results reveal that they achieve a superior performance with a human emotion detection accuracy of 97%, which ensures superiority among the efforts in this field.

Introduction

1.1 Introduction:

Facial emotions are important factors in human communication that help us understand the intentions of others. In general, people infer the emotional states of other people, such as joy, sadness, and anger, using facial expressions and vocal tone. According to different surveys, verbal components convey one-third of human communication, and nonverbal components convey two-thirds. Among several nonverbal components, by carrying emotional meaning, facial expressions are one of the main information channels in interpersonal communication. Therefore, it is natural that research of facial emotion has been gaining lot of attention over the past decades with applications not only in the perceptual and cognitive sciences, but also in affective computing and computer animations. Interest in automatic facial emotion recognition (FER) (Expanded form of the acronym FER is different in every paper, such as facial emotion recognition and facial expression recognition. In this project, the term FER refers to facial emotion recognition as this study deals with the general aspects of recognition of facial emotion expression.) has also been increasing recently with the rapid development of artificial intelligent techniques, including in human-computer interaction (HCI) , virtual reality (VR) , augment reality (AR) , advanced driver assistant systems (ADASs) , and entertainment . Although various sensors such as an electromyograph (EMG), electrocardiogram (ECG), electroencephalograph (EEG), and camera can be used for FER inputs, a camera is the most promising type of sensor because it provides the most informative clues for FER and does not need to be worn. The proposed approach consists of four phases: pre-processing, feature extraction and selection, feature decomposition, and classification. Feature extraction and selection is carried out by MediaPipe face mesh algorithm. *is algorithm is based on real-time deep learning. In addition, the feature decomposition phase is performed by PCA. This phase is deployed to enhance the accuracy of emotion detection. It is required to decompose the extracted features using the Singular Value Decomposition (SVD). Finally, the obtained features are enrolled into a selected classifier. In addition, an MLP deep neural network is utilized. The introduced techniques are assessed on different datasets with the help of different evaluation metrics. Moreover, this project report introduces a hardware implementation of the proposed models.

1.2 Motivation:

Two studies examined an unexplored motivational determinant of facial emotion recognition: observer regulatory focus. It was predicted that a promotion focus would enhance facial emotion recognition relative to a prevention focus because the attentional strategies associated with promotion focus enhance performance on well-learned or innate tasks - such as facial emotion recognition. In Study 1, a promotion or a prevention focus was experimentally induced and better facial emotion recognition was observed in a promotion focus compared to a prevention focus. In Study 2, individual differences in chronic regulatory focus were assessed and attention allocation was measured using eye tracking during the facial emotion recognition task. Results indicated that the positive relation between a promotion focus and facial emotion recognition is mediated by shorter fixation duration on the face which reflects a pattern of attention allocation matched to the eager strategy in a promotion focus (i.e., striving to make hits). A prevention focus did not have an impact neither on perceptual processing nor on facial emotion recognition. Taken together, these findings demonstrate important mechanisms and consequences of observer motivational orientation for facial emotion recognition. FER systems have broad applications in various areas, such as computer interactions, health-care systems and social marketing. However, facial expression analysis is incredibly challenging due to subtle and transient movements of the foreground people and complex, noisy environment of the background in the real-world images/videos . here are three main challenges caused by illumination variation, subject-dependence, and head pose-changing; these widely affect the performance of the FER system. The state-of-the-art techniques in FER systems are effective for use in controlled laboratory environments but not for applications in real-world situations.

1.3 Problem Statement & Objective:

Problem Statement:

Facial Emotion Recognition (FER) is the technology that analyses facial expressions from both static images and videos in order to reveal information on one's emotional state. The complexity of facial expressions, the potential use of the technology in any context, and the involvement of new technologies such as artificial intelligence raise significant privacy risks.

Objective:

The main idea of this approach is to deploy deep learning as an automatic key point generator using MediaPipe technique. Hence, a sensitive mathematical process is performed to encode the generated key points into a set of distinguishable features. In addition, different machine learning techniques are implemented on the extracted features to perform the classification task. The proposed approach consists of four main phases. The first phase is image pre-processing in which a super-resolution task is carried out using SRGAN. In the second phase, we deploy MediaPipe to generate key landmarks on the face images. Furthermore, we present a key landmark analysis and an angular encoding module. This module contains three subphases (key landmark selection, emotional mesh generation, and mesh angular encoding). The main idea of this module is to generate an emotional mesh that connects the selected key landmarks. Hence, the obtained mesh is encoded into angular values to generate a feature map The recorded reactions were subsequently compared to the reaction of the image that was expected. The results of the experiment have shown several imperfections of the face analysis system. The system has difficulties classifying expressions and cannot detect and identify inner emotions that a person may experience when shown the image. Face analysis systems can only detect emotions that are expressed externally on a face by physiological changes in certain parts of the face..

LITERATURE SURVEY

2.1 Literature Survey:

[1] Mehrabian A., "Communication without words", Psychology Today, Vol. 2, No. 4, 1968, pp. 53-56.
[2] Ekman, P and Friesen, W, "Facial Action Coding System: A Technique for the Measurement of Facial Movement", Consulting Psychologists Press, Palo Alto, 1978 pp. 10-11.
[3] Samad, Rosdiyana, and Hideyuki Sawada. "Extraction of the minimum number of Gabor wavelet parameters for the recognition of natural facial expressions." Artificial Life and Robotics 16, no. 1 (2011) Springer: pp. 21-31.
[4] Meher, Sukanya Sagarika, and Pallavi Maben. "Face recognition and facial expression identification using PCA." In Advance Computing Conference, 2014 IEEE International, pp. 1093-1098. IEEE, 2014.
[5] Samad, Rosdiyana, and Hideyuki Sawada. "Edge-based Facial Feature Extraction Using Gabor Wavelet and Convolution Filters." In MVA, pp. 430-433. 2011.
[6] Sisodia, Priya, Akhilesh Verma, and Sachin Kansal. "Human Facial Expression Recognition using Gabor Filter Bank with Minimum Number of Feature Vectors." International Journal of Applied Information Systems, Volume 5 – No. 9, July 2013 pp. 9-13.
[7] Thai, Le Hoang, Nguyen Do Thai Nguyen, and Tran Son Hai. "A facial expression classification system integrating canny, principal component analysis and artificial neural network." arXiv preprint arXiv: 1111.4052 (2011).
[8] Abdulrahman, Muzammil, Tajuddeen R. Gwadabe, Fahad J. Abdu, and Alaa Eleyan. "Gabor wavelet transform based facial expression recognition using PCA and LBP." In Signal Processing and Communications Applications Conference, 2014 22nd, pp. 2265-2268. IEEE, 2014.
[9] Sobia, M. Carmel, V. Brindha, and A. Abudhahir. "Facial expression recognition using PCA based interface for wheelchair." In Electronics and Communication Systems, 2014 International Conference on, pp. 1-6. IEEE, 2014.
[10] Poon Bruce, M. Ashraful Amin, and Hong Yan. "Performance evaluation and comparison of PCA based human face recognition methods for distorted images." International Journal of Machine Learning and Cybernetics 2, no. 4 (2011): 245-259.
[11] Rahulamathavan, Yogachandran, RC-W. Phan, Jonathon A. Chambers, and David J. Parish. "Facial expression recognition in the encrypted domain based on local fisher discriminant analysis." Affective Computing, IEEE Transactions on 4, no. 1 (2013): 83-92.
[12] Sarawagi, Varsha, and K. V. Arya. "Automatic facial expression recognition for image sequences." In Contemporary Computing, 2013 Sixth International Conference on, pp. 278-282. IEEE, 2013.
[13] Chao, Wei-Lun, Jun-Zuo Liu, Jian-Jiun Ding, and PO-Hung Wu. "Facial expression recognition using expression-specific local binary patterns and layer denoising mechanism." In Information, Communications and Signal Processing, 2013 9th International Conference on, pp. 1-5. IEEE, 2013.
[14] Vaibhavkumar J. Mistry, Mahesh M. Goyani, "A literature survey on Facial Expression

2.2 Limitations of existing systems:

The emotion recognition techniques provide results in diverse models of emotion representation. Facial expression analysis usually provide the results using Ekman's six basic emotions model extended with neutral state – usually a vector of seven values is provided, each value indicating an intensiveness of: anger, joy, fear, surprise, disgust, sadness, neutral state. Emotion recognition from facial expressions is susceptible to illumination conditions and occlusions of the face parts Facial expression analysis has a major drawback – mimics could be to some extent controlled by humans and therefore the recognition results might be intentionally or unintentionally falsified.

Self-report on emotions, although subjective, is frequently used as a "ground truth" and this approach will be applied in this study. The second approach from the literature is multi-channel observation and consistency. Another approach is manual tagging by qualified observers or physiological observations, but this approach was not used in this study. The abovementioned results influenced decisions on the design of this study, especially use of more than one observation channel and improving illumination conditions.

All automatic emotion recognition algorithms are susceptible to some disturbances and facial expression analysis is not an exception – suffers from face oval partial cover, location of the camera, insufficient or uneven illumination. When compared to a questionnaire (self-report), all automatic emotion recognition methods are more independent on human will and therefore might be perceived as a more reliable source of information on affective state of a user, however inconsistency rate is alarming

2.3 Mini Project Contribution:

We distributed the entire project in subdivisions, mentioned below:

TEAM MEMBER 1:
Saheel Chavan has worked on the model training part and explored various models to be applied on our project.

TEAM MEMBER 2:
Malhar Gudekar has worked on the frontend part and performed data cleaning on the dataset used to train the model.

TEAM MEMBER 3:
Swetha Iyer has performed data pre-processing and also explored and trained various face recognition models.

PROPOSED SYSTEM

3.1 Introduction:

The main idea of this approach is to deploy deep learning as an automatic key point generator using MediaPipe technique. Hence, a sensitive mathematical process is performed to encode the generated key points into a set of distinguishable features. In addition, different machine learning techniques are implemented on the extracted features to perform the classification task. The proposed approach consists of four main phases. The first phase is image preprocessing in which a super-resolution task is carried out using SRGAN. In the second phase, we deploy MediaPipe to generate key landmarks on the face images. Furthermore, we present a key landmark analysis and an angular encoding module. This module contains three subphases (key landmark selection, emotional mesh generation, and mesh angular encoding.

Face Detection CNN

The human faces share a similar shape and texture, the representation learned from a representative proportion of faces can generalize well to detect the others, which are not used in the network training process. The performance of the trained model depends on many factors such as number of images in the training dataset, data augmentation, the CNN architecture, loss function, hyper-parameters adjusting, transfer learning, fine-tuning, evaluation metrics, etc., which leads to a complex set of actions in order to develop the entire pipeline.

The face detector was used to localize faces in images and to align them to normalized coordinates afterwards. The CNN architecture is made up of a backbone network, a localization CNN and the fully connected layers for classification. The backbone networks used for facial recognition were the Inception V2 for SSD architecture and ResNet-InceptionV2 for Faster R-CNN, respectively. In terms of the localization network there were two types of architecture: SSD and Faster R-CNN. The SSD architecture adds auxiliary CNNs after the backbone network, while the Faster-RCNN uses a regional proposal network (RPN) for proposing regions in the images, which were further sent to the CNN, which was also used as a backbone. The classification layers of both face detectors were at the top of the architecture and used the softmax loss function together with L2 normalization in order to adjust the localization of the face region and to classify the image.

Inception Based SSD Model

The SSD architecture uses an Inception V2 pretrained model as a backbone. The Inception model was designed for solving high variation in the location of information, thus is useful for localization and object detection when is added as base network for SSD and RPN architectures. Different types of kernels with multiple sizes (7 × 7, 5 × 5 and 3 × 3) were used.

Larger kernels can look for the information that is distributed more globally, as the smaller one search in the information that is not as sparse. There is one important filter that is also used for this type of CNN, which is the 1 × 1 convolution for reducing or increasing the number of feature maps. The network is 22 layers deep when counting only layers with parameters (or 27 layers with pooling). The overall number of layers (independent building blocks) used for the construction of the network is about 100. The SSD architecture adds auxiliary structure to the network to produce multiscale feature maps and convolutional predictors for detection. At prediction time, the network generates scores for the presence of each object category in each default box and produces adjustments to the box to better match the object shape. Additionally, the network combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes

Pipeline for Deep Face Detection CNN

Transfer learning and fine-tuning has been used for training SSD and Faster R-CNN models pretrained on COCO and Open Image databases. The input for the fine-tuned training consisted of 2800 face images randomly selected from the Open ImageV4dataset images . The dataset was split into three categories: 2000 images for training, 400 for validation and 400 for testing.

The hyperparameters for training each network are described below. The values were obtained under experiments and represent the best configuration in terms of accuracy, generalization and inference speed obtained after training each model. Due to the limitations of the graphical process unit used for training the number of operations was reduced for each training epoch, which implied a low number of images per batch, especially for Faster R-CNN model.

Facial Emotion Recognition CNN

The database used for training FER CNN is a selection of uncontrolled images from FER2013 and laboratory controlled images from CK+, JaFFE and KDEF . This dataset was divided in 24,336 training, 6957 validation and 3479 test sets. The labelling of the training and testing dataset were previously made by the authors of the database and, in addition, were verified by our team in order to avoid bias. A small amount of the images that did not meet our criteria in terms of class annotations and distinctiveness were dropped or moved to the corresponding class. In addition, data augmentation was used during training in order to increase the generalization of the model. A series of rotations, zooming, width and height shifting, shearing, horizontal flipping and filling were applied to the training dataset. The FER2013 images represents the majority of the dataset, around 90% for each emotion.

The facial expressions were divided in seven classes: angry, disgust, fear, happy, neutral, sad and surprise. Training classes distribution over the dataset was angry 10%, disgust 2%, fear

3%, happy 26%, neutral 35%, sad 11% and surprise 13%. The validation and test set followed the same distribution.
The CNN models used for FER were pretrained on the ImageNet database and could recognize objects from 1000 classes. We did not need a SSD or RPN architecture as the face localization was already achieved with face detection CNN. ImageNet did not provide a class related to face or humans but there were some other classes (e.g., t-shirt or bowtie) that helped the network to extract these kinds of features during the prelearning process. This is related to the fact that the network needs these features in order to classify emotions related to classes. Taking into account that only some bottleneck layers will be trained, we would use transfer learning and fine-tuning, as follows: in the first phase the fully connected layers of the pretrained model were replaced with two new randomly initialized FC layers that were able to classify the input images according to our dataset and classes. During this warm-up training all the convolutional layers were frozen allowing the gradient to back-propagate only through the new FC layers. In the second stage the last layers of the convolutional networks were unfrozen, where high-level representations were learned allowing the gradient to back-propagate through these layers but with a very small learning rate in order to allow small changes to the weights.

Data-Base description

Cohn–Kanade [CK+]

The CK+ dataset consists of 593 video sequences from 123 participants. Each sequence contains images beginning from onset (neutral frame) and progressing to the peak expression (last frame). The label associated with each sequence is depicted from the peak expression. The dataset contains images for seven different expressions: anger, contempt, fear, disgust, happiness, surprise, and sadness. The images have a resolution of 640 × 480 pixels. In this work, the images are cropped into 48 × 48 pixels to focus on the subject face.

Japanese Female Facial Expression (JAFFE)

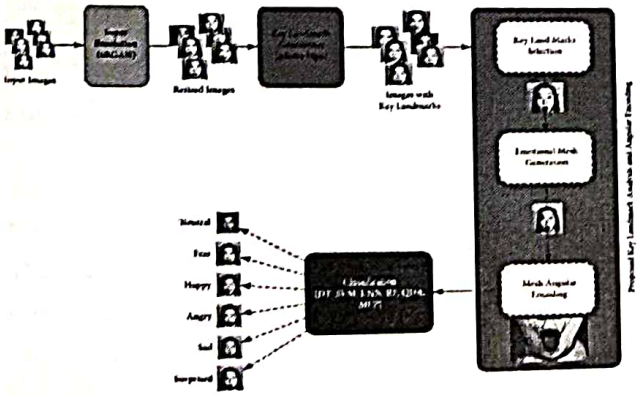
The JAFFE dataset has 213 photos of ten different female actors posing for seven different facial expressions. There are six primary expressions: happiness, sadness, surprise, anger, disgust, and fear, plus one neutral expression. The images have a resolution of 256 × 256 pixels.

Real-World Affective Face Database (RAF-DB)

RAF-DB contains 15,339 facial images with uncontrolled poses and illumination from thousands of individuals of different ages and races. The images within the RAF-DB are labelled by approximately 40 annotators. The database includes six basic expressions plus a neutral expression.

3.2 Architecture/Framework:

This module contains three subphases (key landmark selection, emotional mesh generation, and mesh angular encoding). The main idea of this module is to generate an emotional mesh that connects the selected key landmarks. Hence, the obtained mesh is encoded into angular values to generate a feature map. Moreover, the generated feature map is enrolled into a classifier to be discriminated into six categories.



Pre-processing

Generally, the images that are captured by robotic vision devices have a limited resolution due to the hardware limitations of cameras involved in such systems. Furthermore, most of the available datasets for human emotion recognition are down-sized because of the storage limitations. Therefore, the first module in the proposed approach is the super-resolution. In addition, the proposed approach involves angular feature extraction from the geometry of the face images, which requires a clarified representation of the landmarks and boundaries of the face images to allow proper facial emotion recognition. SRGAN , a Generative Adversarial Network (GAN) for image Super-Resolution (SR), is employed in the current research to increase the perceptual quality of images prior to further processes. With SRGAN, the images are super-resolved with a 4x upscaling factor, while minimizing the Mean SquareError (MSE) between the super-resolved and original images and maximizing the Peak Signal-to-Noise Ratio (PSNR).

Figure 5 illustrates the preprocessing step by employing the SRGAN. The figure displays an original image selected from the CK+ dataset and the corresponding super-resolved image after SRGAN. The original image size is 48×48 pixels, and the super-resolved image size is 192×192 pixels.

Key Landmark Generation

The process of key landmark generation is performed using deep MediaPipe technique. MediaPipe is an open-source ML framework developed by Google and devoted to building real-life computer vision applications. MediaPipe capabilities allow developers to focus on algorithm or model development, while using MediaPipe to iteratively improve their application with results that are consistent across different devices and platforms . Solutions that are currently implemented with MediaPipe include face detection, face mesh annotation, iris localization, hand detection, pose estimation, hair segmentation, object detection and tracking, and 3D object detection (Objectron). These solutions are released in different platforms: mobile (Android and IOS), C++, Python, and JS.

Proposed Key Landmark Analysis and Angular Encoding

This project report presents a key landmark analysis and an angular encoding module. This module contains three subphases (key landmark selection, emotional face mesh generation, and mesh angular encoding). The main idea of this module is to generate an emotional mesh, which connects the selected key landmarks. Hence, the obtained mesh is encoded into angular values to generate a feature map. In the following subsections, a discussion for each step in this module is presented.

Key Landmark Selection

As discussed earlier, the MediaPipe face mesh solution provides face detection capability and 468 facial landmarks spread over the face, along with their locations (x and y coordinates for each detected landmark). In the proposed model, only 27 key landmarks are selected from the 468 detected landmarks. These key landmarks are used later to define the vertices of the emotion face mesh. s. The selection of the key landmarks and their locations is based on the Facial Action Coding System (FACS) , which encodes movements of individual facial muscles. It can be used to describe facial actions that make up an expression based on changes in facial muscles regardless of emotion. The movement of particular facial muscles, known as Action Units (AUs), is encoded by FACS. This requires unique instantaneous changes in facial appearance . Table 2 describes the facial emotion-related AUs and the corresponding FACS names.. Each key landmark location is chosen such that it is more probably affected by a specific emotion-related AU, which seeks better recognition of facial expressions.

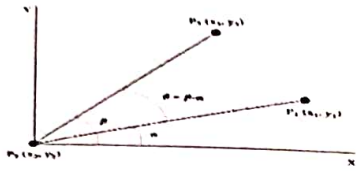
Emotional Mesh Generation

After selection of the key landmarks, emotion face mesh is created, consisting of 27 vertices inferred from the selected key landmarks. Edges of the emotion face mesh, which define the connections between vertices, are drawn to establish a closed mesh structureThe mesh yields 27 vertices and 38 edges. Deformation of emotion face mesh measured by the deviation of angles between edges reflects facial muscle contraction and relaxation, which will be used to identify facial emotions.

Mesh Angular Encoding

After acquiring the key landmarks and establishing the emotion face mesh, we use the mesh to extract the relevant features for emotion classification. The relevant features employed are geometric features, since most emotions can be detected from geometric changes. Ten features are extracted, defining angles between specific edges of the emotion face mesh. The angles are represented in degrees in the range of (0°, 360°). These features are then fed to the ML classifiers to learn from them to identify each emotion. The low dimensionality of features (10 features) makes them more resistant to local facial changes. In addition, the classifiers can be trained in a much shorter time. Moreover, the overall complexity of the proposed framework is significantly reduced.

The angle between the three vertices can be computed as follows



The angle θ between the line (edge) connecting P_2 and P_3 and the line (edge) connecting the points P_2 and P_1 is unknown.

The angle β between the line P_2-P_3 and the X-axis can be computed as

$$\beta = \tan^{-1}\left(\frac{y_3 - y_2}{x_3 - x_2}\right). \quad (1)$$

Similarly, the angle α between the line P_2-P_1 and the X-axis can be computed as

$$\alpha = \tan^{-1}\left(\frac{y_1 - y_2}{x_1 - x_2}\right). \quad (2)$$

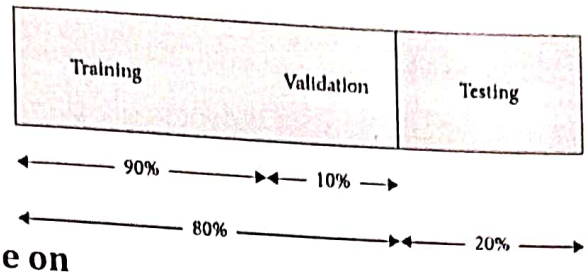
Hence, the angle θ will be

$$\theta = \beta - \alpha = \tan^{-1}\left(\frac{y_3 - y_2}{x_3 - x_2}\right) - \tan^{-1}\left(\frac{y_1 - y_2}{x_1 - x_2}\right). \quad (3)$$

Using the above procedure, ten angles between prescribed edges in the emotion face mesh are computed, and then used for classification. Angle values are all positive, where negative values can be avoided by adding 360° to the values. Furthermore, the generated feature maps are redistributed using PCA to enhance their distribution.

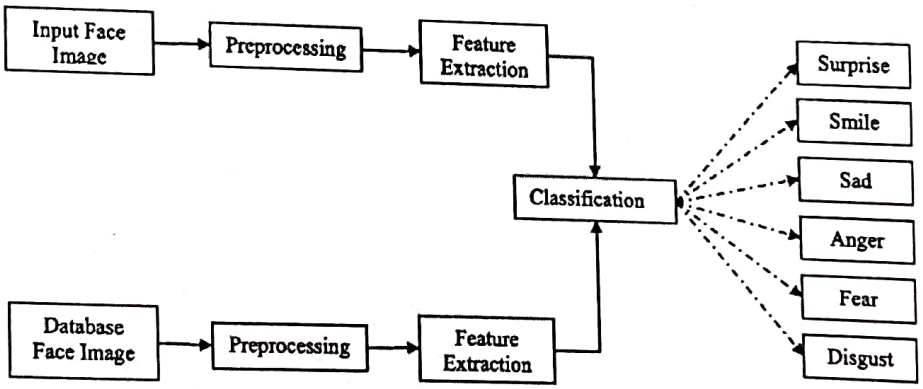
Classification

In this work, we develop an automated facial expression identifier to recognize human emotions for robotic vision applications. Discriminant features extracted from a face are fed to classifiers to recognize the emotion in the given face. DT, KNN, a multiclass SVM , Gaussian NB, MLP with backpropagation, QDA, RF, and LR classifiers are used for classification. The trial-and-error method and grid-search are conducted to identify the optimal structure and hyperparameters of classifiers. In addition, 10-fold cross-validation is employed to estimate the optimal hyperparameter combinations to avoid overfitting. The images in the dataset are divided into two parts: training part and testing part. The training part is used to train/validate the classifier, and the testing part is used to test the performance of the classifier. The 10-fold cross-validation adopted in the current model employs further splitting of the training part into ten folds (subsets). After that, nine folds are used to train the classifier, while the remaining fold is used to validate the training. This process continues until each of the ten folds is used exactly once for validation. The optimal configurations identified in the training stage are then applied in the testing stage.

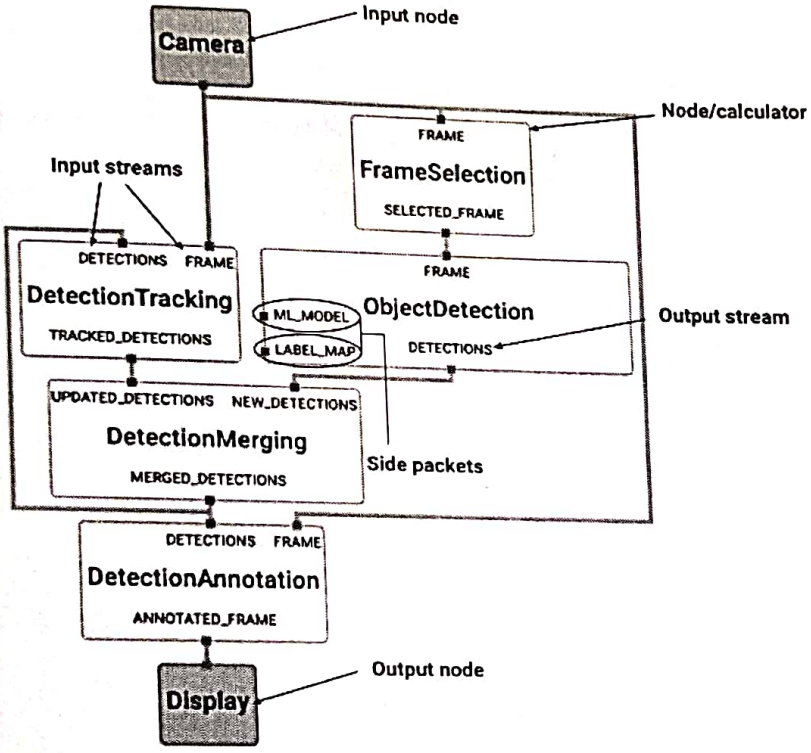


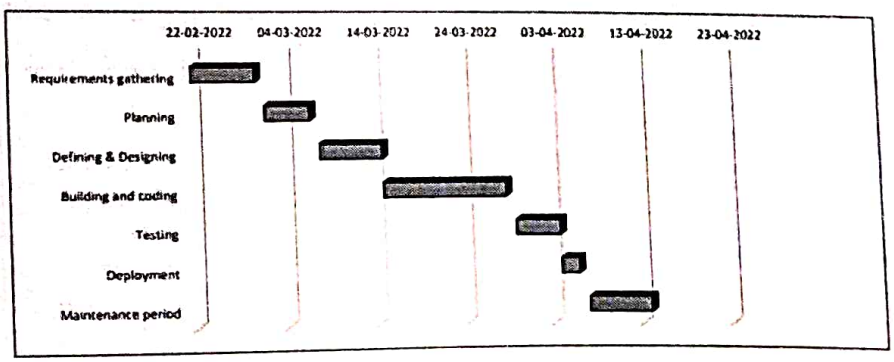
Detailed example on
https://www.hindawi.com/journals/cin/2022/8032673/

3.3 Algorithm & Process Design: Algorithm



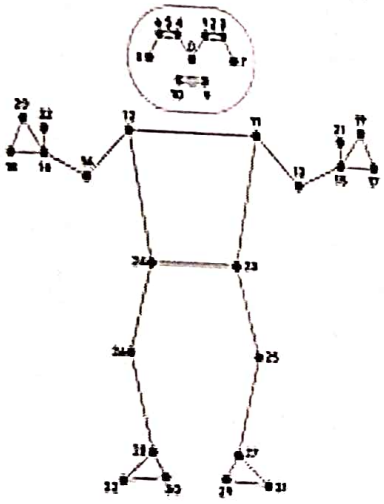
Mediapipe Algorithm





Grant Chart

Mediapipe basic Landmarking



0. nose
1. left_eye_inner
2. left_eye
3. left_eye_outer
4. right_eye_inner
5. right_eye
6. right_eye_outer
7. left_ear
8. right_ear
9. mouth_left
10. mouth_right
11. left_shoulder
12. right_shoulder
13. left_elbow
14. right_elbow
15. left_wrist
16. right_wrist
17. left_pinky
18. right_pinky
19. left_index
20. right_index
21. left_thumb
22. right_thumb
23. left_hip
24. right_hip
25. left_knee
26. right_knee
27. left_ankle
28. right_ankle
29. left_heel
30. right_heel
31. left_foot_index
32. right_foot_index

3.4 Details of Hardware & Software:

PYTHON:

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Its language constructs and objectoriented approach aim to help programmers write clear, logical code for small- and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

Python's large standard library, commonly cited as one of its greatest strengths, provides tools suited to many tasks. For Internet-facing applications, many standard formats and protocols such as MIME and HTTP are supported. It includes modules for creating graphical user interfaces, connecting to relational databases, generating pseudorandom numbers, arithmetic with arbitrary-precision decimals, manipulating regular expressions, and unit testing.

Python is meant to be an easily readable language. Its formatting is visually uncluttered, and often uses English keywords where other languages use punctuation. Unlike many other languages, it does not use curly brackets to delimit blocks, and semicolons after statements are allowed but rarely used. It has fewer syntactic exceptions and special cases than C or Pascal.

Python consistently ranks as one of the most popular programming languages.

JUPYTER NOTEBOOK:

Jupyter Notebook (formerly IPython Notebooks) is a web-based interactive computational environment for creating notebook documents.

A Jupyter Notebook document is a browser-based REPL containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media. Underneath the interface, a notebook is a JSON document, following a versioned schema, usually ending with the ".ipynb" extension.
Jupyter Notebook can connect to many kernels to allow programming in different languages. A Jupyter kernel is a program responsible for handling various types of requests (code execution, code completions, inspection), and providing a reply. Kernels talk to the other components of Jupyter using ZeroMQ, and thus can be on the same or remote machines. Unlike many other Notebook-like interfaces, in Jupyter, kernels are not aware that they are attached to a specific document, and can be connected to many clients at once. Usually kernels allow execution of only a single language, but there are a couple of exceptions.[citation needed] By default Jupyter Notebook ships with the IPython kernel.

A Jupyter Notebook can be converted to a number of open standard output formats (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python) through "Download As" in the web interface, via the nbconvert library or "jupyter nbconvert" command line interface in a shell. To simplify visualisation of Jupyter notebook documents on the web, the nbconvert library is provided as a service through NbViewer which can take a URL to any publicly available notebook document, convert it to HTML on the fly and display it to the user.

OpenCV:

OpenCV (Open Source Computer Vision Library) is an open source computer vision and machine learning software library. OpenCV was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in the

February of 2010. In addition to offering code analysis, PyCharm features: A graphical debugger

- An integrated unit tester
- Integration support for version control systems (VCSs)
- Support for data science with Anaconda

The main reason Pycharm for the creation of this IDE was for Python programming, and to operate across multiple platforms like Windows, Linux, and macOS. The IDE comprises code analysis tools, debugger, testing tools, and also version control options. It also assists developers in building Python plugins with the help of various APIs available. The IDE allows us to work with several databases directly without getting it integrated with other tools. Although it is specially designed for Python, HTML, CSS, and Javascript files can also be created with this IDE. It also comes with a beautiful user interface that can be customized according to the needs using plugins.

Numpy:

NumPy is a Python library used for working with arrays.It also has functions for working in domain of linear algebra, fourier transform, and matrices.NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.NumPy stands for Numerical Python.

In Python we have lists that serve the purpose of arrays, but they are slow to process.NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.Arrays are very frequently used in data science, where speed and resources are very important.

Mediapipe:

MediaPipe Face Detection is an ultrafast face detection solution that comes with 6 landmarks and multi-face support. It is based on BlazeFace, a lightweight and well-performing face detector tailored for mobile GPU inference. The detector's super-realtime performance enables it to be applied to any live viewfinder experience that requires an accurate facial region of interest as an input for other task-specific models, such as 3D facial keypoint estimation (e.g., MediaPipe Face Mesh), facial features or expression classification, and face region segmentation. BlazeFace uses a lightweight feature extraction network inspired by, but distinct from MobileNetV1/V2, a GPU-friendly anchor scheme modified from Single Shot MultiBox Detector (SSD), and an improved tie

resolution strategy alternative to non-maximum suppression. For more information about BlazeFace, please see the Resources section.
Naming style and availability may differ slightly across platforms/languages.

MODEL_SELECTION:An integer index `0` or `1`. Use `0` to select a short-range model that works best for faces within 2 meters from the camera, and `1` for a full-range model best for faces within 5 meters. For the full-range option, a sparse model is used for its improved inference speed. Please refer to the model cards for details. Default to `0` if not specified.

MIN_DETECTION_CONFIDENCE: *Minimum confidence value ([0.0, 1.0]) from the face detection model for the detection to be considered successful. Default to `0.5`.*

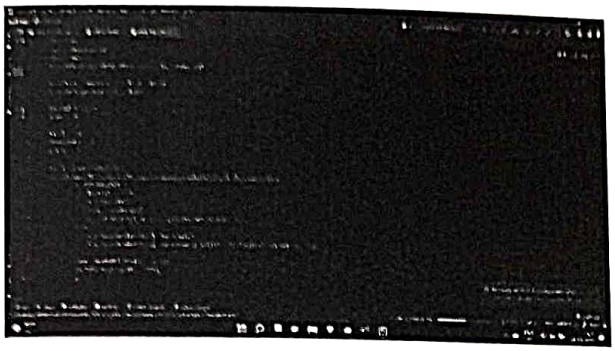
Live perception of simultaneous human pose, face landmarks, and hand tracking in real-time on mobile devices can enable various modern life applications: fitness and sport analysis, gesture control and sign language recognition, augmented reality try-on and effects. MediaPipe already offers fast and accurate, yet separate, solutions for these tasks. Combining them all in real-time into a semantically consistent end-to-end solution is a uniquely difficult problem requiring simultaneous inference of multiple, dependent neural networks. The MediaPipe Holistic pipeline integrates separate models for pose, face and hand components, each of which are optimized for their particular domain. However, because of their different specializations, the input to one component is not well-suited for the others. The pose estimation model, for example, takes a lower, fixed resolution video frame (256x256) as input. But if one were to crop the hand and face regions from that image to pass to their respective models, the image resolution would be too low for accurate articulation. Therefore, we designed MediaPipe Holistic as a multi-stage pipeline, which treats the different regions using a region appropriate image resolutionFirst, we estimate the human pose (top of Fig 2) with BlazePose's pose detector and subsequent landmark model. Then, using the inferred pose landmarks we derive three regions of interest (ROI) crops for each hand (2x) and the face, and employ a re-crop model to improve the ROI. We then crop the full-resolution input frame to these ROIs and apply task-specific face and hand models to estimate their corresponding landmarks. Finally, we merge all landmarks with those of the pose model to yield the full 540+ landmarks.

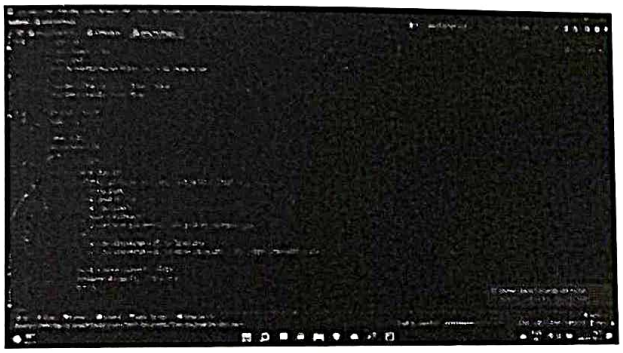
3.5 Experiment & Result:

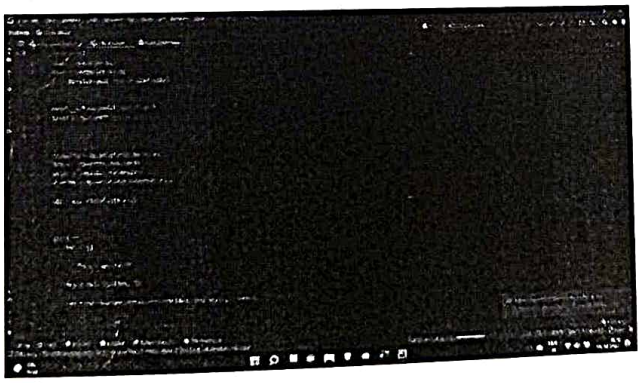
Experiment:

1. Collection File
2. Training File

3. Inference File

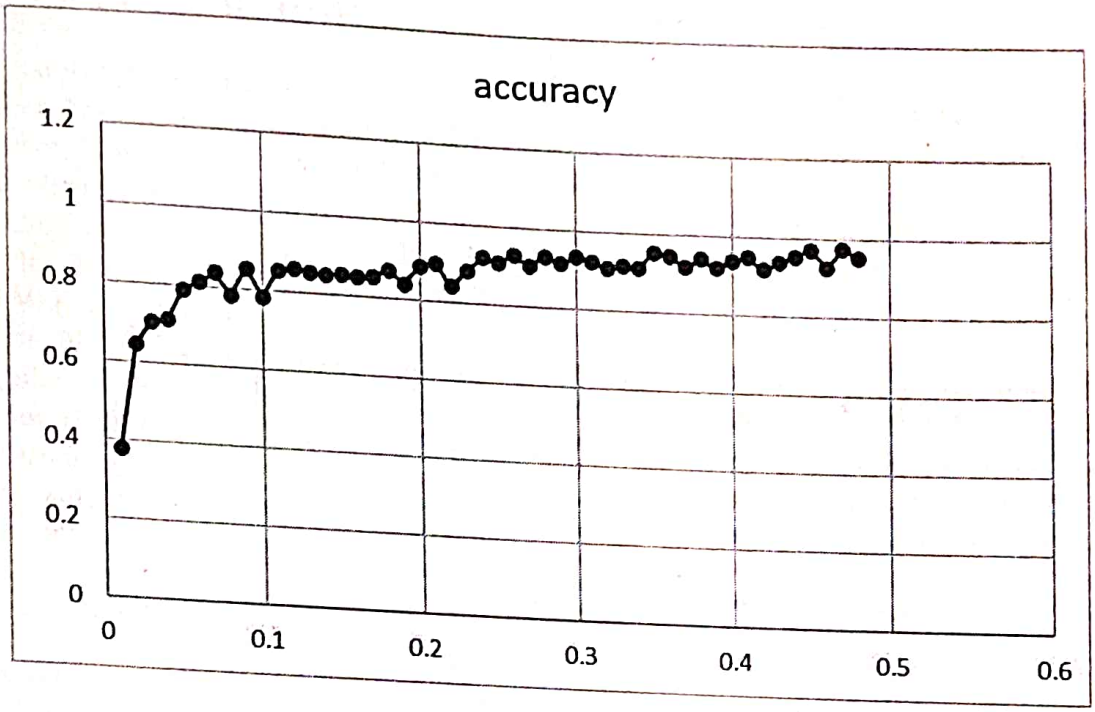






Result:

Accuracy:


accuracy
1.2
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6

3.7 Conclusion:

The issue of Human-Robot Interaction (HRI) has been discussed in this project. As a solution, the project presented a novel approach for facial expression recognition. This proposed approach consists of four phases, which are carried out to extract key points from facial images using a real-time algorithm (MediaPipe). Furthermore, these key points are enrolled into a sequence of selection, mesh generator, and angular encoding modules. Moreover, the generated feature maps are classified using several classification algorithms, including SVM, KNN, RF, QDA, NB, LR, DT, and MLP. The novelty of the proposed approach is highlighted in the proposed key point analysis and angular encoding algorithm. This algorithm is efficient, because it generates only ten features (angular values), which are discriminative for different emotional classification categories. The proposed approach has been evaluated on CK+, JAFEE, and RAF-DB datasets. It reveals a superior performance in terms of accuracy of detection and processing time evaluation metrics. Furthermore, the low dimensionality of extracted features enables the ML-based approaches to reach an optimum performance in a short time with much lower computational cost than those of the DL-based approaches, which require more time for convergence and need much computational cost.

In addition, the future work that can be deduced from this Project is introducing a method for emotion detection from other modalities such as videos, spoken words, and written text. Furthermore, hardware implementation of the proposed approach is a research trend, which we are working on. Moreover, further machine learning techniques such as dictionary learning and semi-supervised learning can be performed to solve this issue

Reference

- **https://www.hindawi.com/journals/cin/2022/8032673/**
- https://www.koreascience.or.kr/article/JAKO20213155946 4400.page
- https://ieeexplore.ieee.org/abstract/document/9460142/references#references

- https://google.github.io/mediapipe/solutions/face_detection.html
- https://aip.scitation.org/doi/pdf/10.1063/5.0042221
- https://www.sciencedirect.com/science/article/pii/S2468227620302039
- https://link.springer.com/article/10.1007/s42452-020-2234-1#article-info
- https://towardsdatascience.com/face-detection-recognition-and-emotion-detection-in-8-lines-of-code-b2ce32d4d5de
- https://www.researchgate.net/publication/349612830_Detection_of_Emotion_Intensity_Using_Face_Recognition
- https://www.frontiersin.org/articles/10.3389/fpsyg.2021.759485/full
- https://youtu.be/ZxZSGRdTLtE
- https://youtu.be/He_oZ-MnIrU
- https://google.github.io/mediapipe/solutions/face_detection.html