INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

Project Report

Human Emotion Recognition from Facial Expressions and Body Gestures using CNN, Openpose and One Shot Learning

---

by

Anuj Kumar Singh - IIT2017015

Abhishek Sharma - IIT2017044

Under the supervision of

Prof. Anupam Agrawal

Candidate's Declaration

We hereby declare that the project work entitled "**Human Emotion Recognition from Facial Expressions and Body Gestures using CNN, Openpose and One Shot Learning**" submitted at Indian Institute of Information Technology, Allahabad, is the bonafide work of Anuj Kumar Singh (IIT2017015) and Abhishek Sharma (IIT2017044). It is a genuine record of our study carried out under the guidance of Prof. Anupam Agrawal. Due acknowledgements have been made in the text to all the materials used.

November 17, 2020 Prof. Anupam Agrawal

Contents

1. Introduction

Emotion is defined as a conscious mental reaction (such as anger or fear) subjectively experienced as a strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body. Emotion recognition is a complex task often proving out to be difficult for even the best of psychologists.

Emotion detection using facial and speech recognition is quite common a tool used these days and has several applications such as designing artificially intelligent systems that are mentally and physically equivalent to humans.

However not much has been done in the field of emotion detection using facial and body gestures. Our main goal of the project is to learn new computer science technologies and concepts while creating some practical applications.

We plan on developing a system to detect the emotion of an individual on the basis of his/her facial features and body gestures. Our system will support all three formats of recognition including those using images, videos and live webcams.

The applications of our project include identifying people in a crowd, monitoring citizens for suspicious behavior by tracking gender, age, identity and present emotional state. It can be used to preemptively stop criminals and potential terrorists and other surveillance techniques.

Currently present models either detect using Facial Expressions or using Body Gestures, the aim of our project is to combine the two approaches for improved accuracy and prediction.

2. Problem Definition and Objective

Humans have a variety of emotions, which can now be recognized by machines and computers thanks to advanced algorithms. The large number of emotions makes the **Human Emotion Recognition from Facial Expressions and Body Gestures using CNN, Openpose and One Shot Learning** a complex problem.

The problem we aim to solve involves identifying the emotion of an individual using their facial expressions and body language/gestures. In this project we will recognize the emotions using both the techniques individually at first and then combine the emotions in a way which gives the best suitable output considering both the individual outputs.

Our system will support all three formats of recognition including those using images, videos and live webcams. We have considered frontal faces and the future scope of the project involves supporting for those at certain angles.

3. Literature Review

Sr. No	Title of the paper	Year of Publication	Paper Description	Methodology Discussed	Datasets Used
1	Emotion Recognition using Deep Convolutional Neural Networks[1]	2016	Authors: V.Enrique Correa, Arnoud Jonker, Michael Ozo, Rob Stolk Year of Conference: 2016.	This Paper is predicting the emotions using the trained CNN model from the given list of emotions (like Angry, Disgusted, Fearful, Happy,Sad,Surprised, Neutral).	1. Facial Expression Recognition Challenge (FERC-2013) 2.Cohn Kanade (CK+) 3. Radboud Face Database (RaFD).
2	Emotion Recognition based on Multi-view Body Gestures[2]	2019	Authors: Zhijuan Shen, Jun Cheng, Xiping Hu, Qian Dong.Z Date of Conference: 22-25 Sept. 2019.	According to this paper, emotional body gestures are more closely related to the velocity and acceleration of key joints.	Datasets namely FABO, GEMEP, HUMAINE are being used in this paper.
3	Sitting Posture Recognition Based on OpenPose[3]	2019	Authors: Kehan Chen Year of Conference: 2019.	This paper helps identify the correct sitting posture using parts and pairs detected using Openshot. Based on this CNN is used to classify the posture as correct/incorrect.	Self made dataset made by using openpose library.

4	Siamese Neural Networks for One-shot Image Recognition[4]	2015	Authors: Gregory Koch, Richard Zemel, Ruslan Salakhutdinov. Presented at ICML 2015 Deep Learning Workshop.	1.Paper compares similarity between different languages. 2.Two identical neural networks are used for this.	This paper used the Omniglot data set.
5	Emotion Recognition Using Feature-level Fusion of Facial Expressions and Body Gestures[5]	2019	Authors: Tanya Keshri, Suja Palaniswamy Presented at ICCES 2019, IEEE Conference Record # 45898;	The basic steps in emotion recognition are - 1. image pre-processing 2. feature extraction & selection 3. classification.	In this paper dataset used is Amrita Emotion Database(AED)-2.

1. Emotion Recognition using Deep Convolutional Neural Networks [1]

1.1. In this paper the following idea has been implemented:

1. Input image is passed to the Input layer.

2. Used FERC-2013 dataset as it has 20000 approx. pictures

3. This layer is followed by one convolutional layer and a max pooling layer resp.

4. The network is finished with two more convolutional layers and one fully connected layer, connected to a soft-max output layer. This led to the accuracy as 63%.

Fig. 1: Emotion recognition using deep convolutional neural networks, retrieved from https://datarepository.wolframcloud.com/resources/FER-2013

The base paper suggests 3 datasets namely Facial Expression Recognition Challenge (FERC-2013), Extended CohnKanade (CK+), and Radboud Faces Database (RaFD)[1]. These datasets mainly differ in the number of images, their quality and cleanliness. Fig. 1 represents the data set images.

1.2. Merits

1. Have high accuracy percentages, both testing as well as validation accuracy.

2. Comparison between different datasets result in better results.

1.3. Modifications

There are several modifications that we will do from our base paper such as:

1. The base paper project is only working on the pictures but we will modify it with the use of a webcam and our project is giving outputs at realtime.

2. In the base paper only the emotion of one person in a picture is getting detected but we will modify it to detect all the faces in the picture, this will be helpful in group photos.

2. Emotion Recognition based on Multi-view Body Gestures [2]

2.1. In this paper, we introduce an exploratory experiment to recognize emotion using deep learning only from body gestures. 43,200 videos of simplified body gestures and their neutral control groups are captured from 80 humans using Hikvision network cameras to support the experiment.

According to this paper, emotional body gestures are more closely related to the velocity and acceleration of key joints, which is significantly different from the trajectories of limbs in action recognition.

In our approach, we use the publicly available OpenPose toolbox to estimate the location of the joints first.

Paper suggested works on various datasets namely FABO, GEMEP, HUMAINE[2] which were based on various factors like naive geometrical representations, such as orientation of hands, displacements,velocity and acceleration which was created on their own. Fig. 2 represents how actions are related to emotions.

Fig. 2: Emotional recognition using multi view body gestures [2]

2.2. Merits

1. In this paper linked different different body gestures with the emotions like

i) jump => happy

ii) squat => sadness

iii) throw => anger

iv) stand => surprise

v) recede => fear

vi) turn and walk away => disgust

So in this way body gestures can be linked with emotions and can predict 1 person emotion according to this.

2.3. Modifications

1. We can use One Shot Learning Techniques for improvements

One Shot learning is a classification task where one example (or a very small number of examples) is given for each class, that is used to prepare a model, that in turn must make predictions about many unknown examples in the future.In the case of one-shot learning, a single exemplar of an object class is presented to the algorithm.

2. We can use Siamese Neural Networks for One-shot Image Recognition. A Siamese network is an architecture with two parallel neural networks, each taking a different input, and whose outputs are combined to provide some prediction.

3. Sitting Posture Recognition Based on OpenPose [3]

3.1. In this paper, it is predicting whether the posture is the correct sitting posture or not. It is done by CNN. This paper used openpose for making a dataset and then CNN model.

CNN model is composed of 19 layers, the input layer takes the images as input. The output layer is the Classification layer, using the SoftMax Classifier, which tells whether the posture of the human in the image is proper or not.

It is done in this way :

1. Start with an image.

2. Labelling the correct sitting postures as '0' else with '1'.

3. Used CNN model which consisted of 19 layers. This model is being trained for 100 epochs and achieves an accuracy rate of 90%.

3.2. Merits

This paper is using the skeleton view for predicting whether that is the sitting posture or not. This is based on openpose (OpenPose is an open-source real-time system for multi-person 2D pose detection, including body, foot, hand, and facial key-points).

3.3. Modifications

This paper is implemented for sitting postures but we will extend this approach for finding the emotions using the body gesture as mentioned above(i.e - for 7 different emotions which are sad, happy, angry, neutral, surprised, disgusted and fearful).

4. Siamese Neural Networks for One-shot Image Recognition [4]

4.1. A Siamese network is an architecture with two parallel neural networks, each taking a different input, and whose outputs are combined to provide some prediction.

Two identical networks are used, one taking the known signature for the person, and another taking a candidate signature. The outputs of both networks are combined and scored to indicate whether the candidate signature is real or a forgery. This paper used the Omniglot data set.

The Siamese Network is interesting for its approach to solving one-shot learning by learning feature representations (feature vectors) that are then compared for verification tasks. Fig. 3 represents the Siamese Neural Network diagram along with its basic working.

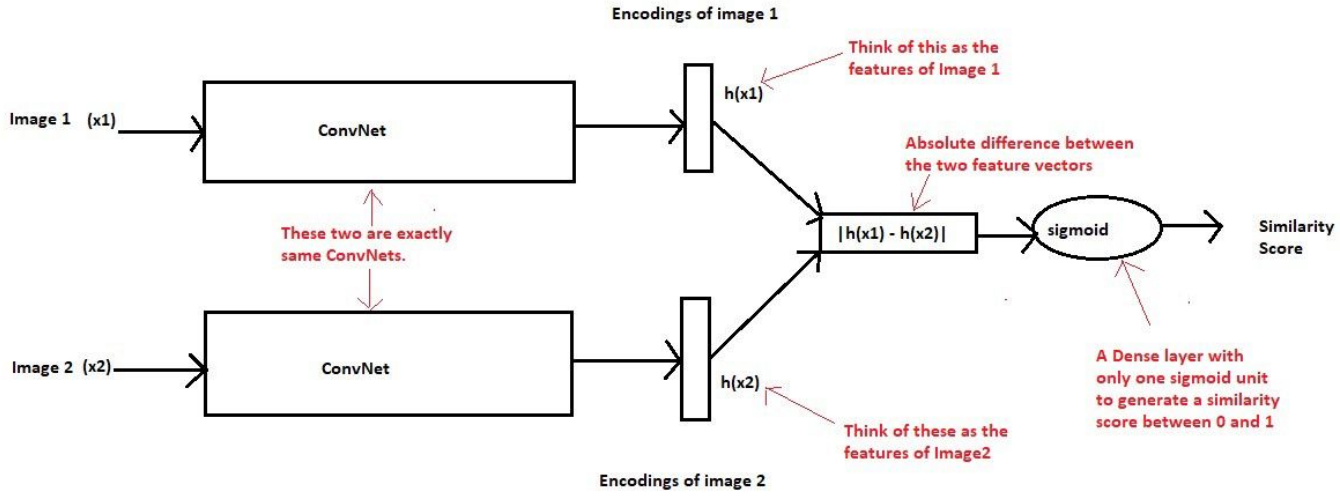


Fig. 3: Siamese Neural network for one shot image recognition, retrieved from https://towardsdatascience.com/, January 2019

4.2. Merits

The matching accuracy is very high in this model and outputs more accurately. Test accuracy is around 70% for this model. Accuracy is calculated using n-way one shot learning. It also doesn't require a very large training data set like other similarity classifiers.

4.3. Modifications

We can use the siamese neural network for body gesture emotion recognition as it gives more accuracy . In this paper it is used for matching languages but in our project we can use it for matching them with the different body gestures skeleton view obtained using open-pose and based on it we can predict the emotion.

5. Emotion Recognition Using Feature-level Fusion of Facial Expressions and Body Gestures [5]

5.1. Most of the work done so far is based on single modality. Our objective is to efficiently integrate emotions recognized from facial expressions and upper body pose of humans using images. Our work on bimodal emotion recognition provides the benefits of the accuracy of both the modalities.

The basic steps in emotion recognition are image preprocessing, feature extraction & selection and classification. All these phases are discussed in detail in this section.

We have developed an algorithm multi SVM classifier for classification of seven emotions. Multi SVM is described as follows:

• Initially SVM is trained for dataset AED-2 for 7 different emotions using svmtrain. It will generate train data containing classes with their feature set.

• For each class i, where i<=n (where n =number of emotions, i.e. 7)

• Execute svmclassify. The extracted features are classified using the procedure described and results.

5.2. Merits

1. Have a high accuracy percentage for all the emotions.

2. Proposed method proves that when we combine two modalities it can achieve better accuracy.

5.3. Modifications

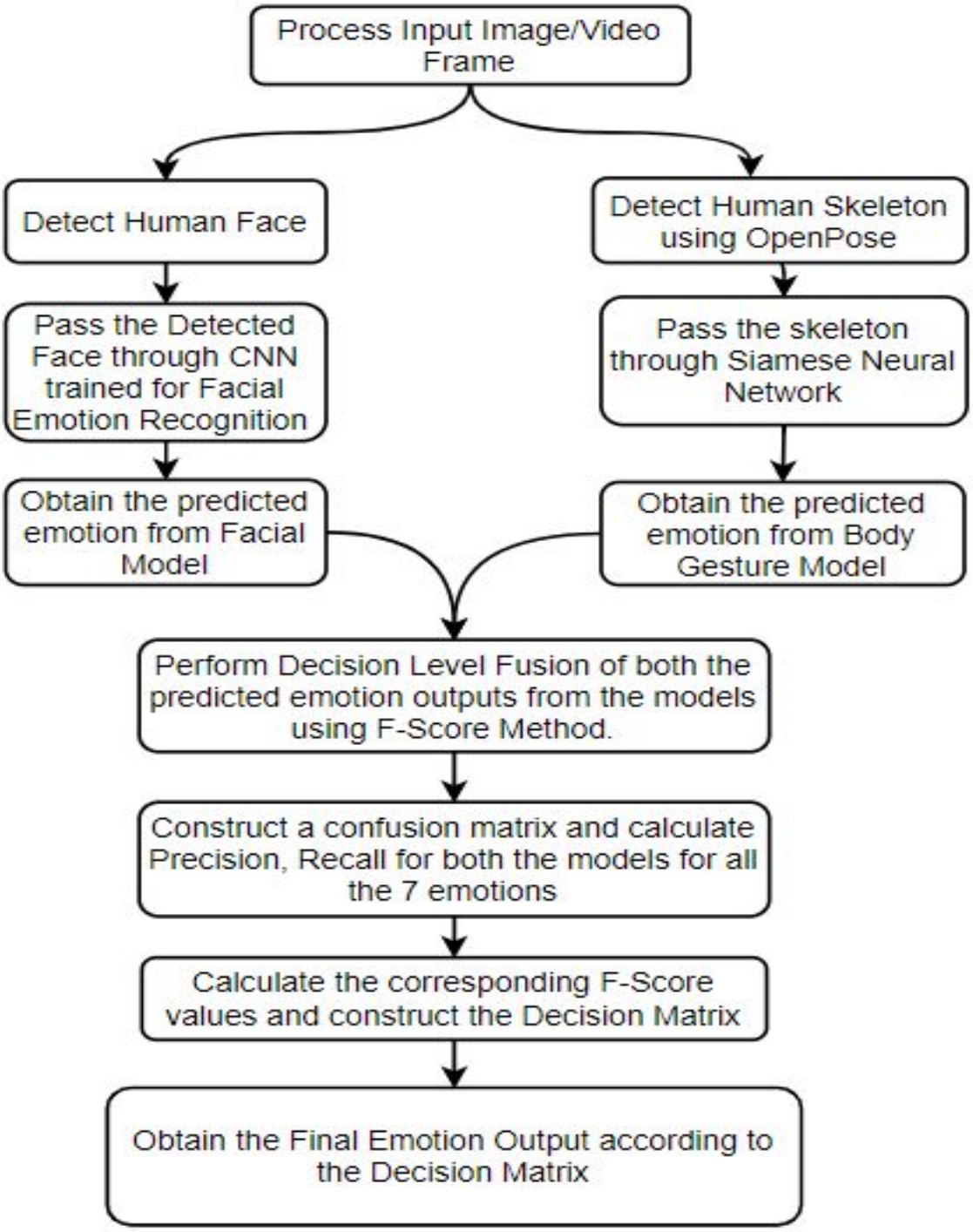
In future, this work can be extended using deep learning and for images with pose & illumination variations. More number of subjects will be added to the dataset.

4. Proposed Methodology

Method we are following involve following steps:
We use the methodology described in paper [1], i.e CNN to detect the emotion corresponding to the facial expressions of the person. We get the output vector denoting the probabilities of the emotion belonging to any of the 7 classes.

We use the methodology described in paper [2], [3] and first produce a skeleton of the human in consideration using Openpose. Using paper [4] we train our siamese model to identify the similarity between the two images that would be provided as input, one would be the emotions in our database and the second would be the image which is to be evaluated. We can construct a similarity vector corresponding to all 7 classes.

Based on the output vectors obtained using the two models we can combine both of our models using decision based fusion to get our final emotion output.


Process Input Image/Video Frame
Detect Human Face
Detect Human Skeleton using OpenPose
Pass the Detected Face through CNN trained for Facial Emotion Recognition
Pass the skeleton through Siamese Neural Network
Obtain the predicted emotion from Facial Model
Obtain the predicted emotion from Body Gesture Model
Perform Decision Level Fusion of both the predicted emotion outputs from the models using F-Score Method.
Construct a confusion matrix and calculate Precision, Recall for both the models for all the 7 emotions
Calculate the corresponding F-Score values and construct the Decision Matrix
Obtain the Final Emotion Output according to the Decision Matrix

4.1 Using Facial Expression

In the first phase we recognize the emotion using the facial expressions of the user in the frame.
We have implemented our program for working of all three modes i.e on images, videos, and web-cameras.

- **Training CNN (Convolutional Neural Network):** In this stage we train our CNN model from paper[1] on FER (Facial Emotion Recognition) dataset.

- **Detecting Frontal Faces:** We have to provide the path to our input image/video stream as an argument while running the program. Our program then does the following:
    - The first step involves the validation phase i.e it checks whether the image path is valid and an image exists at the given path.
    - If image exists then use haarcascades to detect frontal faces from the image
        - If no faces are detected then we don't output anything.
        - For each face we predict the emotions from the 7 classes of emotions that we have taken into consideration(Angry, Happy, Sad, Fearful, Surprised, Disgusted and Neutral).

- **Extracting Emotion from Facial Expressions:** Once we completed the face detection from input image/video stream then we pass all the detected faces to our CNN model to extract emotions of the faces using facial expressions.

4.2 Using Body Gestures

In this part we use Openpose and One Shot Learning.

4.2.1 Openpose

In the first step we use Openpose, which is a pre-trained model used to convert a body into its skeleton. Internally it consists of 2 CNN which parse the image or video frame provided as input parallely, the first CNN predicts where the points in a body are located, the second CNN predicts the degree of association i.e the probability of forming an edge between two points.

Openpose gives out the human skeleton corresponding to the input image. We use Openpose for two things, first is for converting the images of our dataset used for body gesture based emotion detection into a dataset consisting of the corresponding skeletons. The second thing is to use it for processing our input image and converting it to a skeleton which we pass further to the Siamese Neural Network.

4.2.2 One Shot Learning (Siamese Neural Network)

In the second step we use Siamese Neural Network, it consists of two CNN each of it taking one image each, first is the original image and second is the image corresponding to the emotion to which we need to find its similarity to. Both the CNN produce a feature vector, for the two images to have a greater similarity their feature vectors need to be similar too, therefore we find the absolute difference of the values of the two output vectors obtained from CNN1 and CNN2. We then pass it through a sigmoid function which generates a similarity value between 0 and 1, 0 denoting least similar and 1 denoting maximum similar. Finally, we compare and see for which class of emotion the similarity score is maximum and that class is our final output class for that emotion.

4.3 Decision Level Fusion

Decision level fusion is a form of fusion that combines the decisions of several models or classifiers into a single decision about the activity of an event that has occurred.

According to definition [12], F-score "In statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy."

It considers of 2 parameters:

Precision: p

Precision is the fraction of true positive examples among the examples that the model classified as positive. In other words, the number of true positives divided by the number of false positives plus true positives.

$$p = \frac{TP}{TP + FP}$$

Recall: r

Recall, also known as sensitivity, is the fraction of examples classified as positive, among the total number of positive examples. In other words, the number of true positives divided by the number of true positives plus false negatives.

$$r = \frac{TP}{TP + FN}$$

where ,

TP: True positive, equivalent to Hit i.e. the number of true positives classified by the model.

FP: False positive, equivalent to false error (Type 1 error) i.e. the number of false positives classified by the model.

FN: False negative, equivalent to miss (Type 2 error). i.e the number of false negatives classified by the model.

The formula for the standard F1-score is the harmonic mean of the precision and recall. A perfect model has an F-score of 1.

$$F_1 = \frac{2}{\frac{1}{\text{recall}} \times \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$
$$= \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

Confusion Matrix

Using the results of each classifier we develop a confusion matrix. A confusion matrix is a table classifying each of the observations in a tabular format.

In our case we have 7 classes of emotions corresponding to which we will be having a 7 * 7 matrix with all the emotions across the rows and columns as actual and predicted respectively. Each cell in this matrix of ours gives the number of observations for each predicted class to actual class. Using the values from the confusion matrix as input we calculate recall and precision for all of our 7 classes.

Once precision and recall have been calculated we compute the f-values/scores for all the 7 classes.

Decision Matrix

The decision matrix is also an 7 * 7 matrix with the ith row and jth cell represented as d(i,j). Suppose C1, C2 represent our first and second classifiers respectively, Ei, Ej represent the emotion detected by the classifiers C1 and C2. We define D(i,j) as

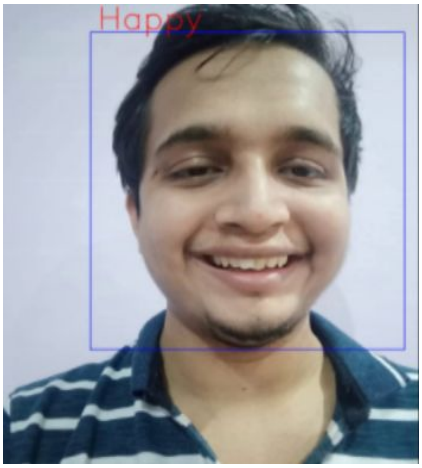
$$D(i,j) = \begin{cases} E_i & \textit{if } \textbf{F-score}(C_1, E_i) \geq \textbf{F-score}(C_2, E_j) \\ E_j & \textit{otherwise} \end{cases}$$

Once we have the Decision matrix, we can predict the final i.e. the bimodal output.

5. Results

We have obtained the below results for our bimodal emotion recognition system.

5.1 For Facial Emotion Detection



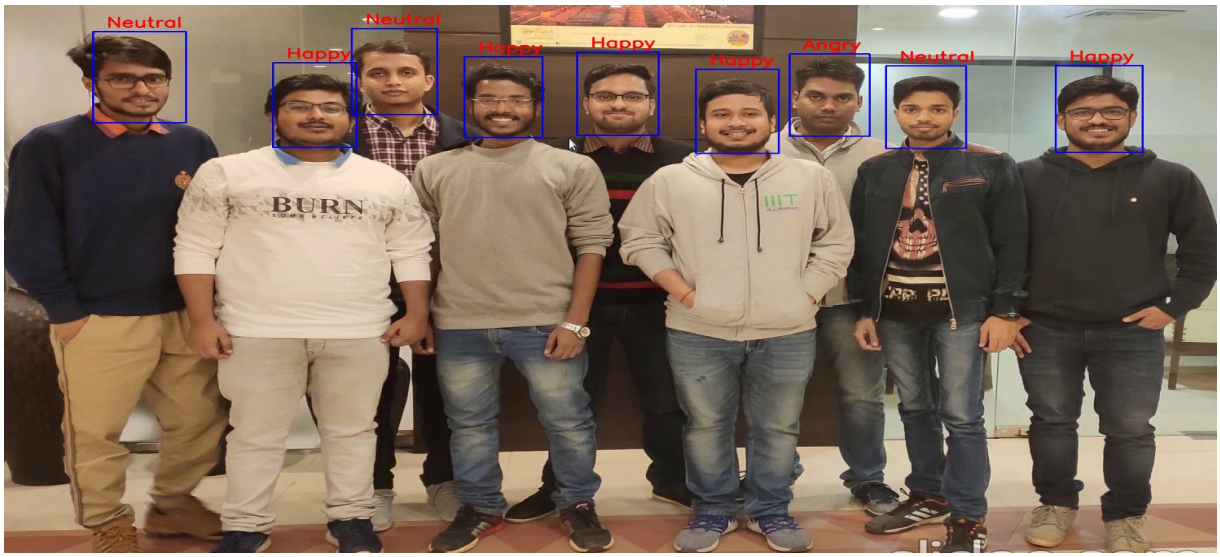


Fig. 4: Facial emotion recognition using CNN

5.2 For Body Gesture Emotion Detection

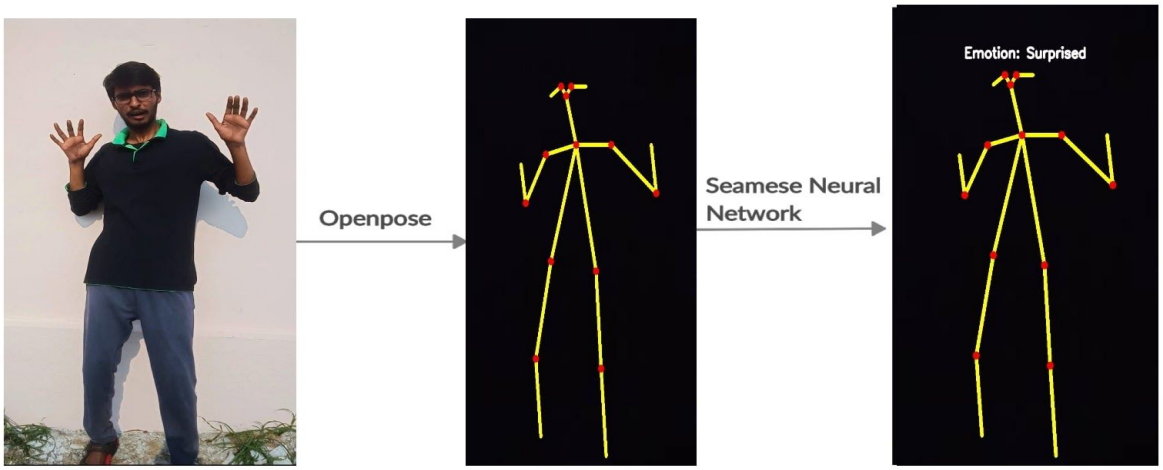


Fig.5: Body emotion recognition using Openpose and One Shot Learning

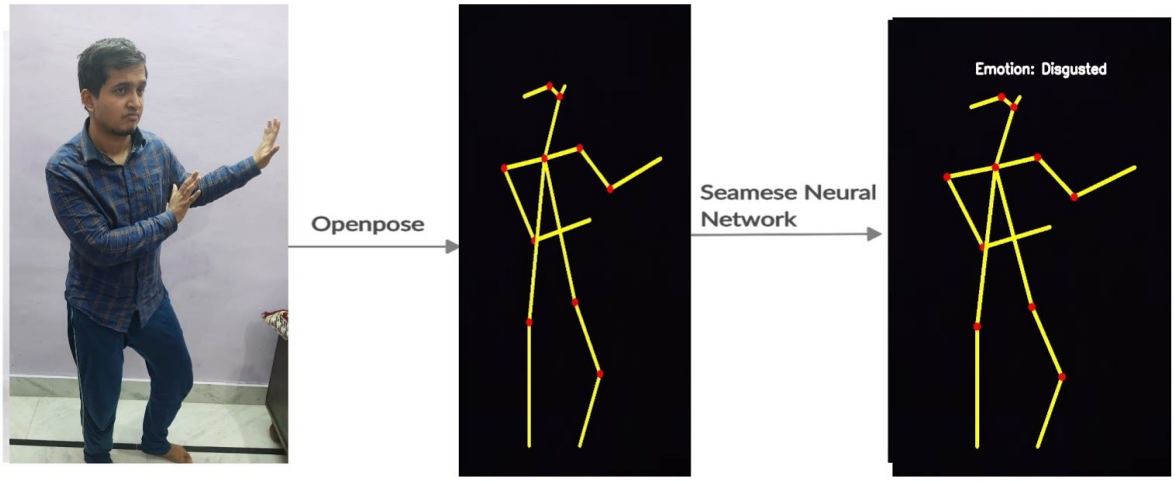


Fig.6: Body emotion recognition using Openpose and one shot learning

5.3 Bimodal detection

In table 1, C1 and C2 are the classifiers for body and face respectively. Following is the correspondence between E_i and emotion in table 1: E_1 = happy; E_2 = sad; E_3 = angry; E_4 = disgust; E_5 = fearful; E_6 = surprise; E_7 = neutral.The decision matrix in table 1 is now used to decide the emotion in a image file in case there are different results from classifiers for the corresponding face and gesture modality. It must be kept in mind however that a different decision matrix will be produced when the classifier is trained on a different dataset.

C1 / C2	E1	E2	E3	E4	E5	E6	E7
E1	E1	E1	E1	E1	E1	E6	E1
E2	E2	E2	E3	E4	E5	E2	E2
E3	E3	E3	E3	E4	E5	E3	E3
E4	E4	E2	E4	E4	E4	E4	E4
E5	E5	E5	E5	E5	E5	E5	E5
E6	E6	E6	E6	E4	E5	E6	E6
E7	E1	E2	E3	E4	E5	E6	E7

Table 1: Decision Matrix

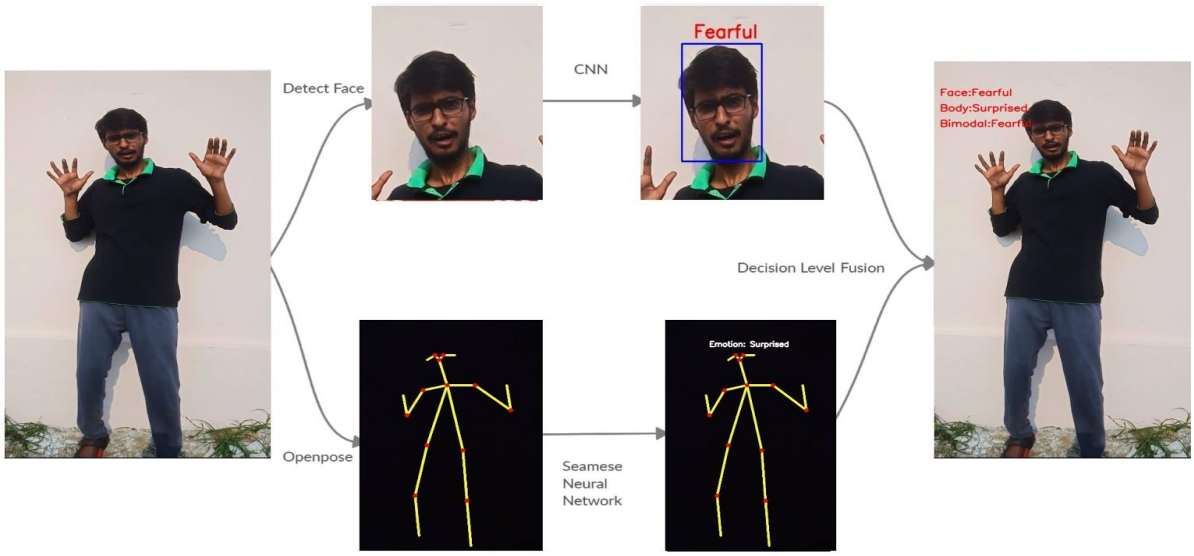


Fig.7: Human Emotion Recognition from Facial Expressions and Body Gestures using CNN, Openpose and One Shot Learning

The final recognition of emotion using the decision matrix of table 1 has been demonstrated in various cases:-

Case 1: Face incorrect, body correct and fusion correct

The approach used in [1] which uses facial expression shows the emotion as Angry(C2-E3). For the same image, using approach in [3] & [4] which used body gestures shows the emotion as Fearful(C1-E5). The fusion module decides the final emotion as fearful based on the F-score. Since the F-score value for Fearful(C1-E5) is higher according to the decision matrix than the F-score value for Angry(C2-E3), the fusion result gives the final emotion as Fearful which is the actual emotion. Thus it is shown that the accuracy of the fusion module is better than facial emotion and body gesture considered separately.

Fig. 8: Face incorrect, body correct result after fusion of two models

Case 2: Face correct, body incorrect and fusion correct

The approach used in [1] which uses facial expression shows the emotion as Fearful(C2-E5). For the same image, using approach in [3] & [4] which used body gestures shows the emotion as Surprised(C1-E6). The fusion module decides the final emotion as Fearful based on the F-score. Since the F-score value for Fearful(C2-E5) is higher according to the decision matrix than the F-score value for Surprised(C1-E6), the fusion result gives the final emotion as Surprised which is the actual emotion. Thus it is shown that the accuracy of the fusion module is better than facial emotion and body gesture considered separately.

Fig. 9: Face correct, body incorrect result after fusion of two models

6. Dataset Description

For facial expression based emotion recognition we use the FERC-2013[1] data set (from Kaggle) because of its large sample of images i.e approximately 32000 images. It contains images of 7 different emotions i.e - angry, disgustful, fear, happy, sad, surprised, neutral.

For body gesture based emotion recognition we will be training our siamese model[4] using our self generated data set of emotions.

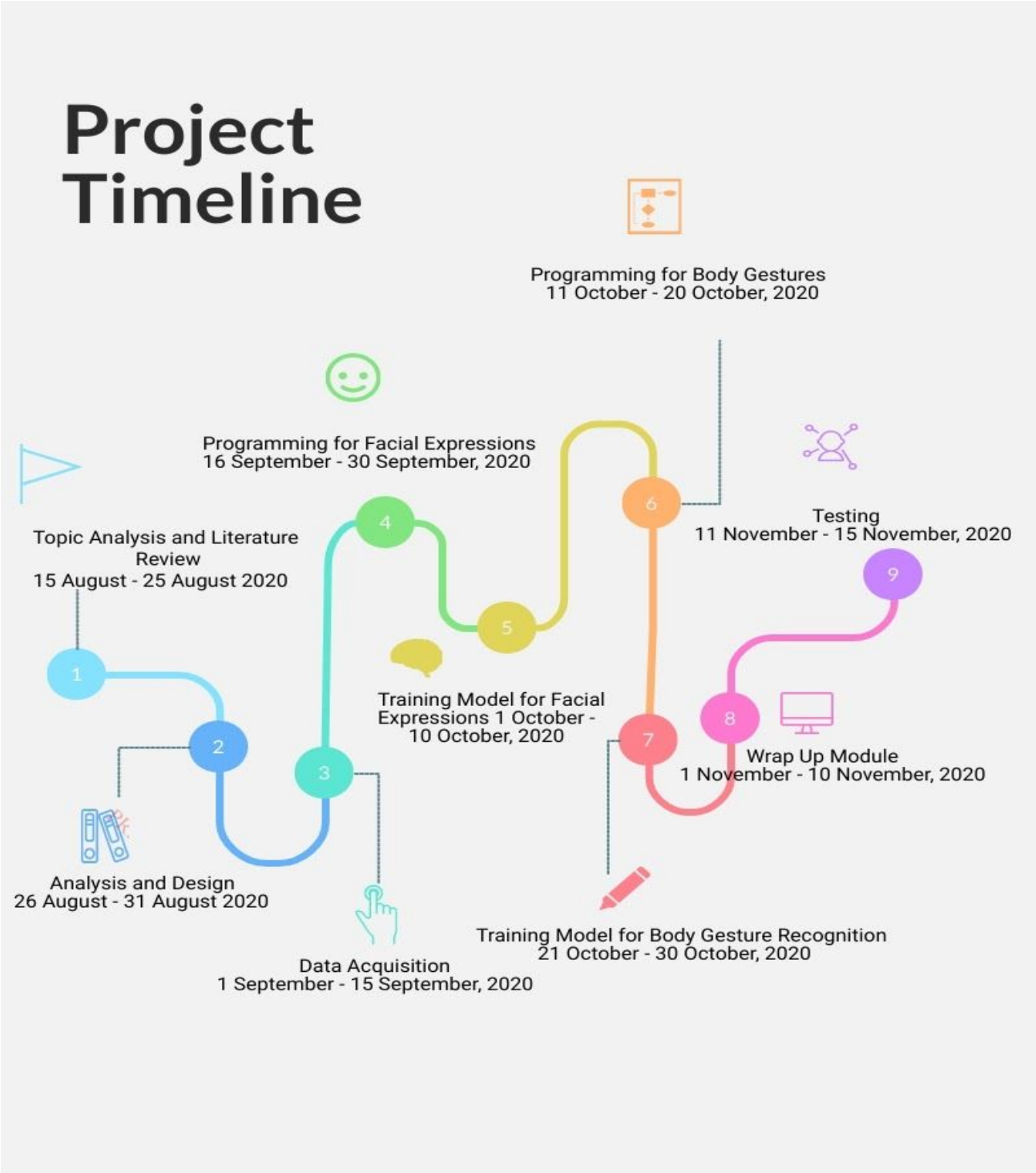
7. Language and Tools to be used in Implementation

- **Language Used:** We used python language for implementing our entire recognition system.
- **Tensorflow:** TensorFlow is the open-source library for a number of various tasks in machine learning.
- **Keras:** Keras is a neural network library in python.
- **Numpy:** Numpy is the core library for scientific computing in Python. It is widely used for image processing and to manipulate image data.
- **OpenCV:** OpenCV is a library of programming functions. It is a free open source library used in real-time image processing. It's used to process images, videos, and even live streams.
- **Openpose:** A library developed by CMU(Showing results for cmu university) for pose detection.
- **Scikit-learn:** It is a free software machine learning library for the Python programming language.

8. Suggestive Hardware Requirements

- **Processor:** Intel Core i5 or above.
- **Operating System:** Windows or Ubuntu.
- **RAM:** 8GB and above.

9. Activity Schedule



10. References

[1] V.Enrique Correa, Arnoud Jonker, Michael Ozo, Rob Stolk, “Emotion Recognition using Deep Convolutional Neural Networks”, Year of Conference: 2016.

[2] Zhijuan Shen, Jun Cheng, Xiping Hu, Qian Dong.Z, "EMOTION RECOGNITION BASED ON MULTI-VIEW BODY GESTURES".Date of Conference: 22-25 Sept. 2019.

[3] Kehan Chen, "Sitting Posture Recognition Based on OpenPose", Year of Conference: 2019.

[4] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov. " Siamese neural networks for oneshot image recognition.", Presented at ICML 2015 Deep Learning Workshop.

[5] Tanya Keshri, Suja Palaniswamy, “Emotion Recognition Using Feature-level Fusion of Facial Expressions and Body Gestures”, Presented at the Fourth International Conference on Communication and Electronics Systems (ICCES 2019) IEEE Conference Record # 45898.

[6] Anupam Agrawal , Nayaneesh Kumar Mishra, "Fusion Based Emotion Recognition System", 2016 International Conference on Computational Science and Computational Intelligence (CSCI)

[7] Liyanage C. De Silva, Pei Chi Ng, “Bimodal Emotion Recognition”, The National University of Singapore

[8] Tomasz Sapinski, Dorota Kaminska, Adam Pelikant, Gholamreza Anbarjafari, “Emotion Recognition from Skeletal Movements”, Entropy June 2019

[9] Prabhakar, Salil, and Anil K. Jain. "Decision-level fusion in fingerprint verification." Pattern Recognition. Vol. 35, no. 4, pp. 861-874, 2002.

[10] Viola, Paul, and Michael J. J, "Robust real-time face detection," International journal of computer vision, Published by Springer, vol. 57, no. 2,pp. 137-154, 2004.

[11] "Challenges in Representation Learning: A report on three machine learning contests." I Goodfellow, D Erhan, PL Carrier, A Courville, M Mirza, B Hamner, W Cukierski, Y Tang, DH Lee, Y Zhou, C Ramaiah, F Feng, R Li, X Wang, D Athanasakis, J Shawe-Taylor, M Milakov, J Park, R Ionescu, M Popescu, C Grozea, J Bergstra, J Xie, L Romaszko, B Xu, Z Chuang, and Y. Bengio. arXiv 2013.

[12]https://en.wikipedia.org/wiki/F1_score

[13]https://www.kaggle.com/ashishpatel26/tutorial-facial-expression-classification-keras/notebook

[14]https://github.com/CMU-Perceptual-Computing-Lab/openpose

[15]https://github.com/tensorfreitas/Siamese-Networks-for-One-Shot-Learning