Emotion Detection via Facial Expression Recognition

10.21.2022

Group 2

Vansh Agarwal
Kushal Agrawal
Andrea Corrales
Aparna Ramachandran
Vishal Shrivastava

Overview

Just like verbal communication, body language is a type of communication that is used to express information. The human face can convey a number of different emotions through the use of their eyes, eyebrows, mouth, and facial muscles. Without saying a word, a lot can be said from someone's facial expression. In this project, we present a machine learning system with the ability to detect human emotion through facial expressions.

Our team will be utilizing multiple datasets (CK+, FER-2013, and RAF-DB) for this task. Methods include detecting the face and its landmarks, utilizing linear and non-linear models, and results will be evaluated using standard multiclass classification metrics.

Motivation

In human interactions or human-computer interactions (HCI), recognizing the emotion is a vital phase. *In human interaction, 7% of the affective information is conveyed by words, 38% is conveyed by speech tone, and 55% is conveyed by facial expressions* [4]. Therefore, facial expression analysis is an ideal way of recognizing human emotions and this can aid and enhance applications in various fields. In this project, our goal is to study various datasets to recognize and classify emotions such as Neutral, Fear, Happy, Sad and so on. Our project can easily adapt to numerous applications, based on the problem we are trying to solve. Applications of our project include:

- **Driver assist systems in self-driving cars or trucks:** To detect whether the drivers are distracted and perform possible actions.
- **Physician assist systems:** To help diagnose early psychological disorders by recognizing a person's inability to express certain facial expressions.
- **Cognitive therapy:** To assist with stress and anxiety disorders among war veterans and equity traders, who are constantly under emotional strain.
- **Video game testing** for user enjoyment detection. This can be extended for user experience detection for any valid products.
- **Security systems and law enforcement** for detection of emotional incongruities or assisting with interrogation of suspects.
- **Recruitment:** To assist in interviews, especially for the positions where the person must showcase emotional intelligence.

Recognizing a facial expression as representation of a certain emotion can be difficult even for humans. Therefore, facial expression analysis and classification is a challenging area for AI and it provides us with a lot of opportunities for improving state-of-the-art systems.

Doing this successfully enables us to stretch the capabilities of a machine to characterize what is a very instinctive and innate human ability.

As with any machine learning and deep learning algorithms, emotion recognition will require a lot of data, specifically training on images. However, there can be several shortcomings in such a dataset as they might be randomly captured images, or screenshots from videos with various angles, backgrounds, with people of different genders, ethnicities, etc. Our project will also aim to address and resolve these issues such as:

- Multiple orientations of a face to the camera.
- Low resolution images.
- Gender, ethnicity or racial biases in the dataset used for training.

Data Acquisition and Preprocessing

Datasets

We will be utilizing the following existing datasets:

1. **CK+ (Extended Cohn-Kanade Dataset) [1]:** This dataset contains **593** *video sequences* from a total of 123 different subjects, ranging from 18 to 50 years of age with a variety of genders and heritage. Each video shows a facial shift from the neutral expression to a targeted peak expression, recorded at 30 frames per second (FPS) with a resolution of either **640x490 or 640x480 pixels**. Out of these videos, *327* are labeled with one of seven expression classes **(Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise)**. The CK+ database is widely regarded as the most extensively used laboratory-controlled facial expression classification database available, and is used in the majority of facial expression classification methods.

Fig 1. Example video frames from the CK+ dataset.

2. **FER-2013 (Facial Expression Recognition 2013 Dataset)** [2]**:** The data consists of **35,685 48x48 pixel** *grayscale images* of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. The task is to categorize each face based on the emotion shown in the facial expression into one of seven categories **(Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral)**.

Fig 2. Example images from the FER-2013 dataset.

3. **RAF-DB (Real-world Affective Faces Database)** [3]**:** This is a large-scale facial expression database with around 30K great-diverse *facial images downloaded from the Internet*. Based on the crowdsourcing annotation, each image has been independently labeled by about 40 annotators. Images in this database are of great

variability in subjects' age, gender and ethnicity, head poses, lighting conditions, occlusions, (e.g. glasses, facial hair or self-occlusion), post-processing operations (e.g. various filters and special effects), etc.

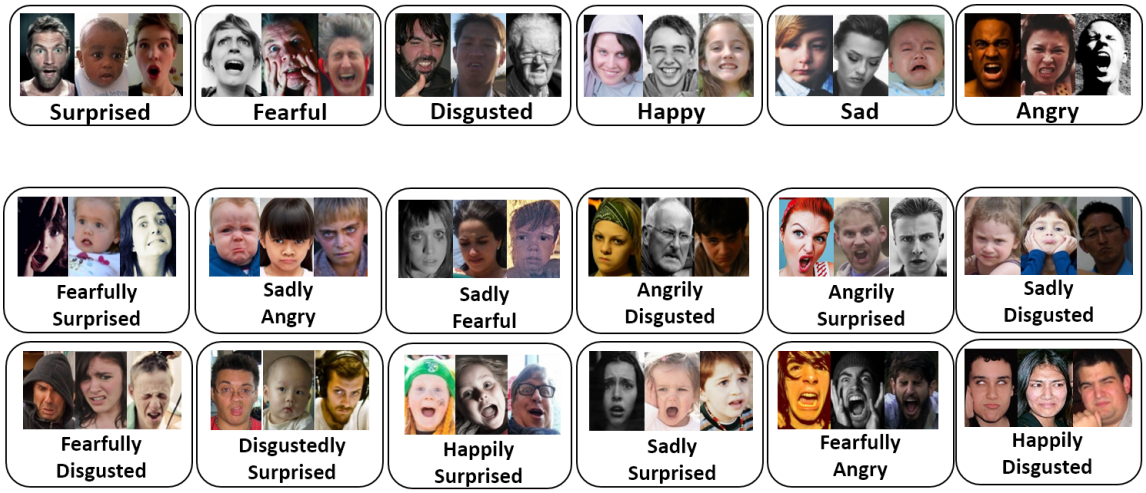

Fig 3. Example images from the RAF-DB dataset.

RAF-DB has large diversities, large quantities, and rich annotations, including:

a. **29,672** number of real-world images,
b. a **7-dimensional expression distribution vector** for each image,
c. **two different subsets**: single-label subset, including *7 classes of basic emotions* **(Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral)**; two-tab subset, including *12 classes of compound emotions*,
d. 5 accurate landmark locations, 37 automatic landmark locations, bounding box, race, age range and gender attributes annotations per image,
e. baseline classifier outputs for basic emotions and compound emotions.

Image Preprocessing

The following preprocessing techniques can normalize and augment our dataset and help generate useful features and insights:

- **Image Super Resolution using GANs**
- **Global contrast normalization**

- **Fast Fourier Transform and contrast-limited adaptive histogram equalization for handling poor illumination**

- **Cropping/resizing using bilinear interpolation to extract equal numbers of features from each image**

Pending Work

The following tasks need to be accomplished for our dataset to be complete:

- **Access to RAF-DB:** Access to RAF-DB is free for researchers, but requires them to email the original authors and request a password. We have done this, and are awaiting their response.

- **Normalization to a common resolution:** Since the images/videos are of varying resolutions (FER-2013 is just 48x48), we must find a practical way to feed them to our models. Simply scaling all images down to 48x48 is not a good idea, since we will lose a large amount of image fidelity.

- **Label Unification:** 6 of the 7 classes are the same across the three datasets. However, CK+ contains the emotion "Contempt" instead of which FER-2013 and RAF-DB contain neutral. We need to come up with a method to unify these.

- **Video Frame Selection:** Which frame(s) of the videos in CK+ should be utilized? In other words, when does the expression "peak" in the video?

Methods

Feature Extraction

There are several classes of emotions that we humans can express (and identify) through our facial expressions, including, but not limited to Amusement, Anger, Awe, Doubt, Pain, Concentration, Confusion, Contempt, Desire, Disappointment, Elation, Interest, Sadness, Surprise, Triumph, etc. For our analysis, we'll be considering the seven most common of these classes which are considered as the universal emotions. [5]

After pre-processing the dataset, we need to extract relevant features from the images. We'll be using the deep MediaPipe[6] technique developed by Google. Since we are using

python as our programming language, we can utilize several open-source ML solutions provided in MediaPipe, such as face detection, face mesh generation, hand tracking, and pose estimation, etc. For our analysis, we'll primarily be using the face mesh generation technique to detect the faces in the input images, generate boundary boxes, and identify several key landmarks. MediaPipe's face mesh generation technique is capable of inferring 3D surface geometry using just a single static image without any specialized depth sensor.

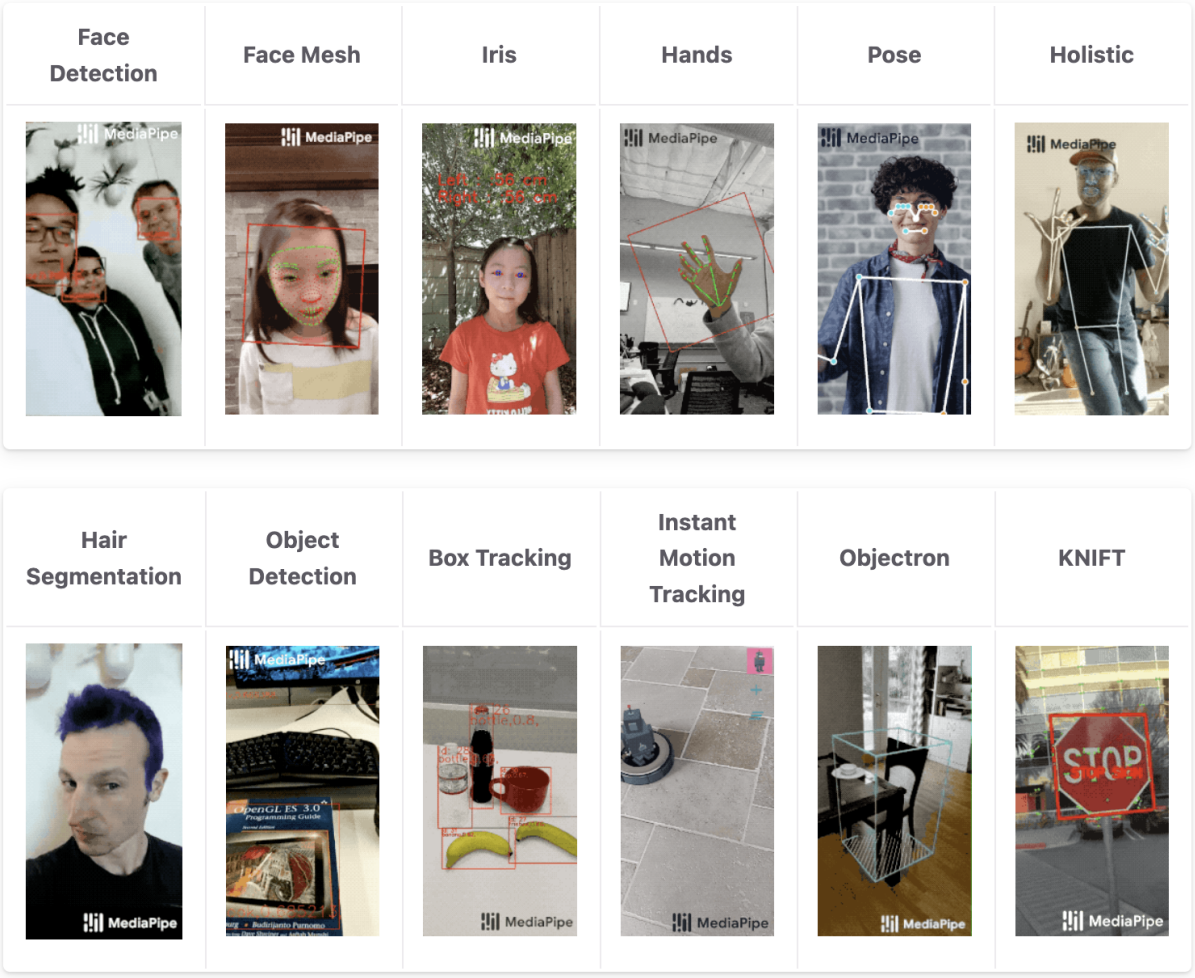

Fig 4. Examples of ML solutions provided by MediaPipe (source: [MediaPipe])

More specifically, we'll be using the attention mesh model provided by MediaPipe that applies attention [7] to semantically meaningful regions on the face. Consequently, this model is better at predicting landmarks around the lips, eyes, and irises at the cost of some additional computational complexity.

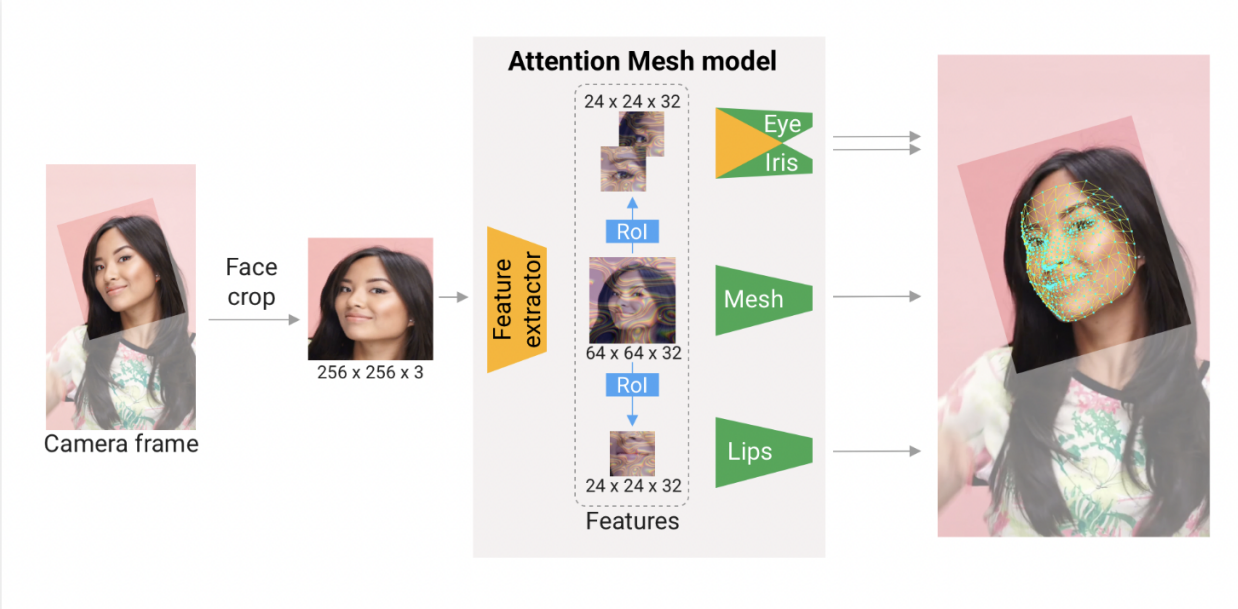

Fig 5. Attention Mesh: Overview of model architecture (source: MediaPipe)

Face mesh provides us with roughly 468 key landmarks all over the face along with their x and y coordinates. However, we don't need all those landmarks for expression analysis. Instead, we are going to use the Facial Action Coding System (FACS) developed by Carl-Herman Hjortsjö in 1969 [8] and further enhanced by Paul Ekman, and Wallace Friesen in 1978 [9] and again in 2002 [10] as our guide to select relevant landmarks that are important for describing the facial muscle movements or Action Units (AU) [10].

FACS encodes the individual facial muscle movements as identified by slight changes in the face appearance. Action Units (AUs) are defined as the fundamental units of the muscle movement generated by contraction or relaxation of the concerned muscle or the muscle groups.

AU number	FACS name
0	Neutral face
1	Inner brow raiser
2	Outer brow raiser
4	Brow lowerer
5	Upper lid raiser
6	Cheek raiser
7	Lid tightener
8	Lips toward each other
9	Nose wrinkler
10	Upper lip raiser
11	Nasolabial deepener
12	Lip corner puller
13	Sharp lip puller
14	Dimpler
15	Lip corner depressor
16	Lower lip depressor
17	Chin raiser
18	Lip pucker
19	Tongue show
20	Lip stretcher
21	Neck tightener
22	Lip funneler
23	Lip tightener
24	Lip pressor
25	Lips part
26	Jaw drop
27	Mouth stretch
28	Lip suck

AU number	FACS name
51	Head turn left
52	Head turn right
53	Head up
54	Head down
55	Head tilt left
M55	Head tilt left
56	Head tilt right
M56	Head tilt right
57	Head forward
M57	Head thrust forward
58	Head back
M59	Head shake up and down
M60	Head shake side to side
M83	Head upward and to the side

AU number	FACS name
61	Eyes turn left
M61	Eyes left
62	Eyes turn right
M62	Eyes right
63	Eyes up
64	Eyes down
65	Walleye
66	Cross-eye
M68	Upward rolling of eyes
69	Eyes positioned to look at other person
M69	Head or eyes look at other person

Fig 6. List of general AU codes and corresponding FACS descriptors (Source: Facial Action Coding System)

After choosing the key landmarks, we'll create an emotional mesh where each landmark will be represented as a vertex and the edges will define the connection between those landmarks. The angle deformations between these edges will represent the facial muscle contraction or relaxation and they will serve as our geometrical features to identify different emotional states. The range of these angles will be [0^0, 360^0].

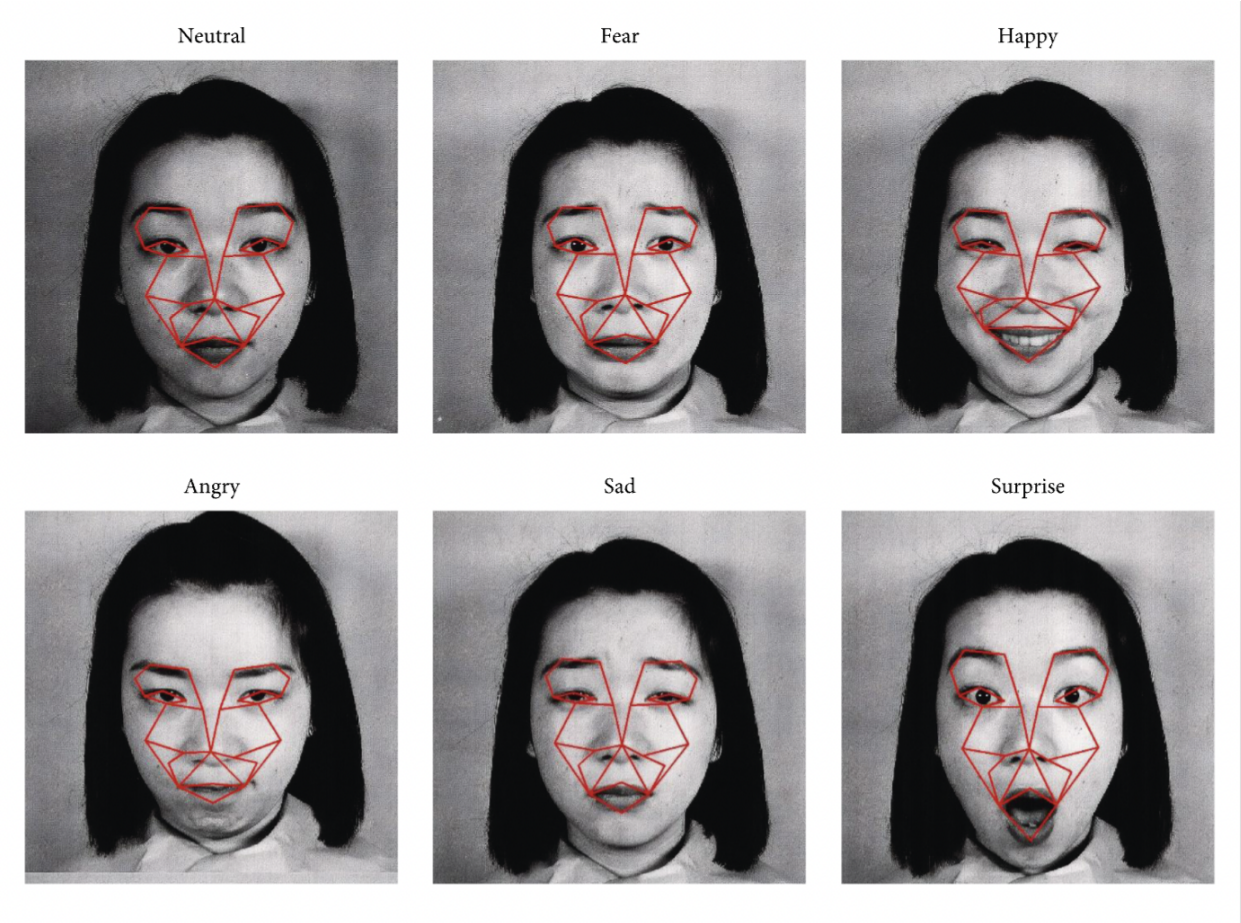

Fig 7. *Samples of emotion face mesh for different emotions (Source: https://www.hindawi.com/journals/cin/2022/8032673/fig9/)*

Classification & Evaluation

We'll be feeding these geometrical features to several linear and non-linear classifiers, such as K-Nearest Neighbors (KNN), Decision Trees (DT), Random Forest (RF), Multi-Layer Perceptrons (MLP), Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Gaussian Naive Bayes classifier (NB), and Quadratic Discriminant Analysis (QDA) model.

In order to evaluate the different classification models, we'll compare the standard classification evaluation metrics for each model. Since this is a multiclass classification problem, for each class (expression) we will generate a confusion matrix and use the classification accuracies, precision, recall, F-1 score, and area under the curve (AUC) metrics for quantitative comparisons.

References

1. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, *"The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression,"* in Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition- Workshops, pp. 94–101, San Francisco, CA, USA, June 2010.

2. Li S, Deng W. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. IEEE Trans Image Process. 2019 Jan;28(1):356-370. doi: 10.1109/TIP.2018.2868382. Epub 2018 Sep 3. PMID: 30183631.

3. I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. 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. *Challenges in representation learning: A report on three machine learning contests.* Neural Networks, 64:59--63, 2015. Special Issue on "Deep Learning of Representations"

4. A. Mehrabian, "Communication without words," in *Communication Theory*, pp. 193–200, Routledge, UK, London, 2017. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315080918-15/communication-without-words-albert-mehrabian

5. https://www.paulekman.com/resources/universal-facial-expressions/ [Accessed: 10/21/2022]

6. https://google.github.io/mediapipe/ [Accessed: 10/21/2022]

7. https://en.wikipedia.org/wiki/Attention_(machine_learning) [Accessed: 10/21/2022]

8. Hjortsjö CH (1969). Man's face and mimic language. http://diglib.uibk.ac.at/ulbtirol/content/titleinfo/782346 [Accessed: 10/21/2022]

9. P. Ekman, Facial Action Coding System, Consulting Psychologists Palo Alto, Palo Alto, CA, USA, 1978.
10. Ekman P, Friesen WV, Hager JC (2002). Facial Action Coding System: The Manual on CD ROM. Salt Lake City: A Human Face.