Human Emotion Detection using Machine Learning Techniques

Katariya Dharmesh Ashok, Student, dharmessh.ashok.katariya@gmail.com, University of Mumbai
Dongre Rasik Baburao, Student, rasikdongare77@gmail.com, University of Mumba
Satyendra Pal, profsatyen@gmail.com, Navneet College of Arts, Science & Commerce

Abstract
The process of converting an image into digital form and applying various procedures to it is known as image processing. This is done in order to improve the image or to glean information from it. One way to communicate nonverbally is through facial expressions. Neutral, joyful, sad, angry, contemptuous, disgusted, afraid, and surprised are among the eight universal face expressions. Therefore, it is crucial to recognize these facial emotions. A technologically based monitoring system for senior citizens that uses video images to identify emotions. In order to monitor the living conditions of elderly people in real time, our suggested solution incorporates video analysis technologies. The system will send a message to their children and family in the event of an emergency.

Keywords:

Convolutional Neural Networks (CNN), Local Binary Pattern Histogram (LBPH) technique, facial expression recognition.

INTRODUCTION

One way to describe a facial expression is as the movement of muscles under the face's skin. One way to communicate nonverbally is through facial expressions. Without uttering a single word, the human face could express a wide range of emotions. Furthermore, these facial expressions are universal and understandable by all types of people, in contrast to certain nonverbal
communication methods that are not. People from many cultures use the same facial expressions to indicate happiness, grief, anger, surprise, fear, and disgust.

People can see a person's emotions through their bodily movements. They serve as the channel for spreading social knowledge between humans, but the majority of other mammals and a few other animal species also experience them. Both voluntary and involuntary facial expressions can be adopted by humans. Expressions that people produce when they are sick, wounded, or uncomfortable are known as involuntary expressions.

Figure 1: Illustrations of the six fundamental emotions

The study of systems and the creation of tools and systems that can detect, understand, process, and mimic human emotions is known as affective computing. Through the use of sensors, microphones, and cameras, affective computing technologies are able to detect the user's emotions and react by carrying out certain predetermined aspects of a product or service. Human-computer interaction, in which a technology can recognize and react to the emotions displayed by its users, is one way to view effective computing.

Emotion recognition can be used to screen for mental illness and emotion-related physiology as well as to monitor users' emotional well-being. Emotions are manifested through a range of physiological alterations in addition to psychological behavioral performance. Humans have no control over these physiological changes. As a result, physiological cues may accurately represent participants' actual emotions. The electrocardiogram (ECG), respiratory suspended particulate (RSP), electroencephalogram (EEG), galvanic skin response (GSR), and blood volume pulse (BVP) are among the physiological markers that have been effectively used for emotion identification. The link between the heartbeat and emotional fluctuations is reflected in these physiological signs, most notably the ECG. A lot of research has been done on ECG-based emotion recognition. One of the crucial factors in emotion

detection is heart rate variability (HRV), which is taken from an ECG.

Our objective is to operate in real time, identifying emotions from photos taken by a live webcam. The camera will now play a video, and faces will be identified in the frames based on facial landmarks, which include the mouth, nose, eyes, eyebrows, and corners of the face. Subsequently, the traits that will be used to identify facial emotions were taken from these facial landmarks, or dots. Following the identification of the emotions, we use image processing tools to check for any discomfort.
The study of systems and the creation of tools and systems that can detect, understand, process, and mimic human emotions is known as affective computing. Through the use of sensors, microphones, and cameras, affective computing technologies are able to detect the user's emotions and react by carrying out certain predetermined aspects of a product or service. Human-computer interaction, in which a technology can recognize and react to the emotions displayed by its users, is one way to view effective computing.

Emotion recognition can be used to screen for mental illness and emotion-related physiology as well as to monitor users' emotional well-being. Emotions are manifested through a range of physiological alterations in addition to psychological behavioral performance. Humans have no control over these physiological changes. As a result, physiological cues may accurately represent participants' actual emotions. The electrocardiogram (ECG), respiratory suspended particulate (RSP), electroencephalogram (EEG), galvanic skin response (GSR), and blood volume pulse (BVP) are among the physiological markers that have been effectively used for emotion identification. The link between the heartbeat and emotional fluctuations is reflected in these physiological signs, most notably the ECG. A lot of research has been done on ECG-based emotion recognition. One of the crucial factors in emotion detection is heart rate variability (HRV), which is taken from an ECG.

Our objective is to operate in real time, identifying emotions from photos taken by a live webcam. The camera will now play a video, and faces will be identified in the frames based on facial landmarks, which include the mouth, nose, eyes, eyebrows, and corners of the face. Subsequently, the traits that will be used to identify facial emotions were taken from these facial landmarks, or dots. Following the identification of the emotions, we use image processing tools to check for any discomfort.

BACKGROUND

- MACHINE LEARNING

  A technique called machine learning aids in our understanding of the models and algorithms that a computer system can employ to carry out a particular activity without the need for outside guidance. Machine learning algorithms create mathematical models using training data, which are samples of data that can be utilized to make judgments or predictions without the influence of outside variables on task performance.

  There are various categories into which machine learning can be divided. The method in supervised learning creates a mathematical model from a set of data that includes the necessary outputs as well as the inputs. Supervised learning includes techniques for regression and classification. When we wish to confine the output to a specific set of values, classification methods can be employed. The reason regression algorithms are designated for continuous output is that their values fall within a specific range. Algorithms for semi-supervised learning build mathematical models from training data that is not fully supervised, meaning that some sample input is labeled. Without any intended output labels, an algorithm in unsupervised learning creates a mathematical model from a collection of input data.

There are various categories into which machine learning can be divided. The method in supervised learning creates a mathematical model from a set of data that includes the necessary outputs as well as the inputs. Supervised learning includes techniques for regression and classification. When we wish to confine the output to a specific set of values, classification methods can be employed. The reason regression algorithms are designated for continuous output is that their values fall within a specific range. Algorithms for semi-supervised learning build mathematical models from training data that is not fully supervised, meaning that some sample input is labeled. Without any intended output labels, an algorithm in unsupervised learning creates a mathematical model from a collection of input data.

- IMAGE RECOGNITION

The technology that recognizes locations, logos, people, objects, buildings, and other elements in pictures is known as image recognition. The ability to recognize and detect an object in a digital video or image is known as image recognition, and it is a component of computer vision. The field of computer vision encompasses techniques for collecting, processing, and evaluating data from real-world video or still images. Such sources provide high-dimensional data that yields decisions

that are either symbolic or numerical in nature. In addition to image identification, computer vision encompasses object recognition, event detection, learning, video tracking, and picture reconstruction.

The visual cortex in our brain processes the impulses that the human eye interprets as a picture. The goal of image recognition is to replicate this process precisely. A computer can interpret a picture as either vector or raster. Vector images are a collection of color-annotated polygons, whereas raster images are a series of pixels with distinct numerical values for the colors in the images. In order to evaluate images, the geometric encoding is converted into constructs that represent objects and physical properties. The computer then analyzes these constructs logically. Classification and feature extraction are two aspects of data organizing. In order to classify an image, the initial step is to simplify it by removing only the crucial information that is required while excluding other information.

Building a prediction model that can be utilized with a classification algorithm is the second phase. We must train the classification algorithm by displaying thousands of project-related subject and non-subject photos before it can function. We employ neural networks to create a prediction model. Similar to our brain, a neural network is a system that combines hardware and software to estimate functions based on a vast number of unknown inputs. Numerous image classification algorithms, including logistic regression, K-nearest neighbors (KNN), support vector machines (SVM), and facial landmark estimation, are available for image recognition.

Image recognition is the third phase. Both training and test picture data are arranged. We distinguish between training and test data by eliminating duplicates. Images are recognized by the model that receives this data. In order to determine the closest match with the subject, we must now train a classifier that uses measures from a fresh test image. Only milliseconds are needed for this classifier. Either subject or non-subject is the classifier's output.

- FEATURE EXTRACTION

  Feature extraction is a dimensional reduction procedure that reduces an original raw data collection for processing purposes. An image's behavior is defined by its features. In essence, features are patterns in an image, such edges or points. When you need to minimize processing resources while maintaining the essential and pertinent data, feature extraction is a helpful technique. Feature extraction can reduce the amount of redundant data. The sampled image is subjected to image preprocessing methods including thresholding, scaling, normalizing, binarization, etc., after which the features are extracted. To obtain features for picture classification and recognition, feature extraction techniques are used.

Among the feature detection techniques are Color Gradient Histogram and ORB. In reality, the ORB (Oriented FAST and Rotated BRIEF) algorithm combines the concepts of FAST and BRIEF. ORB approach effectively locates the image's corners. Areas of the image with a stark contrast in brightness are recognized as features by the FAST component. A spot is marked as a feature if more than eight neighboring pixels are either darker or brighter than the pixel in question. In order to convey this, BRIEF transforms the retrieved points into binary feature vectors. By simply measuring an image's red, green, and blue values, the Color Gradient Histogram technique locates images with comparable color proportions. Binding the values allows you to adjust the Color Gradient Histogram.

Computer systems utilize a technique called corner detection to extract characteristics. The contents of a picture are inferred using those features that were extracted. Motion detection, picture registration, video tracking, image mosaicing, 3D modeling, and object recognition are a few of the applications that use corner detection. For corner identification, the Harris and Shi-Tomasi corner detectors are frequently utilized. A technique called Harris Corner Detector may identify which windows, when moved in both X and Y directions, result in very noticeable changes in intensity (i.e. gradients). A score R is calculated for each such window that is discovered. Important corners are chosen and marked after a threshold is applied to this score. An other technique for locating corners is the Shi-Tomasi corner detector.

Similar to the Harris Corner detector, it calculates a score (R) and allows us to identify the top N corners, which may be helpful if we don't want to detect every corner. R is computed using the following formula:

It is categorized as a corner if R exceeds the threshold. Because it takes less time and provides higher accuracy, the Viola Jones algorithm is used for facial feature detection. The Viola Jones detection framework uses basic features called Haar-like features to identify faces or facial traits. Feature boxes are passed over the image, and the difference in summed pixel values between neighboring regions is calculated. After that, the difference is evaluated to a threshold that determines whether or not an object is deemed detected. Thresholds that have been previously taught for various feature boxes and features are necessary for this.

Let's look at the picture below. The top row displays two positive attributes. The first characteristic chosen is that the area around the eyes is darker than the area around the nose and cheekbones. The second element that was chosen emphasizes the fact that the eyes are darker than the nasal bridge. However, it doesn't matter if the same windows are applied to the cheeks or somewhere else.

A target-sized window is positioned over the input image during the detecting phase. The Haar features are computed for every image portion. Different features display different values. After that, the difference is contrasted with a threshold that distinguishes objects from non-objects. A "weak classifier" is any Haar feature that identifies a little better than random guessing. A powerful classifier is created by organizing a large number of Haar characteristics into cascade classifiers, which are necessary to accurately distinguish an item from a non-object.

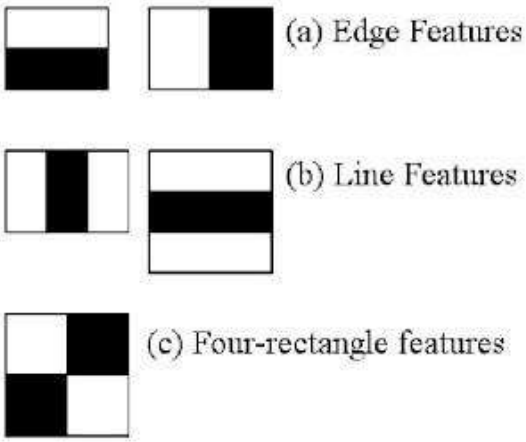
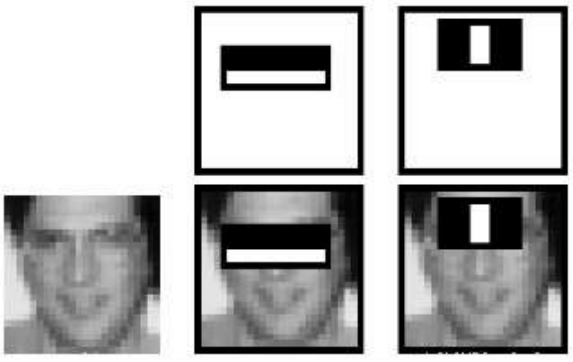


Fig. 2: Feature boxes



There are weak learners in each of the cascade classifier's steps. A method known as boosting is used to train each step. By calculating the weighted average of the choices made by the weak learners, boosting aids in the training of a highly accurate classifier.

Using the sliding window's current position, the classifier assigns a positive or negative label to the region at each stage. A positive result indicates the presence of an object, while a negative result indicates none was discovered. The detector moves the window to the next spot if the label is negative, completing the classification of that specific area. The classifier advances to the following step if the label is affirmative. When the final stage qualifies the region as positive, the detector reports an object found at the current window location.

LEARNING METHODS

A LOCAL HISTOGRAM OF BINARY PATTERNS

An image's pixels are labeled using the Local Binary Pattern (LBP), which also thresholds the area around each pixel and treats the outcome as a binary integer. Combining LBP with the Histogram of Oriented Gradients has been shown to enhance detection performance. LBPH employs four parameters:

Radius: the radius, which is typically set to 1, represents the radius surrounding the central pixel and can be used to construct the circular local binary pattern.

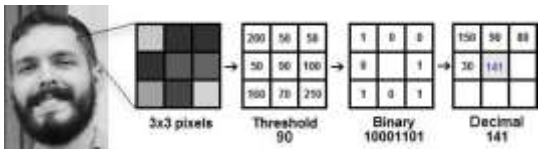
Neighbors: The number of sample points needed to construct the circular local binary pattern is known as the neighbors. The computational cost increases with the number of sample points contained, and it is typically set at 8.

Grid X: the quantity of cells oriented horizontally. The generated feature vector's dimensionality increases with the number of cells and grid fineness. Typically, it is set to 8.

Grid Y: the quantity of cells oriented vertically. The generated feature vector's dimensionality increases with the number of cells and grid fineness. Typically, it is set to 8.

We must set an ID and utilize a dataset with several facial photos in order to train the algorithm. The same person's photos must match their ID.

Building an intermediate image to more accurately depict the original image is the first stage of LBP.

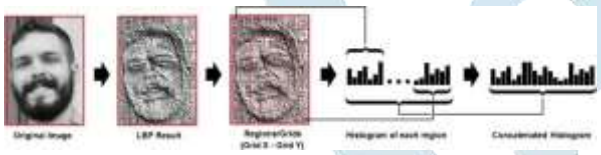


As demonstrated in the figure above, we may extract a portion of a grayscale image into 3x3 pixels. It is a matrix made up of each pixel's intensity (0-255). The matrix's central value is then used as the threshold value. We must set a new binary value for each neighboring value of the core value. Values below the threshold are represented by 0 and values equal to or above the threshold by 1.

The central value will now be ignored and just the binary value will be present in the matrix. Create a new binary value by concatenating all of the binary values from each point. Next, put that binary value as the matrix's central value after converting it to a decimal number. We obtain a new image at the conclusion of the LBP process that has improved features of the original image.

The image above can now be used to create histograms. The image can be divided into several grids using the parameters grid X and grid Y. There are only 256 places (0~255) in each histogram (from each grid) that correspond to the occurrences of each pixel intensity. To create a larger histogram, we should concatenate each one.

Every image from the training dataset is represented by the histogram that has been produced. In order to generate a histogram that accurately depicts an input image, we must repeat the processes for the new image. We only need to compare two histograms and return the image with the closest histogram in order to determine which image matches the input image. The distance between two histograms can be calculated using a variety of methods, such as the euclidean distance, chi-square, absolute value, etc.



Thus, ID with the image that has the closest histogram is the algorithm's output. The computed distance should also be returned by the method.

- CONVOLUTIONAL NEURAL NETWORKS

  Neurons with learnable weights comprise convolutional neural networks. Every neuron takes in multiple inputs, adds them up, and then outputs the result. The architecture of convolutional neural networks differs from that of ordinary neural networks. Regular Neural Networks use a sequence of hidden layers to change an input. Each layer is composed of a collection of neurons, and every layer is completely connected to every other layer's neurons. Lastly, the output layer is the final fully connected layer.

  There are differences with Convolutional Neural Networks. Three dimensions—width, height, and depth—are used to arrange the layers. Not every neuron in the subsequent layer is connected to every other neuron in that layer; instead, a tiny portion of it. Finally, a single vector of probability scores arranged along the depth dimension will be the only remaining portion of the final output.

There are four ideas in CNN: 1. Convolution
2. ReLu
3. Pooling
4. Complete interconnectedness

Convolution as a feature extraction method

CNN uses a filter or a kernel to perform convolution on an input image. Filtering and convolution entail scanning the screen from top left to right, then going down a bit after covering the screen's width and repeating the procedure until the entire screen has been scanned. The image aligns with the person's facial features. The matching feature pixel is multiplied by the picture pixel. The sum of the values is divided by the feature's total pixel count.

Extraction of Features: Non-Linearity
The output we obtain after applying our filter to the original image is then run through an activation function, which is another mathematical function.

ReLu, or Rectified Linear Unit, is the activation function that is typically utilized in CNN feature extraction. which only maintains the positive values while converting all of the negative numbers to 0. Eliminating every negative value from the convolution is the goal. While the negative values go to zero, all of the positive values stay the same.

Pooling for Feature Extraction
Once the feature maps are obtained, a pooling or sub-sampling layer must be added to the CNN layers. The Pooling layer, like the Convolutional Layer, is in charge of shrinking the Convolved Feature's spatial size. In order to handle the data through dimensionality reduction, less computing power will be needed. Additionally, it is helpful in preserving the model's successful training process by extracting dominating features that are both rotational and positional invariant. Pooling prevents over-fitting and reduces training time.

Fully Connected Layer (FC Layer) classification: After transforming our input image into an appropriate format, we will flatten it into a column vector. A feed-forward neural network receives the flattened output, and back propagation is used for each training cycle. Using the Softmax Classification approach, the model can categorize images by differentiating between dominating and specific low-level attributes.

We now have every component needed to construct a CNN. Pooling, ReLU, and Convolution. The classifier we first mentioned, which is often a multi-layer perceptron layer, receives the output of max pooling. These layers are typically utilized repeatedly in CNNs, such as Convolution -> ReLU -> Max-Pool -> Convolution -> ReLU -> Max-Pool, and so forth. For now, we won't talk about the completely connected layer.

CONCLUSION

Based on the video data, the suggested model's output provides an estimated sentiment prediction for the subject. Peers and family members can take action to encourage, motivate, and uplift the subject's emotional stature in the event of "critical" sentiments, resulting in the subject's harmony and peace of mind. The resulting output can be used in a variety of situations, including estimating mental disorders and stress levels. Consequently, these sentiment analysis algorithms are essential for transforming society into a dynamic environment.

RESOURCES

1. In Renuka S. Deshmukh, Shilpa Paygude, and Vandana Jagtap present a machine learning approach to facial emotion recognition.
2. Facial Feature Recognition for Emotion Identification, Pao James
3. "Facial Emotion Recognition System through Machine Learning approach," by Renuka S. Deshmukh, Vandana Jagtap, and Shilpa Paygude, 2017.
4. "The Research of Elderly Care System Based on Video Image Processing Technology," by Dongwei Lu, Zhiwei He, Xiao Li, Mingyu Gao, Yun Li, and Ke Yin, 2017.
5. Shivam Gupta, "Recognition of facial emotions in both static and real-time images," 2018.
6. "Facial Emotion Recognition," by Ma Xiaoxi, Lin Weisi, Huang Dongyan, Dong Minghui, and Haizhou Li, 2017.
7. Mohammad. M. Alyan Nezhadi, Mostafa Muhammadpour, Hossein Khaliliardali.
8. "Facial Emotion Recognition Using Deep Convolutional Networks," by Seyyed Mohammad R. Hashemi, 2017.
9. "Automatic Emotion Recognition Using Facial Expression: A Review," International Research Journal of Engineering and Technology, 2017, Dubey, M., and Singh, P. L.
10. 8/how-to-switch-backend-with-keras-from-tensorflow-to-theano: https://stackoverflow.com/questions/4217765
11. Grammatical+Facial+Expressions: https://archive.ics.uci.edu/ml/datasets
12. Convolutional neural network: