Human Emotion Detection using Machine Learning Techniques

Article *in* SSRN Electronic Journal · January 2020

DOI: 10.2139/ssrn.3591060

Punidha Angusamy
Coimbatore Institute of Technology

Human Emotion Detection using Machine Learning Techniques

Punidha A, Inba. S, Pavithra. K.S., Ameer Shathali. M, Athibarasakthi. M
Coimbatore Institute of Technology, Coimbatore

Abstract---**Image processing is a method to convert an image into digital form and perform some operations on it. This is done to enhance the image or to extract useful information from it. Facial expressions are nonverbal form of communication. There are eight universal facial expressions which include: neutral, happy, sadness, anger, contempt, disgust, fear, and surprise. So it is very important to detect these emotions on the face. A monitoring system for elderly people which is based on technology involving recognizing emotions from video image. Our proposed system includes video analysis technology which includes the data from video is adopted to realize monitoring elders' living conditions in real time. In case of emergency, the system will alert their relatives and children by sending a message.**

Keywords-facial emotion recognition; Local Binary Pattern Histogram(LBPH)algorithm,Convolutional Neural Networks(CNN)

I. INTRODUCTION

A facial expression can be said as the movement of muscles beneath the skin of the face. Facial expressions are a form of nonverbal communication.Human face could convey countless emotions without saying a single word. And unlike some forms of nonverbal communication are not universal but these facial expressions are universal and can be understood by any kind of people. The facial expressions for happiness, sadness, anger, surprise, fear, and disgust are the same across the people of different cultures.

The movements of muscles convey the emotions of individual to people who see them. They are the means through which social information is conveyed between humans but they also occur in most other mammals and some other animal species. Humans can adopt a facial expression voluntarily or involuntarily. Involuntary expression are those that when people make when they are ill, hurt or feeling uncomfortable.

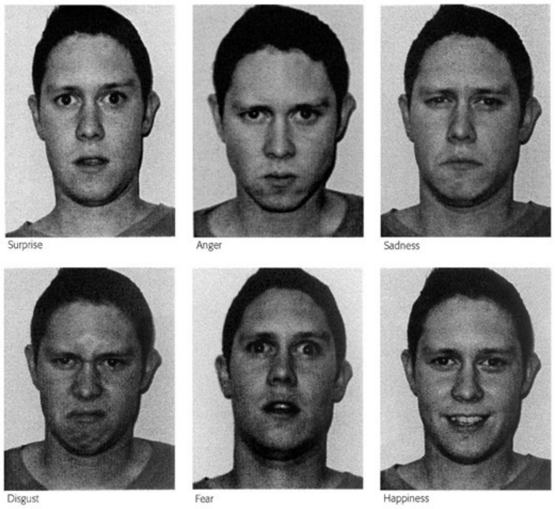


Fig. 1: Example of expression for the six basic emotions

Affective computing is the study of systems and developing systems and devices that can recognize, interpret, process and simulate human affects. Affective computing technologies can sense the emotions of the user through devices such as sensors, microphone, cameras and respond by performing some specific, predefined product/service features One way to look at effective computing is *human-computer interaction* in which a device has the ability to detect and respond to the emotions exhibited by the users.

Emotion recognition can help in monitoring the emotional health of the users and screening for emotion-related physiology and mental disease. Emotions are not only expressed through psychological behavioral performance, but also through a series of physiological changes. These physiological changes are not being controlled by humans. Thus, physiological signals can possibly reflect the true feelings of subjects. There are several kinds of physiological signals that have been successfully applied to emotion recognition, including electrocardiogram (ECG), galvanic skin response (GSR), electroencephalogram (EEG), respiratory suspended particulate (RSP) and blood volume pulse (BVP).These physiological signals, most importantly ECG reflect the relationship between the heart beating and emotion changes. Researchers have performed much work on emotion recognition based on ECG.Heart Rate Variability (HRV) which is extracted from an ECG is considered to be one of the important parameters of emotion recognition.

Our aim is to work in the real time in which we detect the emotions from images that has been captured by live webcam. Now the webcam will be running a video and the faces are going to be detected in the frames according to the facial landmarks which will contain the eyes, eyebrows, nose, mouth, corners of the face. Then the features were extracted from these facial landmarks (dots) faces which will be utilized for the detection of the facial emotions. After the emotions are identified, we look for any discomfort in the emotions through image processing techniques.

II. BACKGROUND

A. MACHINE LEARNING

Machine Learning is a method which helps us to study about algorithms and models that a computer system can use to perform some specific task without external instructions. Machine Learning algorithms build mathematical model based on some sample data which is called as training data which could be used to make predictions or decisions without any external factors affecting the performance of the task.

Machine learning can be classified into several categories. In supervised learning, the algorithm builds a mathematical model from a data set that contains both the inputs and the required outputs. Classification algorithms and regression algorithms come under supervised learning. Classification algorithms can be used when we want to restrict the out to some limited set of values. Regression algorithm are named for continuous output because they have value within a particular range. Semi-supervised learning algorithms create mathematical models from incomplete training data, in which a portion of the sample input does not have any labels. In unsupervised learning, the algorithm builds a mathematical model from a set of input data and it has no desired output labels.

B. IMAGE RECOGNITION

Image recognition refers to the technology that identifies places, logos, people, objects, buildings, and other things in images. Image recognition is a part of computer vision and it is a process that can identify and detect an object in a digital video or image. Computer vision is the one in which it includes the methods of gathering, processing and analyzing data from video or static images that comes from the real world. The data that comes from such source is high-dimensional and produces either numerical or symbolic information in the form of decisions. Apart from image recognition, computer vision also includes event detection, learning, object recognition ,video tracking and image reconstruction.

The human eye see an image as set of signals which will be processed by the visual cortex which is there in our brain. Image recognition tries to exactly recreate this process. Computer perceives an image as either raster or vector image. Raster images has sequence of pixels which has discrete numerical values for the colors in the images while vector images are a set of color-annotated polygons. To analyze images the geometric encoding is transformed into constructs that depicts physical features and objects. These constructs are then logically analyzed by the computer. Data organization involves classification and feature extraction. The first step in image classification is to make the image simple by extracting only the important information that is needed and leaving out other information.

The second step is to build a predictive model for which a classification algorithm can be used. Before classification algorithm works, we need to train it by showing thousands of subject and non-subject images as relative to the project. To build a predictive model we make use of neural networks. The neural network is a system which is a combination of hardware and software similar to our brain and it estimate functions based on huge amount of unknown inputs. There are numerous algorithms for image classification in recognizing images such as support vector machines (SVM), face landmark estimation, K-nearest neighbors (KNN), logistic regression etc.

The third step is image recognition. The image data both training and test data are organized. Training data differs from test data in which we remove duplicates between them. This data is been fed into the model which in turn recognize images. Now we need to train a classifier that takes measurements from a new test image and tells us about the closest match with the subject. This classifier takes only milliseconds. The result of the classifier is either subject or non-subject.

C. FEATURE EXTRACTION

Feature extraction is a process of dimensional reduction by which an initial set of raw data is reduced for some processing purpose. Features define the behavior of an image. Basically features refer to a pattern found in an image such as a point or edge. The process of feature extraction is useful when you need to reduce the number of resources needed for processing while retaining the important and relevant information. The amount of redundant data can be reduced by feature extraction. Image preprocessing techniques such as thresholding, resizing, normalization, binarization, etc. are applied on the sampled image and the features are extracted later. Feature extraction techniques are applied to get features for classifying and recognition of images.

ORB and Color Gradient Histogram are some of the feature detection algorithms. ORB (Oriented FAST and Rotated BRIEF) algorithm is actually a combination of FAST and BRIEF.ORB method efficiently finds the corners of the image. The FAST component identifies features and is identified as areas of the image with a sharp contrast of brightness. If more than 8 surrounding pixels are brighter or darker than a given pixel, that spot is flagged as a feature. BRIEF expresses this by converting the extracted points as binary feature vectors. Color Gradient Histogram method simply measures the proportions of red, green, and blue values of an image and finds an image with similar color proportions. Color Gradient Histogram can be tuned through binning the values.

Corner detection is a method which is used by computer systems to extract the features. Those extracted features are used to infer the contents of an image. Some of the applications where corner detection is used are motion detection, image registration, video tracking, image mosaicing, 3D modelling and object recognition. Harris corner detector and shi-Tomasi corner detector are used widely for corner detection. Harris Corner Detector is a method in which it determines which windows produce very large variations in intensity when moved in both X and Y directions (i.e. gradients).With each such window found, a score R is computed. A threshold is applied to this score and then important corners are selected & marked. Shi-Tomasi corner detector is another method to find the corners. It is almost similar to Harris Corner detector the score (R) is calculated and we can find the top *N* corners, which might be useful in case we don't want to detect each and every corner. R is calculated by the formula:

$R = \min(\lambda_1, \lambda_2)$, If R is greater than the threshold, it is classified as a corner.

The Viola Jones algorithm is implemented for the face feature detection as it does not consumes much time, thus giving greater accuracy[1]. The Viola Jones detection framework identifies the faces or features of the face by using simple features known as Haar-like features. The process involves passing feature boxes over the image and computing the difference of summed pixel values between adjacent regions. The difference is then compared with a threshold which indicates whether an object is considered to be detected or not. This requires thresholds that have been trained in advance for different feature boxes and features[2].

Let us consider the image below. Top row shows two good features. The first feature selected focuses on the property that the region of the eyes is darker than the region of nose and cheeks. The second feature selected focuses on the property that the eyes are darker than the bridge of the nose. But the same windows applying on cheeks or any other place is irrelevant.

During the detection phase, a window which is of target size is moved over the input image. For each subsection of the image the Haar features are calculated. Various features show various values. The difference is then compared to a threshold that separates non-objects from objects. Each Haar feature is a "weak classifier" because it detects slightly better than random guessing. A large number of Haar features are required to distinguish an object from non-object with sufficient accuracy and are therefore organized into *cascade classifiers* to form a strong classifier.

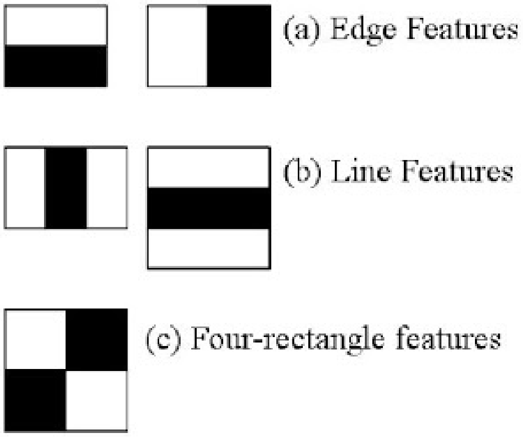
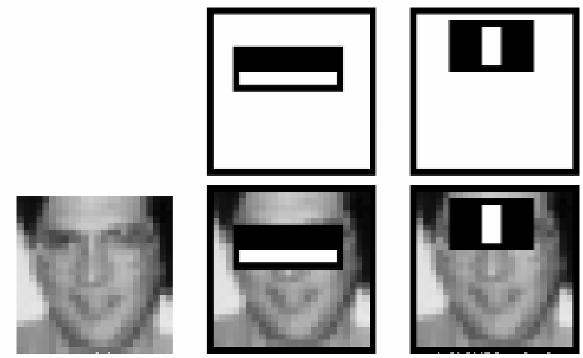


Fig. 2: Feature boxes



The cascade classifier has series of stages, in which each stage has weak learners. Each stage is trained using a technique called as boosting. *Boosting* helps to train a highly accurate classifier by taking weighted average of the decisions made by the weak learners.

In each stage the classifier labels the region by using the current location of the sliding window as either positive or negative. *Positive* means that an object is found and *negative* means no objects were found. If the label is negative, the classification of that particular region is complete and the detector slides the window to the next location. If the label is positive, the classifier passes the region to the next stage. The detector reports an object found at the current window location when the final stage classifies the region as positive.

III. LEARNING METHODS

A. LOCAL BINARY PATTERN HISTOGRAM

Local Binary Pattern (LBP) is the one which labels the pixels of an image and thresholds the neighborhood of each pixel and considers the result as a binary number. It is found that when LBP is combined with Histogram of Oriented Gradients improves detection performance. LBPH uses 4 parameters:

Radius: the radius can be used to build the circular local binary pattern and it represents the radius around the central pixel which is usually set to 1.

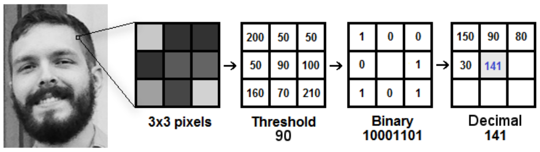
Neighbors: Neighbors are the number of sample points that is required to build the circular local binary pattern. The more sample points included, the higher the computational cost and it is usually set to 8.

Grid X: the number of cells in the horizontal direction. More the cells, finer the grid, the higher the dimensionality of the resulting feature vector. It is usually set to 8.

Grid Y: the number of cells in the vertical direction. More the cells, finer the grid, the higher the dimensionality of the resulting feature vector. It is usually set to 8.

To train the algorithm we need to use dataset with several face images and we need to set an ID. Images of same person must have same ID.

The first step of LBP is to construct an intermediate image to better represent the original image

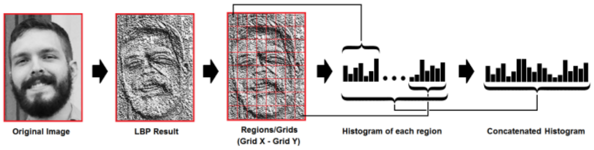


As seen in the above image if we have a grayscale image, we can take part of that image in a 3x3 pixels. It is the matrix that consist of intensity of each pixel (0-255).Then central value of the matrix is taken as threshold value. For each neighbor value of the central value we need to set new binary value.1 for values equal or higher than the threshold and 0 for values lower than the threshold.

Now the matrix will contain only the binary value ignoring the central value. Concatenate all the binary value from each position into a new binary value. Then convert that binary value to decimal value and set it as the central value of the matrix. At the end of the LBP procedure we get a new image with better characteristics of the original image

Now using the above image we can produce histograms. We can use the parameters grid X and grid Y to divide the image into multiple grids. Each histogram (from each grid) will contain only 256 positions (0~255) representing the occurrences of each pixel intensity. We should concatenate each histogram to form a larger histogram

Each histogram created is used to represent each image from the training dataset. So, given an input image, we should perform the steps again for this new image and create a histogram which represents the image. To find the image that matches the input image we just need to compare two histograms and return the image with the closest histogram. Several approaches can be used to compare the histograms (calculate the distance between two histograms), for example: **euclidean distance**, **chi-square**, **absolute value**, etc.



So the output of the algorithm is ID with the image with the closest histogram. The algorithm should also return the calculated distance

B.CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural networks are made up of neurons with learnable weights. Each neuron will accepts several inputs, calculates the sum and produces with an output**.** Convolutional Neural Networks have a different architecture than regular Neural Networks. Regular Neural Networks transform an input by putting it through a series of hidden layers. Every layer is made up of a set of neurons, where each layer is fully connected to all neurons in the layer before. Finally, there is a last fully connected layer called the output layer

Convolutional Neural Networks are different. The layers are organized in 3 dimensions: width, height and depth. The neurons in one layer do not connect to all the neurons in the next layer but only to a small region of it. And at last the final output will be reduced to a single vector of probability scores, organized along the depth dimension.

Four concepts in CNN are:

1.Convolution

2.ReLu

3.Pooling
4.Full connectedness

Feature Extraction: Convolution

Convolution in CNN is performed on an input image using a filter or a kernel. Filtering and convolution will involves with scanning the screen which starts from top left to right and moving down a bit after covering the width of the screen and repeating the same process until we scan the whole screen. The feature from the face of the individual is lined up with the image. The image pixel is multiplied by the corresponding feature pixel. The values are added and divided by total number of pixels in the feature.

Feature Extraction: Non-Linearity

After sliding our filter over the original image the output which we get is passed through another mathematical function which is called an activation function.

The activation function usually used in most cases in CNN feature extraction is ReLu which stands for Rectified Linear Unit. Which simply converts all of the negative values to 0 and keeps the positive values the same. The aim is to remove all the negative values from the convolution. All the positive values remain the same but the negative values changes to zero.

Feature Extraction: Pooling

After a convolution layer once you get the feature maps, we need to add a pooling or a sub-sampling layer in CNN layers. Similar to the Convolutional Layer, the Pooling layer is responsible for reducing the spatial size of the Convolved Feature. This is to decrease the computational power required to process the data through dimensionality reduction. Further, it is useful for extracting dominant features which are rotational and positional invariant, thus maintaining the process of effectively training of the model. Pooling shortens the training time and controls over-fitting.

Classification — Fully Connected Layer (FC Layer):

Now that we have converted our input image into a suitable form, we shall flatten the image into a column vector. The flattened output is fed to a feed-forward neural network and back propagation is applied to every iteration of training. The model is able to distinguish between dominating and certain low-level features in images and classify them using the Softmax Classification technique.

So now we have all the pieces required to build a CNN. Convolution, ReLU and Pooling. The output of max pooling is fed into the classifier we discussed initially which is usually a multi-layer perceptron layer. Usually in CNNs these layers are used more than once i.e. Convolution -> ReLU -> Max-Pool -> Convolution -> ReLU -> Max-Pool and so on. We won’t discuss the fully connected layer right now.

CONCLUSION

The result obtained from the proposed model gives the estimated sentiment prediction of the subject based on the video information. The resulting output can be used in many situations, the mental disorders and stress level is estimated and therefore in case of “critical” sentiments the peers and family members of the subject can take actions to encourage, motivate and uplift the emotional stature of the subject thus resulting in the harmony and peace of mind of the subject. Therefore such sentiment analyses models are are quirement for shaping the society into a happening place.

REFERENCES

[1] Facial Emotion Recognition System through Machine Learning approach, Renuka S. Deshmukh, Shilpa Paygude,Vandana jagtap.

[2] Emotion Detection Through Facial Feature Recognition, James Pao

[3] Renuka S. Deshmukh, Vandana Jagtap, Shilpa Paygude, “Facial Emotion Recognition System through Machine Learning approach” , 2017.

[4] Dongwei Lu, Zhiwei He, Xiao Li, Mingyu Gao, Yun Li, Ke Yin, “The Research of Elderly Care System Based on Video Image Processing Technology”, 2017.

[5] Shivam Gupta, ”Facial emotion recognition in real-time and static Images”, 2018.

[6] Ma Xiaoxi, Lin Weisi, Huang Dongyan, Dong Minghui, Haizhou Li, “Facial Emotion Recognition”, 2017.

[7] Mostafa Mohammadpour, Hossein Khaliliardali, Mohammad. M AlyanNezhadi, Seyyed Mohammad. R Hashemi, “Facial Emotion Recognition using Deep Convolutional Networks”, 2017.

[8] Dubey, M., Singh, P. L., ”Automatic Emotion Recognition Using Facial Expression: A Review,” International Research Journal of Engineering and Tech nology,2016

[9]https://stackoverflow.com/questions/42177658/how-to-switch-backend-with-keras-from-tensorflow-to-theano

[10]https://archive.ics.uci.edu/ml/datasets/Grammatical+Facial+Expressions

[11]https://www.edureka.co/blog/convolutional-neural-network/