March-April 2020
ISSN: 0193-4120 Page No. 18651 - 18656

Image based Emotion Aware Music Recommendation System using Machine Learning

[1]G. Niranjana, [2]Abhash Shrestha, [3]G Suseela, [4]Naresh Bohara

[1,2,4]Department of CSE, SRMIST, Kattankulathur
[3]Department of CSE, Saveetha School of Engineering, SIMATS, Chennai

Article Info
Volume 83
Page Number: 18651 - 18656
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 30 April 2020*

Abstract

Artificial Intelligence is one of the most prominent technologies of the modern world, it has been responsible for the revolution occurring in every major industry, as well as trade. However, it is not without its own set of challenges, and is not a 100% compatible with humans and their varied emotion. In this paper we propose how an AI system can detect and understand a human's emotions, mainly by detection of their facial expression. Human communication can happen in multiple levels, verbal, nonverbal, textual, pictorial, signs and so on. Facial expression, as we know, is a very explicit non-verbal mode of communication in humans. If a machine is capable of detecting and accurately classifying the particular emotion, it will undoubtably improve the human- machine interaction. In this paper we propose a system for lighting and position invariant recognition of facial expression. Information from the human face can be compared with dense model in an iterative manner. We will use a classifier to classify the images into different emotion categories. Variation in illumination, pose, distance from the camera etc. can influence the accuracy of facial recognition and emotion classification. According to the emotion, we can play a suitable musical track.

Keywords: *Emotion categories, expression recognition, Haar, Haar Cascade.*

1. Introduction

Emotion are and irrevocable part of human to human interaction, and plays a major role in everyday life. Emotion are an evolutionary trigger, for adaptive behavior that has helped humans survive the wild and unforgiving conditions that they had to deal with in the past.

Hence, if a machine is able to recognize the emotional state of a user, it would be highly beneficial for the interaction between the human and the machine, it would make the communication seamless and accurate, or even more pleasurable.

Every day, each and every person faces a lot of troubles and music can be a primary way of relieving their stress. Music is a form of art without language, any person could perceive emotions from a song of a language they do not know or speak. As this is the case, it is crucial for the system to be capable of detecting the user's emotion through their facial expression and recommend a suitable list of songs they might like at the particular moment.

Although there are many excellent music playing software in the market today, none of them are capable of recognizing them are capable of recognizing the user's emotion through their facial expression and recommend appropriate music, the user has to do this manually. So, in order to eliminate manual input step for the user, we propose such a software that recognizes the emotion of the user and recommends appropriate music. The emotion are broadly divided into 6 main categories: happy, sad, angry, neutral, surprised, disgust. The experiment with the model shows that detection of a happy face is the most accurate where as anger is the most inaccurate.

Haar cascades are an effective way of detecting objects, the can detect any objects such as watches, cars,

March-April 2020
ISSN: 0193-4120 Page No. 18651 - 18656

houses, and even faces. The Haar classifier is based on the Haar Wavelet Technique used to examine pixels in the image into squares by function. Haar cascades use the Ada-boost learning algorithm, which picks a limited number of crucial features from a large set to provide an accurate and reliable result, then use cascading techniques to detect face in an image. The Voila Jones detection algorithm is responsible for creation and theory behind the Haar cascade classifier.

2. Literature Survey

Artificial Intelligence is a revolutionary technology leading the way in technological advancements in today's world. However, affective computing has been a sort of a bottleneck in realizing emotional machines with advanced artificial intelligence capable of empathy with humans. With this the AI can take into account the emotions of the humans it serves while observing the surrounding environments to act for completing the goal of objective environment as well as the human user's emotion, which is subjective.

In a paper written by **[Chao Gong, Fuhong Lin*, Xianwei Zhou, Xing Lü]** they have proposed that an amygdala inspired affective computing framework can identify various types of human emotions. First, the neural network was able to precisely identify emotions with the help of inputs over a period of time, and not immediate inputs. The Convolutional Neural Network was compressed with the help of pruning as well as hashing tricks for even faster recognition of emergency emotions. According to the experimental results, it has shown a low level of latency and extremely high recognition accuracy. This is a major breakthrough in making A.I capable of understanding and empathizing with humans.

Human computer interaction can be greatly improved with the automatic estimation of human emotion. Continuous emotion tracking and recognition assigns an emotional value to every frame in a sequence. In another paper by **[Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Jiangyan Yi]** propose the use of ConvLSTM model, on the groundwork laid by 3D convolution in order to create and end to end continuous emotion system from video. The 2 D ConvNet, convolutions focus on spatial information, thus lose temporal information of the input signal. However, in 3D convolution and pooling operations have 3D kernel applied to overlapping 3D cubes spatiotemporally, and preserves the temporal information of the input signals resulting in an output volume. The system will use only one network to combine feature extraction and regressor into a unified system. Max as well as temporal pooling are explored in order to optimize the recognition system, The experiment results shows that max pooling can increase the efficiency of the system. Temporal pooling, however, achieves the most desirable results through exceptional performance. Temporal pooling also ends up saving a lot of training time as well as memory. ConvLSTM has shown a better efficiency and performance than 3D ConvNet particularly in arousal, proving its capability in modeling emotional spatiotemporal relationships. This greatly helps to amend the performance of continuous emotion recognition.

Emotions can have a wide range of variety, sometimes 2 different emotions might have almost similar facial features, so it is important to analyze the micro expressions in order to precisely classify the emotion. The emotion detection accuracy might vary depending on the angle of the person with respect to the camera. Normally such expressions might only occur for a very short time, they can be determined from the micro movements of the facial muscles. Micro expressions contain the true information about the human's emotion. In a body of work presented by **[Anna D. Sergeeva, Alexander V. Savin, Victoria A. Sablina, and Olga V. Melnik]** they have proposed the use of 2 algorithms. The first one is the Viola

– Jones [7] proposed in 2001 by Paul Viola and Michael Jones. In this method the Haar features are used in order to detect the face of a person, and detect their eyes as well. At present day, this method is the most optimal in terms of execution time and resource utilization. The second algorithm proposed was mainly dependent on the color and characteristics of the skin in the color space YCrCb and RGB. This method comprises of several image processing and filtering stages. According to the test results both the algorithms are useful for the preliminary processing in the image for detection of micro expressions. The algorithm based on color, lighting, intensity was deemed appropriate for color images and faster for small images. The Voila-Jones on the other hand can be applied to images in grayscale and for images of higher resolution.

For many computer vision algorithms classification of human emotions is a crucial and elusive task. Especially today, which is the age of humanoid robots and AI, and they co-exist with the humans. In a research done by **[Ivona Tautkute1, Tomasz Trzcinski, and Adam Bielski]** they propose a new way of emotion recognition that is dependent of incorporating facial landmarks as a crucial element of classification function. They have also used Deep Alignment Network (DAN) in order to achieve very accurate results in the recognition of facial landmarks. Their approach uses Deep Alignment Network architecture, which was originally proposed for robust face alignment. Its main advantage is that it uses iterative process in comparison to its contemporary methods, in order to adjust the location of facial landmarks. The iterations are included into the Neural network architecture for learning, and the knowledge acquired is passed onto the next stages through the use of facial heatmaps. It can be said that DAN can deal with entire facial images and not pixels or patches. DAN is currently ranked third in terms of facial landmark recognition leaderboard. In this body of work the researchers employed their recognition model as a part of the in car analytics system to be used in self-driving cars

March-April 2020
ISSN: 0193-4120 Page No. 18651 - 18656

or vehicles. A self-driving car's operation can be altered with the emotion of the passenger, for e.g.: fear of speed detected on the passenger's face. This is an ongoing method of emotion recognition that enables the utilization of facial landmarks. This research uses the JAFFE datasets in order to produce the results and shows that there are still improvements to be made. However, this method has been said to have the potential to outperform the existing or other proposed methods. This method can be improved with the incorporation of attention mechanism on facial landmarks.

In most of the cases and model we see the face detection system's accuracy is greatly reduced if the user is wearing anything on the head or the face such as a big headphone, or sunglasses, or even normal glasses. So, in order to overcome this deficiency, in a work proposed by,**[Hwanmoo Yong, Jisuk Lee, and Jongeun Choi,]** they trained the Convolutional Neural Network to detect and classify the emotion of a human wearing a Head Mounted Gear. This system does not take into account the eyes and the eyebrows of the users and solely focuses on the remaining parts such as the mouth and other facial muscles on the forehead and the cheeks. Apparently, there has be no prior research in this field. They used the Roundabout Face Datasets, which consists of 8040 images, with 8 emotional expressions, happy, sad, angry, neutral, disgust, surprise, fear, contempt. Each emotion was shown with 3 different gaze directions and 5 varying angles of the camera. The purposefully cropped the original images around the face and added black rectangles around the eyes and the eyebrows in order to represent Head Mounted Gears. Afterwards the 8040 images were jumbled and splitted into 3 datasets, 5640 for training purpose, 1200 for the validation process and the remaining 1200 images for the test. The result of this experiment was that, 3 CNN were successfully trailed to estimate the emotion from a partially covered human face. From the three DenseNet performed better than ResNet and Inception ResNet V2. However, for the recognition of fear and disgust ResNet was more effective. It can be concluded that the CNN was able to extract features from the lower parts of the face as well, which are known to be less representative of a person's emotion. The CNN was able to predict Happy, Surprised, Disgusted, Fearful emotions better as expression of these emotions require the movement of mouth, while the eyes and eyebrows are used more impactfully in the representation of the remaining emotions.

Emotion recognition is natural trait of human beings which is an area of interest for researchers. Speech is a part containing carrying human emotions or state of mind. Speech is a mixture of utterances.A paper prepared by **[Gustavo Assuncão ,Paulo Menezes and Fernando Perdigão]** analyses the human speech and extracts emotion from it in real-time. In this demonstration a real time emotion recognition system was proposed which extracts speech features to build state-of-art classifiers. This paper has added more interest to the researchers working on the extraction of human based emotions using advanced technology.

Basically, emotions are categorised into six types anger, happiness, surprise, disgust, fear, sadness and neutral. Extraction of emotions is a very vast concept naturally possessed by humans. It is very hard for the machines to determine the emotional state of humans. For this human have attempted to build Speech Recognition System (SER) that extracts emotions based on the properties it is build. These systems are proved to be fruitful in teaching and learning technologies, medical services etc. This reduces the deployment of human resources and expenditure brought along with it.

In this paper proposed by **[Surekha Reddy B, T. Kishore Kumar]** determining of emotions is done using Teager Energy Operator(TEO) and Linear Prediction Coefficient(LPC).

This paper focuses in Stressed Speech Emotion Recognition (SSER).Gaussian Mixture Model (G MM) classifier is used for categorization of emotions. Stressed emotions anger, fear, disgust and sadness are classified taking neutral as the reference. The accuracy observed was 82.7% and 88% which are higher than the previous systems. But system observed some difficulty in recognizing the fear emotion. So, some improvements are to be done the system.

Music is a form of art that carries emotions. They are like rivers that flows in different directions. Everyone has lows and highs in daily life. To rejoice and rejuvenate music is a reliable companion. It would be very effective and handy if Music players would play according to our mood. They are also helpful for the transition of emotions also. It is also stress reliever. This Emotion based player plays songs according to the accordance to the person's mood. Recognizing human emotions is a very difficult task because there is variations in emotions of human. Many things come into act while bringing human emotions as genre, pitch, amplitude etc are also needed. Many works needed to be done like classifiers need to be built, visualization of musical features and finally mapping the features and recommending songs.

This paper proposed by **[S. Deebika, K. A. Indira, Dr. Jesline]** focuses on music player based on emotions of the listener. It uses Convolutional Neural Network (CNN) determining the human neural conditions. Although many algorithms are there but their result is not as predicted, CNN overcomes such gap between them. First, they train the system to make the visualization more informative. Then, Music and songs are classified by Support Vector Machine (SVM). Finally, recommends to the user. This paper enhances the accuracy and efficiency of previously existing systems. Songs are classified using different filters. The basic purpose of the system is to alter or prolong the existing emotional state. This automation is used for differently abled people. This concept can increase the efficiency and correctness of the system.

Speech has been the most effective way of communication between human beings. As huge

March-April 2020
ISSN: 0193-4120 Page No. 18651 - 18656

advancement took place in machines years ago but human-machine interaction has been one of the hardest thing. There is variation between voice , accent etc which makes it even hard to understand human language. Many systems have developed so far but they do produce result as estimated. Many algorithms like Pattern Recognition Neural Network (PRNN) and K-Nearest Neighbour (KNN) are used in the previous systems.

The paper proposed by **[J. Umamaheswari, A.Akila]** depicts the use of hybrid algorithm of Pattern Recognition Neural Network (PRNN) and K-Nearest Neighbour (KNN).

Mel Frequency Cepstral Coefficient (MFCC) and Gray Level Co-occurrence Matrix (GLCM) were used in this system for feature extraction. Weiner filter was used for filtering the noise in speech. Feature extraction is done from the emotional database containing emotional classes Angry, Happy, Sad, Neutral, Surprise and Fear. Many visualization methods for accuracy and precision were taken into consideration and various tables and columns were prepared. Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) like speech recognition techniques were used in the system.

World population is growing in a very rapid manner. With time there are people who needs to be cared like children, elderly people, differently abled people. In this busy world taking care of these age groups has become challenging. But kudos to automation and robotics many humanoid service robots are developed worldwide. But as we know there is a huge communication gap between humans and robots. There is limitation to the communication of robots. Many easy and feasible ways are already developed face recognition is one them. To track the visual gestures of the humans robots need to be trained. If face recognition can make human- robot communication more of human-human interaction then it is feasible to serve more effectively. Humanoid robots are the most preferable type of care givers.

This paper proposed by **[T. M. W. Vithanawasam and B.G. D. A. Madhusanka]** is centered on the idea of face recognition and upper body recognition. According to research

,55% of interaction is body language. For robots eye model web cam and other high resolution cameras were used and focused on the Region of Interest (ROI) to get the result.

Face detection used Haar-cascade classifier used because of its high-level features. OpenCV is platform used for face detection. But for face detection many constraints should be fulfilled like looking into camera etc.

For upper body detection, visual feed of upper- body ROI. The proportionality of the body can be using head's height and head's width. To get more accurate features and emotions huge datasets for every emotion is required. Many obstacles like lighting, position was taken into consideration to build an emotional recognition system. Many emotional expressions like fear, anger was recognized by the system. Despite of many hardships it can be used for the references of more advanced systems.

3. Proposed Work

In the work we have done, we have used the Haar Cascade Algorithm. Object detection using harr cascade classifiers is one of the most accurate ways if effective object detection, this was originally proposed by Paul Voila and Michael Jones in their paper titled" Rapid object detection using a boosted cascade of simple features". It is a method that utilized machine learning where a cascade functions are trained to create a collection of positive as well as negative images. Positive images are images that contain the object that we would like the machine to detect and negative images are those that do not contain the object of desire. After such training is complete, it can be used to detect that particular object on other images and even real time scenarios, similar to our work.

In the Haar algorithm we use, it needs many positive images and negative images first so they can be used to train the classifier. After that comes feature extraction from it. The image is considered in terms of pixels, each feature is a single value obtained by subtracting the sum of pixels within a particular(focused) area from the sum of pixels under the remaining area in the image. This can be simplified as such; the feature value is obtained by subtraction of sum of pixels covered in white from the sum of pixels in black.

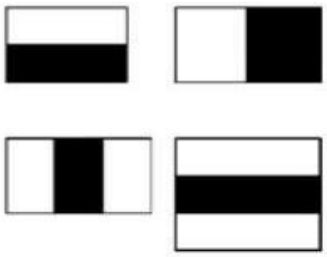

Figure 1: Haar cascade training images

With this method the algorithm can extract enough features. However simply doing this with every pixel would mean a huge amount of computing creating an overload. For e.g.: a 24 x 24 window would result in over 160000 features which would be too much for the computer and also it would take an enormous amount of time to do the same for every single image. So in order to solve this the concept of integrated images was used. Regardless of the number of pixels it will simplify the calculation which would only involve 4 pixels. This would make the whole process efficient as well as faster. But how is the system supposed to select the best features out of thousands of features in the image? This can be done with the help of AdaBoost.

AdaBoost, short for Adaptive Boosting, is a machine learning meta-algorithm. It can be used along with other kinds of learning algorithms to increase the overall

March-April 2020
ISSN: 0193-4120 Page No. 18651 - 18656

efficiency and accuracy. The AdaBoost algorithm is adaptive as the subsequent weal learners are tweaked in favor of those instances misclassified by previous classifiers.

In the system we are using, it determines the best threshold which can sort the images, face images in this case, into positive and negative. There might be many misclassifications and errors, so we will choose features with the least rate of error. This means that they are the most reliable features in order to sort a face image from a non-face image. The final classification is a sum of the weak classifiers. Weak classifiers are those that alone cannot classify the image, but they can with the help of other strong classifiers. So, in theory, we can take an image of 24x24 window an apply a set of 6000 features, and determine if it is a face or not. However, this would take too much time, so we have a crafty solution for this. As they are lot of non-face regions in the images, and when they're detected it is just more efficient to discard them totally and focus on other remaining regions. By doing this no time is wasted on the analysis of a non-face region. So, in order to do this the concept of Cascade classifiers was brought in. Instead of applying all the 6000 features on a window, the features will be sorted into a number of stages and classification and detection is performed on them one-by-one. If the process is not successful at the first phase it is discarded and the remaining features on it are not considered at all. For example: if no eye is detected then the remaining features are not checked and are directly discarded. However if the initial phase is successful, the second stage of features is applied and the process goes on. If a data(image) is able to pass all the phases it is classified as a positive image i.e the object we are trying to detect, a face in this case.

In this work OpenCv is used in order to detect the face as well as the classification of emotion of the person's face. OpenCV contains many pre-trained classifiers in order to detect face through features such as the eyes, mouth etc. These are found in the from of XML files, the one we are using to detect the face is called the "haarcascade_frontalface_default.xml". The first task to perform is to load the xml classifier, and then load our input images, which are taken in real-time through the webcam, into grayscale.

The code for this is as such:

```
import cv2, sys, numpy, os
haar_file = 'haarcascade_frontalface_default.xml'
datasets = 'datasets'  #All the faces data will be present this folder
sub_data = 'neutral'
```

After this initial phase we have to use the webcam in order to record the persons face, the code we have used for this purpose is as such:

```
while count < 100:
    for i in range(0,2):
        if i==0:
            ##WEBCAM
            (_, im) = webcam.read()
```

After recording the images in real-time they need to be converted into grayscale and many functions that we use in OpenCV are compatible with images in grayscale. We also wanted to draw a rectangle around the face of the person in real-time so that they would know that their face is being detected. The code for this is as such:

```
gray = cv2.cvtColor(im, cv2.COLOR_BGR2GRAY)
faces = face_cascade.detectMultiScale(gray, 1.3, 4)
for (x,y,w,h) in faces:
    cv2.rectangle(im,(x,y),(x+w,y+h),(255,0,0),2)
    face = gray[y:y + h, x:x + w]
    face_resize = cv2.resize(face, (width, height))
    font = cv2.FONT_HERSHEY_SIMPLEX
    cv2.putText(im, 'F A C E',(x+5,y-10), font, 0.5, (0, 255, 0), 2, cv2.LINE_AA)
```

4. Proposed Architecture

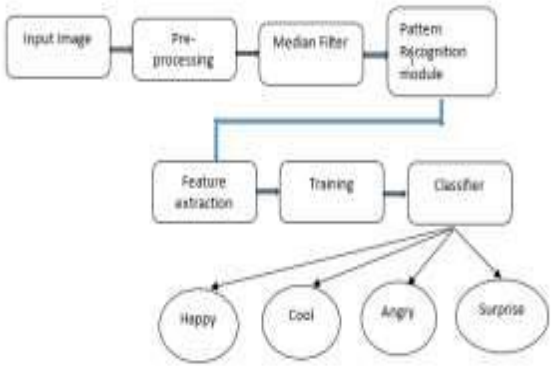

Figure 2: Architecture diagram

Dataset Images

Dataset images are required for the reference and comaprison to determine the emotion state of the person .It compares the captured image.

Table 1: Features used in mood classification [10]

Feature1	Feature2	Feature3
Passionate	Confident	Fiery
Brooding	Cheerful	Aggressive
Tense	anxious	visceral
Rousing	Boisterous	Good natured

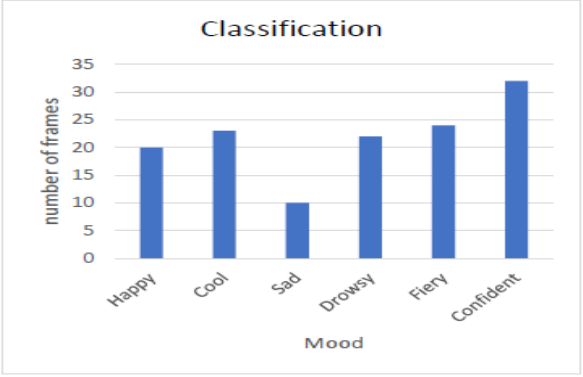

Figure 3: Experimental Results

March-April 2020
ISSN: 0193-4120 Page No. 18651 - 18656

Pre-processing

Image processing for face detection is done in this step. Image is captured in grayscale. Color- based facial image processing and analysis is carried out. De-blurring and super-resolution for accurate face detection.

Median filter

In this step noise, hue act present in image are cleared. Typically it is noise reduction technique. It is used in digital image processing as it refines the image from unwanted features.

Pattern recognition

Every human has similar facial features like nose, ear, mouth, head etc because of which it is easy to reference ones image with other person. As the same pattern is found in every human face differentiation of human face from other objects is feasible

Feature Extraction

Histogram is used to show percentage of emotional feature present in the human face. It shows different emotions present in face using Haar cascade algorithm.

Training

The harr files are trained using a lot of positive and negative images. The Haar XML files will contain the feature sets that when used along with OpenCV can detect and determine the face of a person.

Classifier

It classifies images in six different categories anger, fear, happy, surprise, neutral and sad. Haar classifier is trained to work as to classify the image into any one of the categories.

5. Conclusion

Emotions are an inseparable human interaction. Training a machine to recognize and take the emotion as input can greatly influence the human-computer interaction efficiency making the use of computers reliable. As we know, Music can greatly influence a persons emotional state. In the experiment we have used Haar-cascade algorithm in order to detect and classify the users emotion and according to this we can recommend the appropriate musical tracks

References

[1] Chao Gong, Fuhong Lin*, Xianwei Zhou, XingLü, ”Amygdala-Inspired Affective Computing: to Realize Personalized Intracranial Emotions with Accurately Observed External Emotions**”,** School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

[2] Jian Huang ; Ya Li ; Jianhua Tao ; Zheng Lian ; Jiangyan Y, “End-to-End Continuous Emotion Recognition from Video Using 3D Convlstm Networks”, IEEE

[3] Anna D. Sergeeva, Alexander V. Savin, Victoria A. Sablina, and Olga V. Melnik, “Emotion Recognition from Micro- Expressions Search for the Face and Eyes”, Department of Electronic Computers and Department of Information-Measuring and Biomedical Engineering, RSREU Ryazan, Russia

[4] Ivona Tautkute1, Tomasz Trzcinski, and Adam Bielski, “I Know How You Feel: Emotion Recognition with Facial Landmarks”, Tooploox Polish-Japanese Academy of Information Technology 3Warsaw University of Technology

[5] Hwanmoo Yong, Jisuk Lee, and Jongeun Choi, “Emotion Recognition in Gamers Wearing Head-mounted Display”, Yonsei University School of Mechanical Engineering

[6] Gustavo Assunc¸, Paulo Menezes, Fernando Perdig, “Importance of speaker specific speech features for emotion recognition”, Dept. of Electrical & Computer Eng. University of Coimbra Coimbra, Portugal

[7] Surekha Reddy B. T. Kishore Kumar, Emotion Recognition of Stressed Speech using Teager Energy and Linear Prediction Features”, ECE Department, NIT Warangal, India

[8] S. Deebika, K. A. Indira, Dr. Jesline, “A Machine Learning Based Music Player by Detecting motions”, St. Joseph’s College of Engineering Chennai, Tamilnadu, India

[9] J.Umamaheswari, A.Akila, “An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN”, Department of Computer Science, School of Computer Science, Vels Institute of Science Technology & Advance Studies, Chennai-6000117, Tamilnadu, India.

[10] T. M. W. Vithanawasam, B. G. D. A. Madhusanka, “Face and Upper-Body Emotion Recognition Using Service Robot’s Eyes in a Domestic Environment”, Department of Mechanical Engineering The Open University of Sri Lanka Nugegoda, Sri Lanka

[11] Patra, B. G., Das, D., & Bandyopadhyay, S. (2013). Automatic music mood classification of hindi songs. Proceedings of the 3rd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2013), pp. 24-28.