Implementation of Convolutional Neural Network to Realize a Real Time Emotion based Music Player

**P.H. Abhirami, Elizabath Saba, Regina Mathew,
E. Jacob Sebastian, Cerene Mariam Abraham**

Abstract--- The ability to represent the world as a nested hierarchy of concepts, by defining each concept in relation to abstract representations has promoted deep learning to be widely used as a processing model for solving data science tasks. The era of digitalization has allowed the deep learning technology to flourish and machines with the ability to analyse huge amount of complex data would now be able to give progressively exact outcomes due to its supremacy in terms of accuracy when trained with massive amount of data. Convolutional Neural Networks(CNN), being a deep neural network with their ability to develop an internal representation of a two-dimensional image, allows the model to learn position and scale invariant structures in the data, which is important when working with images. For realizing emotion aware applications, the system must be highly accurate and in real time. In this paper, we provide the design and implementation details of a real time emotion based music player using CNN with the aim to reduce human effort and invoke the feasibility of Human Computer interaction(HCI).

Keywords--- Convolutional Neural Networks, Deep Learning, Face Detection, Emotion Detection.

I. INTRODUCTION

The focus of current research lies within the evolution of computer programs adapting at the exposure to new data. Deep learning being a part of broader aspect of machine learning, uses computer algorithms to create autonomous learning from data and information.

It has been the area of active research that aims to evaluate its function and to illuminate how its methods are affecting traditional approaches of machine learning. Feature learning and scalability are two known benefits of deep learning methods.

Learning categories incrementally by a deep learning technique is through a hidden layer architecture and this eliminates the need of domain expertise and hardcore feature extraction. A major advantage with deep learning and a key part in understanding its popularity is that it's fuelled by massive amount of data, therefore the number of opportunities for new innovations in the field is advancing with the technology.

The requirements of deep learning thus involve substantial computing power and large amount of labelled data. Applications of deep learning from the industrial field ranges from automated driving to medical devices.

A standout amongst the most prevalent types of deep learning techniques include Convolutional neural networks(CNN). A CNN convolves learned features with the input data and makes use of 2D convolutional layers, making this architecture appropriate to handling 2D data, like images. A CNN uses a framework much like a multilayer perceptron that has been intended for reduced processing requirements. As CNN extracts features directly from images, the need for manual feature extraction is eliminated. The relevant features are not pre-trained, instead they are learned while the network trains on a collection of images. The detection of different features of an image is learnt by CNN using tens or hundreds of hidden layers each of which increases the complexity of the learnt image features. This automated feature extraction makes models of deep learning highly accurate for computer vision tasks including object classification. The application of relevant features enables CNN to successfully capture the spatial and temporal dependencies in an image. Due to reduction in the number of parameters involved and re-usability of weights, the architecture performs a better fitting to the image dataset; hence the network can be trained to comprehend the refinement of image better.

Human emotions which constitute a major part of our non-verbal communication are discrete and consistent responses to internal or external events. Emotions have the ability to decide how we think and behave. The extensive possibilities of applications are challenging in the field of computer science. Social intelligence in machines can be incorporated be automatic detection of human emotions. Researches that address the emotions and affective states requires construction of adequate empirically proven learning strategies to respond effectively to detected individual emotions. In this paper,4 emotions are considered for classification, which includes happy, anger, sad and neutral. Extraction of facial features plays the crucial role in identifying each emotion and it works on facial feature deformations. The emotions are recognized based on visual information by the system and it may not be the sole indicator of the same. Finding regularities in the data set being analysed and pattern recognition are the essential parts of facial emotion recognition.

The availability of appropriate datasets for training is an

Manuscript received September 16, 2019.

P.H. Abhirami, Department of Computer Science and Engineering, Muthoot Institute of Technology and Science, Kochi, Kerala. India

Elizabath Saba, Department of Computer Science and Engineering, Muthoot Institute of Technology and Science, Kochi, Kerala. India

Regina Mathew, Department of Computer Science and Engineering, Muthoot Institute of Technology and Science, Kochi, Kerala. India

E. Jacob Sebastian, Department of Computer Science and Engineering, Muthoot Institute of Technology and Science, Kochi, Kerala. India

Cerene Mariam Abraham, Department of Computer Science and Engineering, Muthoot Institute of Technology and Science, Kochi, Kerala.India

essential requirement for the identification of emotions by applying a deep learning algorithm. The datasets used for the work includes Extended Cohn Kanade(CK+), Kaggle facial expression dataset, IMM face dataset and JAFFE. After exhaustive training with these prelabelled datasets, CNN is used for feature extraction and emotion detection. To build a robust system that recognizes basic emotions, various tasks are required to be carried out including face detection, facial component detection, emotion extraction and detection. Factors including pose and lighting conditions contribute to the challenges faced in designing this system. The emotion detection module in the work is followed by song classification wherein the detected emotion is used to find the most suitable song corresponding to that emotion. The music player works in real time and reduces the task of manually browsing through a playlist to select a song matching the person's current mood thereby giving better user experience.

The system implemented in the work includes 4 main modules, namely- face detection, feature extraction, emotion detection and song classification. Viola Jones algorithm is used for face detection due to its low false positive detection rates and robustness. The feature extraction module and emotion detection is handled using CNN where the input to this module is the Region of interest(ROI) of face detected. Finally, the emotion that has been recognized is fed as input into the music player which plays songs accordingly.

II. RELATED WORKS

Numerous algorithms have been developed by researchers for the construction of an efficient face detection system. Out of these, the most commonly used and proven technology is the Viola Jones algorithm proposed by Paul Viola and Michael Jones and the same is used in the face detection module of this work due to its known benefits. Shakhnarovich et al. in their paper proposed a system using Viola jones face detection algorithm which is used for robust, real time and integrated face detection [1].

The detected region of interest is passed to a demographic classifier which finds the gender and ethnicity. They proposed a system which is extremely fast. Castrilln et al. in their paper presents a comparison of many public domain classifiers in openCV for detecting faces which are trained on different conditions [2]. Face detection from static images and video streams are different. Temporal coherence is also taken into account while creating the real time model proposed by Castrilln .M et.al in their paper [3].

To use viola jones algorithm for face detection in mobile platforms, various optimization techniques are needed due to relatively limited processing and memory capabilities. Window shift approach is used for finding multiple faces at different resolutions from a video streams. The combination of different cues available in video streams due to temporal coherence is window based approach. It also finds application in human machine interface where the human face is detected and tracked throughout the interaction. Viola jones algorithm is used for detection and eigen images and PCA used for recognition [4].

Joint haar like features are used to detect face from an image. Joint haar like feature is the co-occurrence of multiple haar like features. Stage wise selection of joint haar like feature using adaboost is used for face detection [5]. Cascade architecture built on convolutional neural networks (CNNs) model is used to address the challenges such as the visual variations in image due to the pose, expression and lightning conditions [6].

After each cascaded detection stage, a CNN based caliberation stage is introduced. For accurate human detection, haar feature cascade and Histogram of gradients features are integrated [7]. This automatically detects the salient features of face. For face detection another approach can be used ie, to find the skin and non-skin pixels. This is the hybrid skin color model.

Feature extraction plays a major role in analysing the emotion of a person. When it comes to dealing with spatio-temporal characteristics of image content, advanced feature selection methods are necessary [8]. Convolutional neural networks are used for feature extraction in this work because it learns features directly and manual extraction is not required.

Fan, Yin et al. in their paper proposed a system where features are being extracted by convolutional neural network[11]. RNN takes these over individual video frames as input and encodes motion later, while C3D models appearance and motion of video simultaneously describes a rule-based algorithm for robust facial expression recognition combined with robust face detection using a convolutional neural network [9].

CNN model is initialized randomly and pre-trained on a larger dataset. To combine multiple CNN models, two schemes for learning the weights of the network responses: by minimizing the log likelihood loss, and by minimizing the hinge loss are used [10]. Simplifies the problem domain by removing confounding factors from the input images, with an emphasis on image illumination variations. This, is an effort to reduce the amount of data required to effectively train deep CNN models. A novel method to extract features from visual and textual modalities using deep convolutional neural networks [12].

Feeds these features to a multiple kernel learning classifier and perform emotion recognition and sentiment analysis on different datasets. Starting from a network pretrained on the generic ImageNet dataset, supervised fine-tuning on the network in a two-stage process is performed, first on datasets relevant to facial expressions, followed by the contest's dataset [13]. Network consists of two convolutional layers each followed by max pooling and then four Inception layers. The network is a single component architecture that takes registered facial images as the input and classifies them into either of the six basic or the neutral expressions [14].

A zero-bias CNN on facial expression data is trained and achieved, to our knowledge, state-of-the-art performance on two expression recognition benchmarks: the extended Cohn-Kanade (CK+) dataset and the Toronto Face Dataset (TFD). RNNs provide an attractive framework for propagating information over a sequence using a continuous valued hidden layer representation [15]. Hybrid CNN-RNN

architecture for facial expression analysis can outperform a CNN approach using temporal averaging for aggregation.

The audio-video based emotion recognition is based on the Acted Facial Expressions in the Wild (AFEW) database [16]. The group-based emotion recognition is based on the Happy People Images(HAPPEI) database. 3DConvolutional Neural Networks (CNN) and recurrent neural network is being used for this [19].

III. PROPOSED SYSTEM

3.1 Problem Statement

To design and implement a real time music player which plays music according to the emotion of the user, using convolutional neural network. This article aims:

• To provide a solution for the substantial method of using a music player by incorporating emotion detection.
• To learn in detail compare emotion detection using different deep learning techniques.
• To design and implement a reliable CNN classifier to distinguish between different facial expressions.
• To identify the conditions under which the realization of an application for emotion detection can lead to improvements in subjective and/or objective measures of system usability.

3.2 Proposed Solution

As an implementation of convolutional neural network for facial expression recognition, a music player is constructed which plays songs according to the emotion of the user. The system is divided into different modules and each performs specific tasks in an operating sequence. At first, the system detects the face of the user and captures the region of interest of the image. The face detection is done using Viola Jones Algorithm.

The captured image is input to CNN which learns features directly. The features are analysed to determine the current emotion of the user. Each emotion detected will be mapped to the music player which plays corresponding music automatically. The designed music player overcomes the effort associated with manual selection of songs from a playlist, especially if there are more number of songs. Hence the system opts for better user experience by invoking human computer interaction.

IV. METHODOLOGY

4.1 Face Detection Module

Detection of face from a given input image or video is facedetection. There are various algorithms for face detection. Viola jones algorithm is used for face detection. The main steps in Viola Jones algorithm are:

HAAR feature - HAAR features represent some characteristics of the face. Haar features are similar to those convolution kernals which are used to detect the presence of the feature in the given image. Each feature result in a single value which is calculated by subtracting the sum of pixels under white rectangle from the sum of pixels under black rectangle. In the feature, the black region are replaced by plus ones and white region is minus one.

Integral image - In haar feature calculation, as every time window moves need to sum up all pixels of the black region and those of white region. It is a tedious operation and the solution is integral image. It reduces the computation rather than summing up all pixels under a rectangle with just four corner values of the integral image. To find the value of any pixel just sum the values of pixels to the top and left.

Adaboost - Viola jones algorithm makes use of 24*24 window as the base window for starting the evaluation of features in any given image. If we consider all possible parameters of haar features like position scale and type we end up calculating 160,000+ features in this window which is practically impossible .So the basic idea is to eliminate a lot of features which are redundant or which are not useful and select only the features that are very useful .This one by Adaboost eliminating 160 thousand features and narrowing down to only couple of thousands of features which we need to evaluate. The features extracted by Adaboost is weak classifiers. Adaboost constructs a linear combination of the weak classifier.

Cascading - It is the basic principle of Viola Jones face detection algorithm to scan the detector many times through that image itself, each time with a new size. Though an image should contain one or more faces it is clear that an excessive large amount of the evaluated sub-windows might still be negatives. He algorithm should hence concentrate on discarding non-faces quickly. Therefore a single strong classifier formed out of linear combination of all the best features is not good to evaluate on each window because of computation cost.

4.2 Facial Feature Extraction Module

For feature extraction, CNN is used. For the emotion recognition module, we have to train the system using datasets containing images of happy, anger, sad and neutral emotions. In order to identify features from dataset images for the model construction, CNN has the special capability of automatic learning. In other words, CNN can learn features by itself. CNN has the ability to develop an internal representation of a two dimensional image.

This is represented as a three dimensional matrix and operations are done on this matrix for training and testing. Moreover, in some other neural networks like fully connected networks, all nodes in a layer are connected to every node in the next layer.

Associated with each connection, weights are present. This will increase the computational complexity. But in CNN, nodes in a layer are connected to only valid nodes in the next layer. Thus there will be less computational complexity.

This includes various layers in order to train and test input images. Final layer is fully connected which have classification task and thus images can be classified according to emotions.

The emotions detected should be one these labels: angry, happy, sad, and neutral. Before entering the CNN, entire dataset will be divided into two. 80 percent of this will be

used for training and remaining 20 percent will be used for testing.

After training images using CNN, a model will be created [17]. Then testing will be done on this model. During the testing time, accuracy can be calculated by checking whether the images are classified correctly or not.

Accuracy can be increased by increasing the number of epochs or by increasing the number of images in dataset.

The input will be given to convolution layer of the neural network. The process that happens at convolution layer is filtering. Filtering is the math behind matching. First step here is to line up the feature and image patch. Then multiply each image pixel by the corresponding feature pixel. Add them up and divide bythe total number of pixels in the feature. In convolution, one image changes to stack of images after filtering.

ReLU, tanh or sigmoid functions can be used for non-linear operations [18]. In ReLU, every negative value is converted to zero and positive value will not have any change. Leaky ReLU allows small gradient when the unit is not active. Thus information will not be missed. This also fixes the dying ReLU problem. ReLU is better than other two non-linear functions.

The next layer after convolution layer will be pooling layer. This is to shrink the image stack given by convolution layer.This will reduce the number of parameters [20].Here, a window size is picked (usually 2 or 3). Then pick a stride (usually 2).

The number of pixels shifts over the input matrix is a stride. When the stride is 1, we move the filters to 1 pixel at a time and when the stride becomes 2, we move the filters to 2 pixels at a time and so on. Then the window is moved across the filtered images.

There are three different types of pooling. They are max pooling, average pooling and sum pooling. From a rectified feature map, max pooling take the largest element. Taking the average of elements from the window is average pooling. Sum of all elements in the feature map is known as sum pooling.

More numbers of convolution and pooling layers can be added till the required accuracy is attained. After pooling layer, the matrix is being flattened to vector and fed to fully connected layer. The aim of flattening is to transform a two-dimensional matrix of features into a vector of features that is fed into a neural network or classifier. With the fully connected layers, the vector elements combine together to form models. Finally, classification is done by activation functions like Softmax or Sigmoid function.

4.3 Emotion Detection and Song Classification Module

The output of neural network classifier is one of the four emotion labels: happy, anger, sad and neutral. HTML pages with user interface for each emotion are designed for the system in such a way that once the emotion of user is identified, playlist corresponding to that emotion will be displayed in the screen. The first song in the playlist of the page displayed will be played first. Songs are selected such that it reflects the mood of the user.

V. IMPLEMENTATION

For implementing Viola Jones face detection, OpenCV's HAAR Cascades is used to detect faces. Object detection using Haar feature-based cascade classifiers is an effective object detection method.

The face detection can be done in two ways. First from the static images and the second from the dynamic or web cam feed. OpenCV's HAAR Cascades is used for both of them. OpenCV's HAAR Cascades is pre-trained classifiers to detect faces. For static face detection, first we need to load the required XML classifiers, here need only haarcascade_frontalfacedefault.xml file.

For feature extraction, to detect the face the training data used is stored in the XML files. This training data is used to best identify the features that it can consider a face. Then load our input image in grayscale mode. Using detect Multiscale function, face is detected now we find the faces in the image. If faces are found, it returns the positions of detected faces as Rect(x,y,w,h). Once we get these locations, we can create a ROI for the face.

For real time face detection, first the image is sent through the Dlib face detector library, if it didn't detect the face, it is passed through the OpenCV's HAAR Cascades. All the face detectors are set initially ie, both the dlib detector and HAAR detector.

As the image will be in varying resolutions that depends on the device. So the resolution must be changed for the images. The resolution is changed to 380x380 using resize function. The resulting images are converted to grayscale using cvtColor function with parameter color_bgr2gray. The Dlib detector of python finds faces almost all the time. It is based on HOG and linear SVM.

Face detection works by computing the HOG feature and classifying them with SVM.HOG feature descriptor is used to characterize objects through shapes. If the face is detected, the resulting region of interest is stored. If face is not found, then the HAAR cascade is used to detect the face.

For detecting the face detectMutltiscale function is used with parameters as the given image, the scale factor, minimum neighbours needed, minimum and maximum size allowed. This is looped four times as there are four xml files for feature extraction to detect the face. It is looped through these files to detect the face, if face is found the resulting region of interest is stored.

In the emotion detection module, we use the code of CNN. First all the packages that are required throughout the module are imported. Then the images from dataset are read from file. The images are then converted to readable arrays.

This is done by using convert_numpy() function. We give image as argument to this function. This 2D matrix is to be converted to 3D matrix. The third dimension is for representing the depth. Depth implies color intensity of that particular image.

Here in this work, depth of the image is 1 since the input is a grey scale image. Next step is to convert pixel values from int8 to float32 format. This is done for getting 32 bit precision. For scaling the pixel values, it is then divided by

255. 255 is the maximum value of a byte. By dividing by 255, we can ensure that the value is scaled between 0.0 and 1.0. Thus dataset easily fits into RAM.

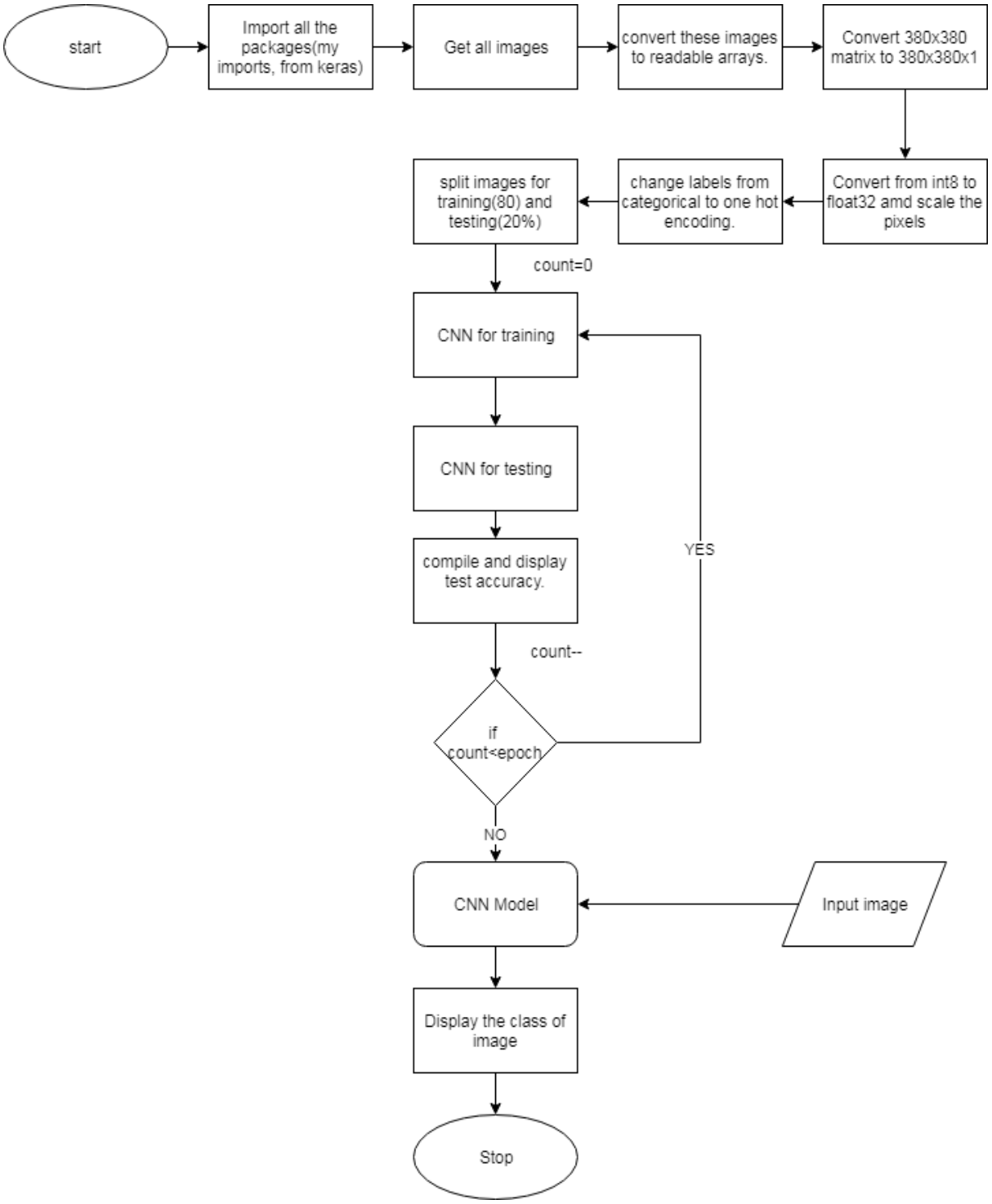

Fig 1: Flowchart of the process flow

The labels have to be converted from categorical to one-hot encoding. In categorical, the values hold real numbers. Here there is more chance for errors, while doing operations using these numbers for classification purpose. In case of one-hot encoding, the values used are 0s and 1s. This will reduce the misclassification error.

The 3D images have to be split into two. That is, 80% for training and remaining 20% for testing. First image from the train set is taken for CNN training and then the model is tested using image from test set. The result obtained is then compared with the true label and accuracy is calculated. This process is continued until number of epochs is completed or expected accuracy is met. After training and testing, the model is created and we input our new image whose class is unknown. The class of the image is then predicted by the model. The first layer of CNN is convolution layer. This layer compares the input images with certain features known as filters and a stack of images are created after this layer.CNN will automatically detect important features without any human supervision. leakyReLu function is used to normalize the matrix. The

values are set in between 0.0 and 1.0. The constant in the argument to leakyReLu multiplies with every single value and prevents dying ReLu problem. Thus information loss can be reduced.

Next layer is pooling layer. This is to reduce the size of the matrix. Max pooling is used here in order to prevent over fitting. Inorder to reduce overfitting, we make use of Dropout() function.

Random neurons get rejected during the training or they are "dropped" out randomly. Flatten and dense layers are used to flattens the matrix to a simple array and apply Softmax function on that layer.

After this we will get a probability representation of classes for the particular input image. The code is compiled using cnn_compile() function to which loss function is given as argument in order to calculate error so that weight can be updated. Emotion detected using convolutional neural network will be fed as input to the music player system.

After training, the model generated will contain the neural network which will be used for testing whether the image gives the correct emotion or not. The emotion detected will be stored in a variable named label.

The label will be checked with each emotion (happy, anger, sad or neutral). The matched emotion will be used to open an HTML file which plays songs that are relevant to the emotion.

HTML files are designed for each emotion so that only the page corresponding to the detected emotion will be displayed. Each HTML file for the emotion will contain a function- MyFunction(), which will play song automatically for the detected emotion of user once the window is opened.

VI. RESULTS AND DISCUSSION

A music player system with design, analysis and implementation of Convolutional neural networks which will plays songs automatically according to the emotion of user is developed.

Training for emotion detection was initiated in our working PC but the memory was insufficient, therefore further training was done using GPU with images from different facial datasets after sorting into the required 4 labels of emotion. The datasets used are CK+, JAFFE, IMM and Kaggle.

All the images from the dataset were sorted into the 4 emotion labels before initiating the training phase using GPU. The graphic representation shown in the following figure depicts the time taken by GPU to complete a given number of image data.

The training was initialized in a standard PC but the execution was halted due to memory inconsistency of the system. So the computational complexity of neural network required that the model be trained using GPU.

Fixed and predetermined dimensions are associated with the input layer. Convolution is used to generate different feature maps for the same input image.

The CNN constructed for facial feature extraction and emotion detection will classify the emotion of user into 4 different labels: happy, anger, sad and neutral. These emotion are identified by the neural network based on different features from the face which are automatically learned by the neural network during the training phase.

Using the members of this work team as models, the following figures show the output associated with each emotion.

Fig. 2-5: Screenshots of detected emotions – Angry, Sad, Happy and Neutral

Emotion detected for a user will be stored in a variable and that is used to compare with each emotion and corresponding HTML file will be open containing songs reflecting that emotion.

The model constructed was able to successfully open the pages matching songs and emotion of user. In the development phase the system took around 3 seconds to automatically play the first song in the playlist which was opened.

The system was reviewed and model was improved with added function for timer to play songs at a lesser time. Following figures depict the pages showing playlist

corresponding to each detected emotion.

Fig. 6: Screenshot of music player for the emotion- angry

Fig. 7: Screenshot of music player for the emotion-happy

Fig. 8: Screenshot of music player for the emotion-neutral

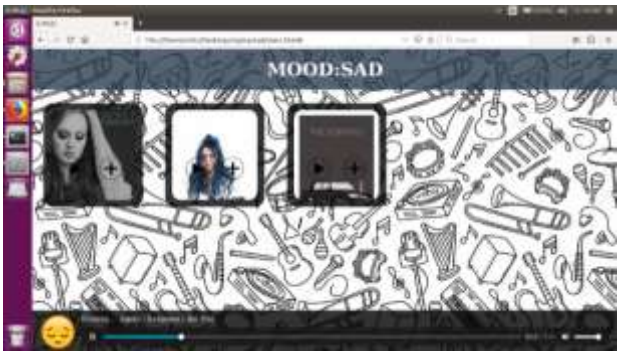

Fig. 9: Screenshot of music player for the emotion-sad

The model constructed is working properly with all the objectives met. Confusion matrix for the system is shown in the following table 1. The confusion matrix is used to describe the performance of classification model that we've constructed for the system using convolutional neural network.

Table 1: Confusion Matrix

	True Positive	True Negative	
Predicted Positive	TP = 56	FP = 24	80
Predicted Negative	FN = 17	TN = 3	20
	73	27	

The loss and accuracy of the model during the training phase is shown in the following figure 10.Loss rate against the number of epochs is given in the first figure as losscurve and accuracy rate against the number of epochs is given in the first figure as accuracy curve. The main feature that helped for attaining the required accuracy was large dataset. The datasets we have used are, CK+ of 1.7GB size, JAFFE of 14 MB, IMM of 48MB and some images of our colleagues. Number of epochs also helped to attain accuracy. We have gone through an existing emotion based music player known as SOLO. It is not an android system. It is a device that can be hung at wall. So it can only be used at that place only. E-MUZ is portable and can be used at any time.

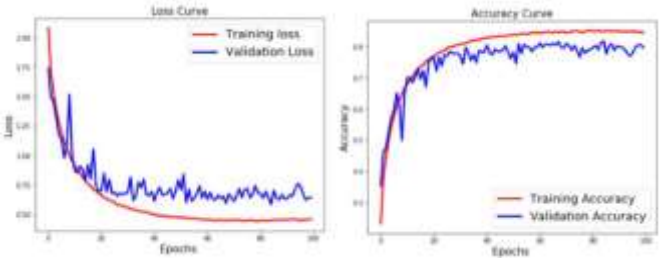

Fig 10: Performance evaluation

VII. CONCLUSION AND FUTURE SCOPE

As proposed in this paper, a music player which plays songs according to the user's emotion has been designed. The system has been divided into different modules for implementation which includes face detection, emotion detection and song classification. The proposed system is designed as an emotion aware application which provides a solution to the tangible approach of manual segregation of large playlists. Implementation of static face detection is done using Viola Jones Algorithm and testing of the same was done using images from different facial datasets. Dynamic face detection will be implemented as future work so that users can analyse emotions real time and such an application involves computational complexity and larger amount of dataset for getting higher accuracy level. The CNN classifier is designed in such a way that 4 emotion labels can be recognized: happy, anger, sad and neutral and more emotions can be worked for in the future.

REFERENCES

1. Shakhnarovich, Gregory, Paul A. Viola, and Baback Moghaddam. "A unified learning frame- work for real time face detection and classification." Proceedings of Fifth *IEEE international conference on automatic face gesture recognition. IEEE*, 2002.
2. Castrilln, Modesto, et al. "A comparison of face and facial feature detectors based on the Vi-olaJones general object detection framework." *Machine Vision and Applications* 22.3 (2011): 481-494.
3. Castrilln, M., et al. "ENCARA2: Real-time detection of multiple faces at different resolutions in video streams." *Journal of visual communication and image representation* 18.2 (2007): 130-140.
4. Ren, Jianfeng, Nasser Kehtarnavaz, and Leonardo Estevez. "Real-time optimization of Viola- Jones face detection for mobile platforms." 2008 *IEEE Dallas Circuits and Systems Work- shop: System-on-Chip-Design, Applications,*

5. *Integration, and Software. IEEE*, 2008.
6. Castrilln-Santana, Modesto, et al. "Multiple face detection at different resolutions for perceptual user interfaces." *Iberian Conference on Pattern Recognition and Image Analysis. Springer*, Berlin, Heidelberg, 2005.
7. Menezes, Paulo, Jos Carlos Barreto, and Jorge Dias. "Face tracking based on haar-like features and eigenfaces." *IFAC/EURON Symposium on Intelligent Autonomous Vehicles*. Vol. 500. 2004.
8. Mita, Takeshi, Toshimitsu Kaneko, and Osamu Hori. "Joint haar-like features for face detection." Tenth *IEEE International Conference on Computer Vision (ICCV'05)* Volume 1. Vol. 2. IEEE, 2005
9. Li, Haoxiang, et al. "A convolutional neural network cascade for face detection." *Proceedingsof the IEEE conference on computer vision and pattern recognition*. 2015.
10. Zhu, Qiang, et al. "Fast human detection using a cascade of histograms of oriented gradients." 2006 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. IEEE, 2006.
11. Luh, Guan-Chun. "Face detection using combination of skin color pixel detection and Viola- Jones face detector." 2014 *International Conference on Machine Learning and Cybernetics*. Vol. 1. IEEE, 2014.
12. Fan, Yin, et al. "Video-based emotion recognition using CNN-RNN and C3D hybrid networks." *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016.
13. Matsugu, M., Mori, K., Mitari, Y., & Kaneda, Y. (2003). Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, 16(5-6), 555-559.
14. Yu, Zhiding, and Cha Zhang. "Image based static facial expression recognition with multiple deep network learning." *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM*, 2015.
15. Levi, Gil, and Tal Hassner. "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns." *Proceedings of the 2015 ACM on international conference on multimodal interaction. ACM*, 2015.
16. Poria, Soujanya, et al. "Convolutional MKL based multimodal emotion recognition and sentiment analysis." 2016 *IEEE 16th international conference on data mining (ICDM). IEEE*, 2016.
17. Ng, Hong-Wei, et al. "Deep learning for emotion recognition on small datasets using transfer learning." *Proceedings of the 2015 ACM on international conference on multimodal interaction. ACM*, 2015.
18. Mollahosseini, Ali, David Chan, and Mohammad H. Mahoor. "Going deeper in facial expression recognition using deep neural networks." 2016 *IEEE Winter conference on applications of computer vision (WACV). IEEE*, 2016.
19. Khorrami, Pooya, Thomas Paine, and Thomas Huang. "Do deep neural networks learn facial action units when doing expression recognition?",*Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015.
20. Ebrahimi Kahou, Samira, et al. "Recurrent neural networks for emotion recognition in video." *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM*, 2015.
21. Dhall, A., Goecke, R., Joshi, J., Hoey, J., & Gedeon, T. (2016, October). Emotiw 2016: Video and group-level emotion recognition challenges. *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 427-432). *ACM.*