This full-text paper was peer-reviewed and accepted to be presented at the IEEE WiSPNET 2017 conference.

Smart Music Player Integrating Facial Emotion Recognition and Music Mood Recommendation

Shlok Gilda,[1] Husain Zafar,[2] Chintan Soni[3] and Kshitija Waghurdekar[4]
Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India
Email: [1]gildashlok@hotmail.com [2]husainzafar1996@gmail.com
[3]chintan.soni4@gmail.com [4]kshiti135@gmail.com

***Abstract*—Songs, as a medium of expression, have always been a popular choice to depict and understand human emotions. Reliable emotion based classification systems can go a long way in helping us parse their meaning. However, research in the field of emotion-based music classification has not yielded optimal results. In this paper, we present an affective cross-platform music player, EMP, which recommends music based on the real-time mood of the user. EMP provides smart mood based music recommendation by incorporating the capabilities of emotion context reasoning within our adaptive music recommendation system. Our music player contains three modules: Emotion Module, Music Classification Module and Recommendation Module. The Emotion Module takes an image of the user's face as an input and makes use of deep learning algorithms to identify their mood with an accuracy of 90.23%. The Music Classification Module makes use of audio features to achieve a remarkable result of 97.69% while classifying songs into 4 different mood classes. The Recommendation Module suggests songs to the user by mapping their emotions to the mood type of the song, taking into consideration the preferences of the user.**

***Index Terms*—Recommender systems, Emotion recognition, Music information retrieval, Artificial neural networks, Multi-layer neural network.**

I. Introduction

Current research in the field of music psychology has shown that music induces a clear emotional response in its listeners [1]. Musical preferences have been demonstrated to be highly correlated with personality traits and moods. The meter, timber, rhythm and pitch of music are managed in areas of the brain that deal with emotions and mood [2].

Undoubtedly, a user's affective response to a music fragment depends on a large set of external factors, such as gender, age [3], culture [4], personal preferences, emotion and context [5] (e.g. time of day or location). However, these external variables set aside, humans are able to consistently categorize songs as being happy, sad, enthusiastic or relaxed.

Current research in emotion based recommender systems focuses on two main aspects, lyrics [6], [10] and audio features [7]. Acknowledging the language barrier, we focus our efforts on audio feature extraction and analysis of modern American and British English songs in order to map those features to four basic moods. Automatic music classification using some mood categories yields promising results.

Facial expressions are the most ancient and natural way of conveying emotions, moods and feelings. For the purpose of this paper, we categorize facial expressions into 4 different emotional categories, viz. happy, sad, angry and neutral.

The main objective of this paper is to design a cost-effective music player which automatically generates a sentiment aware playlist based on the emotional state of the user. The application is designed to consume minimal system resources. The emotion module determines the emotion of the user. Relevant and critical audio information from a song is extracted by the music classification module. The recommendation module combines the results of the emotion module and the music classification module to recommend songs to the user. This system provides significantly better accuracy and performance than existing systems.

Section II briefly describes the past work done in the field. Sections III, IV and V delineate the datasets and models used for the emotion, music classification, and recommendation modules, respectively, and the results obtained for each. Finally, Section VI presents our conclusions and briefly illustrates the potential for further improvements in our methodology.

II. Related Works

Various methodologies have been proposed to classify the behavioural and emotional state of the user. Mase et al. focused on using movements of facial muscles while Tian et al. [8] attempted to recognize Actions Units

(AU) developed by Ekman and Friesen in 1978 using permanent and transient facial features. With evolving methodologies, the use of Convolutional Neural Networks (CNNs) for emotion recognition has become increasingly popular [9].

Music has also been classified using lyrical analysis [6], [10]. While this tokenized method is relatively easier to implement, on its own it is not suitable to classify songs accurately. Another obvious concern with this method is the language barrier which restricts classification to a single language.

Another method for music mood classification is using acoustic features like tempo, pitch and rhythm to identify the sentiment conveyed by the song. This method involves extracting a set of features and using those feature vectors to find patterns characteristic to a specific mood [7], [21].

978-1-5090-4442-9/17/$31.00

III. Emotion Module

In this section, we study the usage of convolutional neural networks (CNNs) in the context of emotion recognition [11], [12]. CNNs are known to simulate the human brain when analyzing visuals; however, given the computational requirements and complexity of a CNN, optimizing a network for efficient computation is necessary. Thus, a CNN is implemented to construct a computational model which successfully classifies emotion into 4 moods, namely, happy, sad, angry and neutral, with an accuracy of 90.23%.

A. Dataset Description

The dataset we used for training the model is from a Kaggle Facial Expression Recognition Challenge, FER2013 [13]. The data consists of 48×48 pixel grayscale images of faces. Each of the faces are organized into one of the 7 emotion classes: angry, disgust, fear, happy, sad, surprise, and neutral. For this research, we have made use of 4 emotions: angry, happy, sad and neutral. There is a total of 26,217 images corresponding to these emotions. The breakdown of the images is as follows: happy with 8,989 samples, sad with 6,077 samples, neutral with 6,198 samples, angry with 4,953 samples.

B. Model Description

A multi-layered convolutional neural network is programmed to evaluate the features of the user image [14], [15]. The convolutional neural network contains an input layer, some convolutional layers, ReLU layers, pooling layers, and some dense layers (aka. fully-connected layers), and an output layer. These layers are linearly stacked in sequence.

1) Input Layer: The input layer has fixed and predetermined dimensions. So, for pre-processing the image, we used OpenCV for face detection in the image before feeding the image into the layer. Pre-trained filters from Haar Cascades along with Adaboost are used to quickly find and crop the face. The cropped face is then converted into grayscale and resized to 48-by-48 pixels. This step greatly reduces the dimensions from (3, 48, 48) (RGB) to (1, 48, 48) (grayscale) which can be easily fed into the input layer as a numpy array.

2) Convolutional Layers: A set of unique kernels (or feature detectors), with randomly generated weights, are specified as one of the hyperparameters in the Convolution2D layer. Each feature detector is a (3, 3) receptive field, which slides across the original image and computes a feature map. Convolution generates different feature maps for the same input image. Distinct filters are used to perform operations that represent how pixel values are enhanced, for example, blur and edge detection. Filters are applied successively over the entire image, creating a set of feature maps. In our neural network, each convolutional layer generates 256 feature maps. Rectified Linear Unit (ReLU) has been used after every convolution operation. After a set of convolutional layers, a popular pooling method, MaxPooling, was used to reduce the dimensionality of each feature map, all the while retaining the critical information. We used (2, 2) windows which consider only the maximum pixel values within the window from the feature map. The pooled pixels form an image with dimensions reduced by a factor of 4.

TABLE I
Confusion Matrix for Emotion Module.

		Predicted class			
		Angry	Happy	Sad	Calm
Actual class	Angry	981	34	13	21
	Happy	27	1497	16	33
	Sad	41	46	1218	111
	Calm	29	56	85	1036

3) Dense Layers: The output from the convolutional and pooling layers represent high-level features of the input image. The dense layer uses these features for classifying the input image into various classes. The features are transformed through the layers which are connected with trainable weights. The network is trained by forward propagation of training data and then backward propagation of its errors. Our model uses two sequential fully connected layers. The network generalizes well to new images and is able to gradually make adjustments until the errors are minimized. A dropout of 20% was applied in order to prevent overfitting of the training data. This helped us control the model's sensitivity to noise during training while maintaining the necessary complexity of the architecture.

4) Output Layer: We used softmax as the activation function at the output layer of the dense layer. Thus, the output is represented as a probability distribution for each emotion class. Models with various combinations of hyper-parameters were trained and evaluated utilizing a 4 GiB DDR3 NVIDIA 840M graphics card using the NVIDIA CUDA® Deep Neural Network library (cuDNN). This greatly reduced training time and increased efficiency in tuning the model. Ultimately, our network architecture consisted of 9 convolutional layers with one max-pooling after every three convolution layers followed by 2 dense layers, as seen in Fig. 1.

C. Results

The final network was trained on 20,973 images and tested on 5,244 images. At the end, the model achieved an accuracy of 90.23%. Table I displays the confusion matrix for the module.

Evidently, the system performs very well in classifying images belonging to the "angry" category. We also note interesting results under "happy" and "sad" category owing to the remarkable differences in Action Units as mentioned by Ekman [9]. The F-measure of this system comes out to be 90.12%.

IV. Music Classification Module

In this section, we describe the procedure that was used to identify the mapping of each song with its mood. We extracted the acoustic features of the songs using LibROSA [16], aubiopitch [17] and other state-of-the art audio extraction algorithms. Based on these features, we trained an artificial

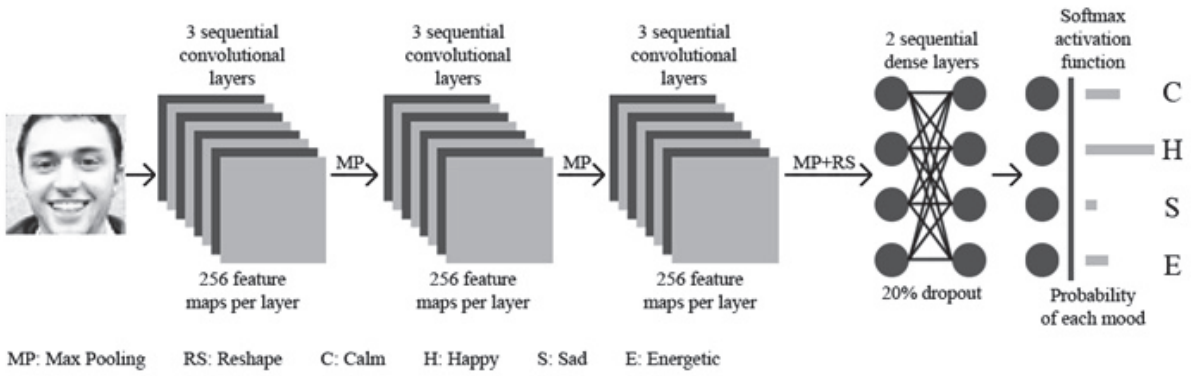

Fig. 1. Final network architecture.

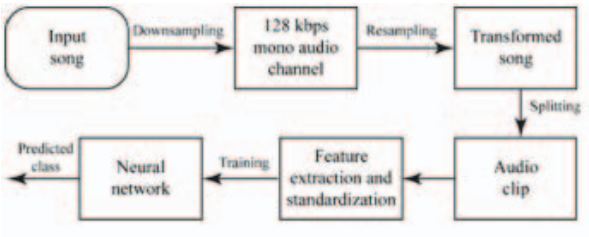

Fig. 2. Music classification process.

neural network which successfully classifies the songs in 4 classes with an accuracy of 97.69%. The classification process is described in Fig. 2.

A. Dataset Description

The dataset comprises of 390 songs spread across four moods. The distribution of the songs is as follows: class A with 100 songs, class B with 93 songs, class C with 100 songs and class D with 97 songs. The songs were manually labelled and the class labels were verified by 10 paid subjects. Class A comprises of exciting and energetic songs, class B has happy and joyful songs, class C consists of sad and melancholy songs, and class D has calm and relaxed songs.

1) Preprocessing: All the songs were down sampled to a uniform bit-rate of 128 kbps, a mono audio channel and resampled at a sampling frequency of 44 100 Hz. We further split each song to obtain clips that contained the most meaningful parts of the song. The feature vectors were then standardized so that it had zero mean and a unit variance.

2) Feature Description: We identified several mood sensitive audio features by reading current works [10] and the results from the 2007 MIREX Audio Mood Classification task [18], [19].

The candidate features for the extraction process belonged to different classes: spectral (RMSE, centroid, rolloff, MFCC, kurtosis, etc.), rhythmic (tempo, beat spectrum, etc.), tonal mode and pitch. All these descriptions are standard. All the features were extracted using Python 2.7 and relevant packages [16], [17].

After identifying all the features, we used Recursive Feature Elimination (or RFE) to select those features that best contribute to the accuracy of the model. RFE works by recursively removing attributes and building a model on those attributes that remain. It uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute. The selected features were pitch, spectral rolloff, mel-frequency cepstral coefficients, tempo, root mean square energy, spectral centroid, beat spectrum, zero-cross rate, short-time Fourier transform and kurtosis of the songs.

B. Model Description

A multi-layered neural network was trained to evaluate the mood associated with the song. The network contains an input layer, multiple hidden layers and a dense output layer.

The input layer has fixed and predetermined dimensions. It takes the 10 feature vectors as input and uses ReLU operation to provide non-linearity to the dataset. This ensured that the model performs well in real-world scenarios as well.

The hidden layer is a traditional multi-layer perceptron, which allowed us to make combination of features which led to a better classification accuracy. The output layer used a softmax activation function which produces the output as a probability for each mood class.

C. Results

We achieved an overall classification accuracy of 97.69% and F1 score of 97.692% after 10-fold cross-validation using our neural network. Table II displays the confusion matrix.

Undoubtedly, the level of performance of the music classification module is exceptionally high.

V. Recommendation Module

This module is responsible for generating a playlist of relevant songs for the user. It allows the user to modify the playlist based on her/his preferences and modify the class labels of the songs as well. The working of the recommendation module is explained in Fig. 3.

TABLE II
CONFUSION MATRIX FOR MUSIC CLASSIFICATION MODULE.

		Predicted class			
		Class A	Class B	Class C	Class D
Actual class	Class A	99	1	0	0
	Class B	2	90	0	1
	Class C	0	1	96	3
	Class D	0	0	1	96

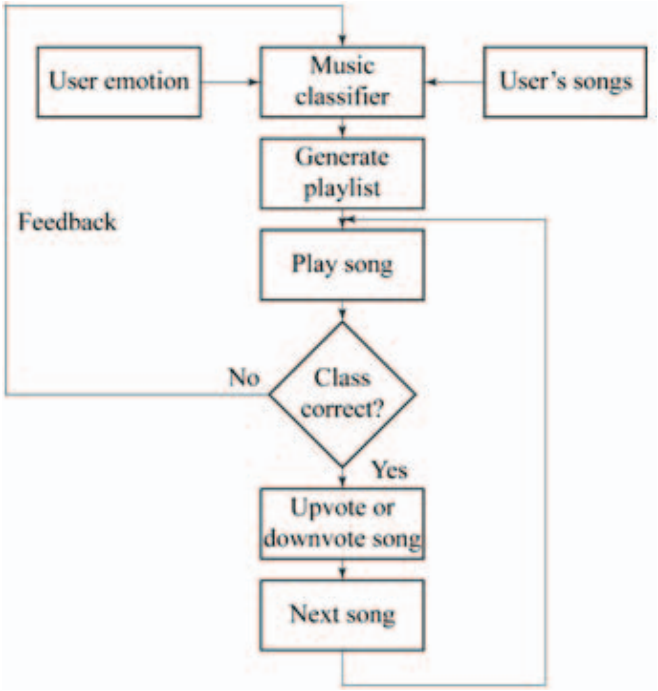

Fig. 3. Recommendation Module Flowchart.

A. Mapping and Playlist Generation

Classified songs are mapped to the user's mood. This mapping is as shown in Fig. 1. The system was developed after referring to the Russell 2-D Valence-Arousal Model and Geneva Emotion Wheel. After the mapping procedure is complete, a playlist of relevant songs is generated. Similar songs are grouped together while generating the playlist. Similarity between songs was calculated by comparing songs over 50 ms intervals, centered on each 10 ms time window. After empirical observations, we found that the duration of these intervals is on the order of magnitude of a typical song note. Cosine distance function was used to determine the similarity between audio files. Feature values corresponding to an audio file were compared to the values (for the same features) corresponding to audio files belonging to the same class label. The recommendation engine has a twofold mechanism; it recommends songs based on:

1. User's perceived mood.
2. User's preference.

Initially, a playlist of all songs belonging to the particular class is generated. The user can mark a song as favorite depending on her/his choice. A favorite song will be assigned a higher priority value in the playlist. Also, the interpretation of the mood of a song can vary from person to person. Understanding this, the user is allowed to change the class label of the songs according to their taste of music (see Table III).

TABLE III
EMOTION-MOOD MAPPING.

User emotion	Music mood
Neutral	Calm and refreshing
Happy	Happy and elated
Sad	Serene and soothing
Energetic	Exciting and energetic

B. Adaptive Music Player

We were able to implement an adaptive music player by the use of a very popular online machine learning algorithm, Stochastic Gradient Descent (SGD) [20]. If the user wants to change the class of a particular song, SGD is implemented considering the new label for that specific user only.

Multiple single-pass algorithms were analyzed for their performance with our system but SGD performed most efficiently considering the real-time nature of the music player. Parameter updates in SGD occur after processing of every training example from the dataset. This approach yields two advantages over the batch gradient descent algorithm. Firstly, time required for calculating the cost and gradient for large datasets is reduced. Secondly, integration of new data or amendment of existing data is easier. The frequent, highly variant updates demand the learning rate α to be smaller as compared to that of batch gradient descent [20].

VI. CONCLUSION AND FUTURE SCOPE

The results obtained above are very promising. The high accuracy and quick response time of the application makes it suitable for most practical purposes. The music classification module in particular, performs significantly well; it achieves high accuracy in the "angry" category, while also performing appreciably well in the "happy" and "calm" categories. Thus, EMP reduces user efforts for generating playlists. by efficiently mapping the user's emotion to the correct song class with an overall accuracy of 97.69%, it achieves optimistic results for the four moods studied.

We also recognize the room for improvement. It would be interesting to analyze how the system performs when all seven basic emotions are taken into consideration; additional songs from different languages and regions can also be added to make the recommendation system more robust. User preferences can be collected to improve the overall system using collaborative filtering. We plan to address these issues in a future work.

REFERENCES

[1] Swathi Swaminathan and E. Glenn Schellenberg, "Current emotion research in music psychology," *Emotion Review*, vol. 7, no. 2, pp. 189–197, Apr. 2015.

[2] "How music changes your mood", Examined Existence. [Online]. Available: http://examinedexistence.com/how-music-changes-your-mood/. Accessed: Jan. 13, 2017.
[3] Kyogu Lee and Minsu Cho, "Mood Classification from Musical Audio Using User Group-dependent Models."
[4] Daniel Wolff, Tillman Weyde, and Andrew MacFarlane, "Culture-aware Music Recommendation."
[5] Mirim Lee and Jun-Dong Cho, "Logmusic: context-based social music recommendation service on mobile device," Ubicomp'14 Adjunct, Seattle, WA, USA, Sep. 13–17, 2014.
[6] D. Gossi and M. H. Gunes, "Lyric-based music recommendation," in *Studies in computational intelligence*. Springer Nature, pp. 301–310, 2016.
[7] Bo Shao, Dingding Wang, Tao Li, and Mitsunori Ogihara, "Music recommendation based on acoustic features and user access patterns," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, Nov. 2009.
[8] Ying-li Tian, T. Kanade, and J. Cohn, "Recognizing lower. Face action units for facial expression analysis," in *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, Mar. 2000, pp. 484–490.
[9] Gil Levi and Tal Hassner, "Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns."
[10] E. E. P. Myint and M. Pwint, "An approach for mulit-label music mood classification," in *2010 2nd International Conference on Signal Processing Systems*, Dalian, 2010, pp. V1-290-V1-294.
[11] Peter Burkert, Felix Trier, Muhammad Zeshan Afzal, Andreas Dengel, and Marcus Liwicki, "DeXpression: Deep Convolutional Neural Network for Expression Recognition."
[12] Ujjwalkarn, "An intuitive explanation of Convolutional neural networks," the data science blog, 2016. [Online]. Available: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/. Accessed: Jan. 13, 2017.
[13] Ian J. Goodfellow et al., "*Challenges in Representation Learning: A report on three machine learning contests*."
[14] S. Lawrence, C. L. Giles, Ah Chung Tsoi, and A. D. Back, "Face recognition: a convolutional neural-network approach," in *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98–113, Jan. 1997.
[15] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel, "Human-Computer systems interaction: back-grounds and applications," 3, ch. *Emotion Recognition and Its Applications*, Cham: Springer International Publishing, 2014, pp. 51–62.
[16] Brian McFee, Matt McVicar, Colin Raffel, Dawen Liang, Oriol Nieto, Eric Battenberg, . . ., and Adrian Holovaty, (2015). librosa: 0.4.1 [Data set]. Zenodo. http://doi.org/10.5281/zenodo.32193.
[17] The aubio team, "Aubio, a library for audio labelling," 2003. [Online]. Available: http://aubio.org/. Accessed: Jan. 13, 2017.
[18] J. S. Downie, *The music information retrieval evaluation exchange (mirex)*. D-Lib Magazine, 12(12), 2006.
[19] Cyril Laurier, Perfecto Herrera, M Mandel and D Ellis, "Audio music mood classification using support vector machine."
[20] "Unsupervised feature learning and deep learning Tutorial," [Online]. Available: http://ufldl.stanford.edu/tutorial/supervised/OptimizationStochasticGradientDescent/. Accessed: Jan. 13, 2017.
[21] A. S. Bhat, V. S. Amith, N. S. Prasad, and D. M. Mohan, "An efficient classification algorithm for music mood detection in western and hindi music using audio feature extraction," in *2014 Fifth International Conference on Signal and Image Processing*, Jeju Island, 2014, pp. 359–364.