ANNALES UMCS INFORMATICA DOI: 10.17951/AI.2016.16.2.1

Music Playlist Generation using Facial Expression Analysis and Task Extraction

Arnaja Sen
Department of Information Technology
K. J. Somaiya College of Engineering
Mumbai, India
e-mail: arnaja.sen@somaiya.edu

Priyanka Kuwor
Department of Information Technology
K. J. Somaiya College of Engineering
Mumbai, India
e-mail: priyanka.kuwor@somaiya.edu

Dhaval Popat
Department of Information Technology
K. J. Somaiya College of Engineering
Mumbai, India
e-mail: d.popat@somaiya.edu

Era Johri
Department of Information Technology
K. J. Somaiya College of Engineering
Mumbai, India
e-mail: erajohri@somaiya.edu

Hardik Shah
Department of Information Technology
K. J. Somaiya College of Engineering
Mumbai, India
e-mail: hns@somaiya.edu

***Abstract*— In day to day stressful environment of IT Industry, there is a truancy for the appropriate relaxation time for all working professionals. To keep a person stress free, various technical or non-technical stress releasing methods are now being adopted. We can categorize the people working on computers as administrators, programmers, etc. each of whom require varied ways in order to ease themselves. The work pressure and the vexation of any kind for a person can be depicted by their emotions. Facial expressions are the key to analyze the current psychology of the person. In this paper, we discuss a user intuitive smart music player. This player will capture the facial expressions of a person working on the computer and identify the current emotion. Intuitively the music will be played for the user to relax them. The music player will take into account the foreground processes which the person is executing on the computer. Since various sort of music is available to boost one's enthusiasm, taking into consideration the tasks executed on the system by the user and the current emotions they carry, an ideal playlist of songs will be created and played for the person. The person can browse the playlist and modify it to make the system more flexible. This music player will thus allow the working professionals to stay relaxed in spite of their workloads.**

Keywords—facial expression analysis, emotion recognition, feature extraction, viola jones face detection, gabor filter, adaboost, k-NN algorithm, task extraction, music classification, playlist generation)

I. INTRODUCTION

Music plays an imperative part in an individual's day to day life and in the current propelling advancements. The inclination, identity and sentiments of people can be perceived through their emotions. Music often times communicates passionate qualities and characteristics of human identity, for example, happiness, sadness, aggressiveness, etc. Additionally, it has a significant part in human culture since it unequivocally brings out emotions and influences social exercises and connections. Consequently, the kind of music a person prefers to listen at a particular time can reflect the state of the mind of a person right then and there. This marvel can be used in a few applications where human feelings assume noteworthy part.

Furthermore, it helps the users to minimize their endeavors in overseeing extensive playlists proficiently. Generally, people have a vast collection of songs in their computer. Manually separating the list of songs and creating a suitable playlist based on an individual's emotional features is an exceptionally monotonous task. Along these lines, the users tend to play arbitrary music which may not suit their current passionate state. We have developed a music player which recognizes user's mood and tasks performed by them at a particular instant, and accordingly manages the playlist for them. Since, the human face assumes a vital part in extraction of an individual's conduct and passionate expressions, this software will capture the user's facial expressions and features to decide the present state of mind of the user. The facial expressions and emotions will be arranged into five unique categories which are anger, disgust, happiness, neutral, and sadness[10]. The images are captured through webcam. The captured image will be spared and passed on to the rendering stage. Simultaneously, the task being performed on the user's computer is extracted and classified into one of the three categories which are browsing, programming and other tasks. Finally, a combination of the recognized

emotion and most recent task's category is provided to the neural networks for training and analysis which then provides the user with an ideal playlist. This playlist will help in improving the user's mood and maximize the efficiency in the task being performed. Furthermore, this will empower the users to relieve their nerves with the assistance of fitting music. The current mood of the user may contrast after some time. Hence, the process will be reiterated after every ten minutes to detect any variations in the emotions or task being performed. The system is constructed in a conventional way and comprises of five phases, which are, face detection, feature extraction, emotion recognition, task extraction and identification, and music playlist generation.

II. Literature Review

Music and its utilization for feeling control procedures, right up till today remains an uncertain question. Numerous test designs and clinical applications crosswise over various societies' landmasses and have protected music as a self regulative device. Feelings can show up in many parts of human-to-human correspondence and regularly provide ancillary data about a message. The clinical and the nonclinical reviews, all exhibit the successful utilization of music as self-regulative device for feelings. In spite of the assorted qualities between the review plans, dynamic music making and listening versus intelligent and non-trial utilization of music, all reviews uncovered the individual uses of music for individual utilization, advancing self-regulative aptitudes for positive modification, which are socially and practically identical between all tried social orders. The accompanying investigations of the current papers bolster the general understanding of this survey that music listening is most regularly utilized with a colossal arrangement of objectives and procedures for enthusiastic control purposes. The takeaways from the papers are as follows:

1. Different systems and methodologies have been proposed and devised to order human enthusiastic condition of conduct. The paper[1] "Human-computer interaction using emotion recognition from facial expressions" composed by F. ABDAT, C. MAAOUI and A. PRUSKI empowers us to understand how the face recognition and emotion detection can be carried out using the support vector machines (SVM) algorithm and Facial Action Coding System (FACS). Besides this, the detection of the various distinct feature points is also clarified in this paper [1]. Moreover, the different distances between the disparate feature face points are considered for recognizing facial expressions.
2. The paper [2] "Playlist environmental analysis for the serendipity-based data mining" composed by Chim Chwee WONG, Emy Salfarina ALIAS and Junichi KISHIGAMI enables us to identify how the playlist can be played by the users' inclination towards the diverse kinds of music of various periods. Moreover, data cleansing and clustering concepts for music are clarified in detail. Further, the various data visualization techniques and serendipity recommendation systems are likewise introduced in this paper [2].

III. Face Detection

The most significant element of an emotion recognition system is the appropriate detection of the face and extraction of the relevant features. Viola Jones algorithm is selected for face detection as it has a very high detection rate which makes it robust, and has an advantage of operating in real time.

The Viola Jones algorithm[4] has four stages which are as follows:

- Haar Feature Selection
- Creation of an Integral Image
- AdaBoost Training Algorithm
- Cascading Classifiers

A. Haar Feature Selection

All the human faces have similar features. These similarities maybe matched using Haar feature selection. The possible types of features are Two, Three and Four rectangle features[4]. These features are represented in figure 1.

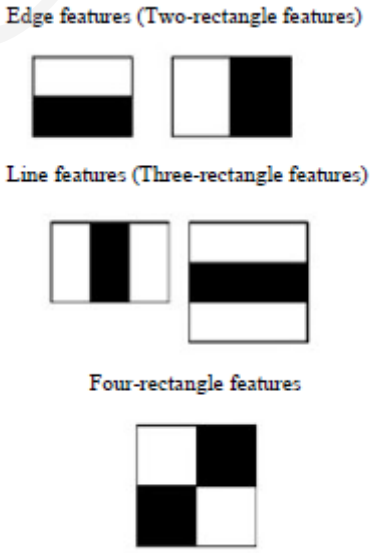

Fig. 1 HAAR Features [4]

The edge feature is useful in detection of contrast between two vertical or horizontal adjacent regions. The line feature is effective in the detection of contrasted region placed between two similar regions. The four-rectangle feature is beneficial in detection of similar regions placed diagonally. The feature value is the difference between the sum of pixels under the white rectangle and the sum of the pixels under the black rectangle.

B. Creation of an Integral Image

Calculation of the areas under the black and the white rectangles in real time is computationally expensive. Hence, the integral image[4] concept is utilized. The integral image at location (x,y) is the sum of the pixels above and to the left of

(x,y) inclusive as illustrated in figure 2. The following pair of recurrences[4] are used for the calculation of an integral image:

$$s(x,y) = s(x,y-1) + i(x,y) \quad (1)$$

$$ii(x,y) = ii(x-1,y) + s(x,y) \quad (2)$$

where

- *ii(x, y)* is the integral image,
- *i(x, y)* is the original image, and
- *s(x, y)* is the cumulative row sum

Original Image

05	02	03	04	01
01	05	04	02	03
02	02	01	03	04
03	05	06	04	05
04	01	03	02	06

Integral Image

05	07	10	14	15
06	13	20	26	30
08	17	25	34	42
11	25	39	52	65
15	30	47	62	81

Fig. 2 Original Image and Integral Image

C. AdaBoost Training Algorithm

In 24*24 pixels sub window, there are approximately 162,000 features, all of which are not relevant. AdaBoost facilitates removal of these irrelevant features by selecting the best features and to train classifiers that use them. This algorithm constructs strong classifiers as a linear combination[4] of weak classifiers as follows:

$$h(x) = sign\left(\sum_{j=1}^{M} a_j\, h_j(x)\right) \quad (3)$$

where,

- each weak classifier is a threshold function based on the feature f j,
- the threshold value θ_j and the polarity s_j are determined in the training as well as the co-efficient α_j, and
- $h_j(x) = \begin{cases} -s_j & \text{if } f_j < \theta_j \\ s_j & \text{otherwise} \end{cases}$

The working of the AdaBoost training algorithm is illustrated in figure 3.

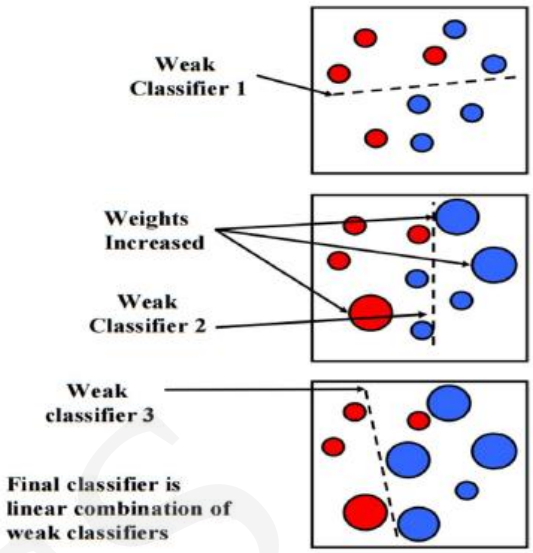

Fig. 3 AdaBoost[4]

D. Cascading Classifiers

A cascade[4] of gradually more complex classifiers is employed in this case. Each classifier involves a pre-decided number of features which allows the differentiation between the faces and non-faces. Hence, as the image goes through each stage in the cascade, the probability of the image containing a face increases. A positive result from the first classifier leads to the evaluation of a second classifier which has been adjusted to achieve a very high detection rate. A positive outcome from the second classifier leads to a third classifier. This process is continued till the last classifier is reached. The sub-window is rejected if a negative outcome occurs at any point. Various stages in the cascade are made by training the classifiers using AdaBoost and then the threshold is adjusted to minimize the false negatives.

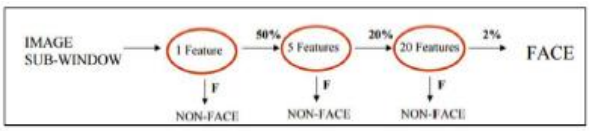

Fig. 4 Cascading[4]

IV. Feature Extraction

A set of Gabor filters [3] with different frequencies and orientations is used to extract useful features from an image. Gabor filter has been adopted for the extraction of features since it is more effective than Geometric Features-Based Filters, and works better in real-world environments. A Gabor filter is essentially a sinusoidal signal with a given frequency and orientation, modulated by a Gaussian. In the discrete domain, two-dimensional Gabor filters [12] are given by:

$$G_c[i.j] = Be^{-\frac{(i^2+j^2)}{2\sigma^2}} \cos\left(2\pi f\,(i\cos\theta + j\sin\theta)\right) \quad (4)$$

$$G_s[i,j] = Ce^{-\frac{(i^2+j^2)}{2\sigma^2}} \sin\left(2\pi f\,(i\cos\theta + j\sin\theta)\right) \quad (5)$$

where,

- B and C are normalizing factors to be determined,
- f is the frequency being looked for in the texture,
- by varying θ, we can look for texture oriented in a particular direction, and
- by varying σ, we can change the support of the basis or the size of the image region being analyzed.

Otsu method [3] is used to compute the threshold level using which a gray level image is reduced to a binary image. This algorithm assumes that the image whose threshold is to be considered contains only foreground and background pixels and then it calculates the optimum threshold separating those two classes so that their intra-class variance is minimal. Otsu threshold method is applied as a pre-processing step in order to remove noise and binarize the image.

To extract the features from images using Gabor filter [3], following steps are carried out:

Step 1: First reduce the size of an RGB input image to 40×40×3.

Step 2: Otsu threshold[3] is applied on each component (R, G, B components) separately.

Step 3: Gabor Wavelet filter is created and the parameters for Gabor wavelet are set as follows:

Gabor kernel size: 24*24
Orientations: 0, π/4, π/2, 3π/4
Scales: 0,1,2,3

Step 4:The kernel designed is composed of real and imaginary parts with 4 orientations and 4 scales [3].

Step 5: Then convolve the image's each component with 16 Gabor wavelets i.e. with real and imaginary part of Gabor filter separately and obtaining 16 real and 16 imaginary responses respectively corresponding to each component (R, G, B).

Step 6: Repeat the above steps for all the images.

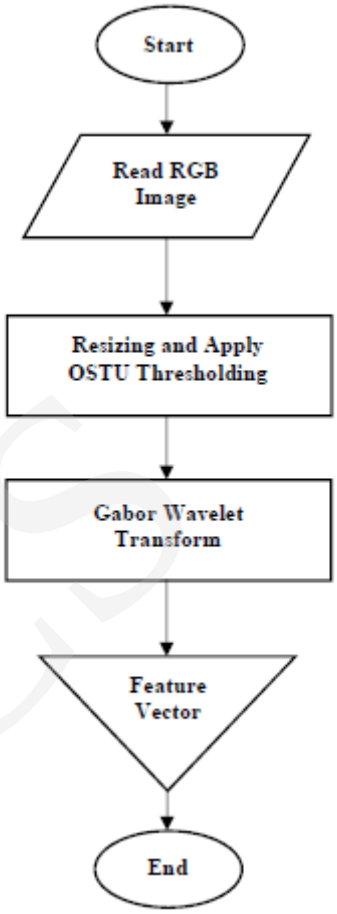

Fig. 5 Flowchart for feature extraction [3]

V. Emotion Recognition

The next task after the features have been extracted is the recognition of emotions based on these features. This task requires a training set consisting of images pertaining to different emotions and features that are distinctive for particular expression. The training set consists of two databases: Japanese Female Facial Expressions (JAFFE) database and YALE database. The JAFFE database[8] contains 213 images of 7 facial expressions (neutral, sadness, surprise, happiness, fear, anger, and disgust) posed by 10 Japanese females. Each image has been classified as per seven emotions by 60 Japanese subjects. The YALE database[9] contains 165 grayscale images of 15 individuals. There are 11 images per subject, one per different facial expression, which are happy, normal, sad, sleepy, surprised, and wink.

VI. Task Extraction And Identification

The tasks performed on the computer are extracted by tracking the processes which is done with the help of Windows Management Instrumentation Command-line (WMIC) tool[5]. WMIC is a command-line and scripting interface that eases the use of Windows Management Instrumentation (WMI) and various systems managed through WMI. WMIC is based on aliases which act upon the WMI namespace in a predefined

manner by taking the simple commands that you enter at the command prompt. Therefore, aliases are friendly syntax intermediaries between you and the namespace. The process ID and the process name is used to track a particular task and classify it into Browsing, Programming, or Other categories.

The algorithm which is used is as follows:

1. Find the process ID of all the processes running on thecomputer using WMIC.
2. Remove the spaces from the output string obtained in step number 1.
3. Create an array of the process ID's retrieved from step number 2.
4. Find the process name for the five recent processes running on the computer using WMIC and the process ID.
5. Finally, categorize the processes as Browsing, Programming, and Other tasks.

The categorized tasks' output is then sent to the neural network[6] along with the analyzed emotion to create the final playlist.

VII. Music Playlist Generation

The last and most important component of this system is playing the songs which soothes the nerves of the user and refreshes them. Based on a combination of the expression recognized and task detected, a playlist is generated from our songs database which contains songs pertaining to each combination of emotion and task.

In the songs database, the songs are classified into various genres [11] using K-Nearest Neighbors (k-NN)[7] after performing an analysis of the lyrics present in the song. The k-NN algorithm provides a class membership to each song where each class corresponds to a different combination of emotion and task. Fifteen such classes are created and the songs are added to them. K-NN algorithm[7] has been employed since it does not require much training and can handle noisy training data very well.

The generated playlist contains three songs at a time. After a duration of ten minutes, the tasks are tracked again and the emotion recognition process is repeated to generate a playlist of three songs. This process continues until the user closes the application.

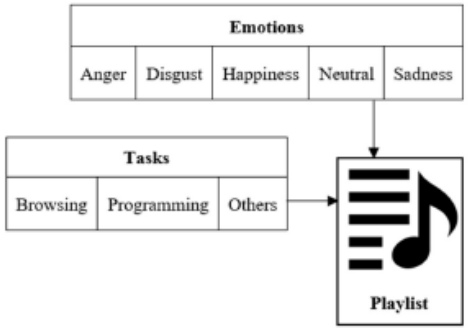

Fig. 6 Playlist generation

VIII. Results

'X-Beats – A Smart Music Player' is created as a desktop application utilizing MATLAB as a development tool. MATLAB version 2016a is used for the implementation of this software. After testing the system for a duration of over two months in a real-world environment, we have attained the following performance measures for our software:

Average accuracy of emotion recognition: 83%
Average accuracy of task detection: 96%

TABLE 1 Accuracy Of Different Emotions

Emotion	Accuracy
Anger	82%
Disgust	86%
Happiness	92%
Neutral	75%
Sadness	78%

TABLE 2 Accuracy Of Different Emotions

Task	Accuracy
Browsing	96%
Coding	93%
Others	98%

TABLE 3 Accuracy Of Different Emotions

Operation Time	Taken
Capture the image and detect face	2 seconds
Recognize the emotion	4 seconds
Detect and categorize task	2 seconds
Extract songs from the database	4 seconds
Playlist generation process	3 seconds

Fig. 7 Intermediate output of our software

IX. Conclusion And Future Work

Extensive efforts have been made to naturally identify the state of mind of the user and present them with a playlist of songs which is reasonable for their current mood. Additionally, the

music player will track the foreground processes which are in execution on the computer. For the emotion recognition module, we have achieved an accuracy of 83% for the testing data. The software has been tested using various sample data in a real-time environment. The task detection module has been tested with an accuracy of 96%. The overall operating time of the system has been reduced by optimizing the algorithms which have been used for implementing the application.

The music genre classification has been beneficial in reducing the searching time of songs, and thereby increasing the overall efficiency of the system. Furthermore, this has proved to be advantageous in swiftly generating the music playlist. The major strengths of the system are complete automation of the software as well as user independence.

In future, we intend to increase the accuracy of the application by training the neural network with no tolerance for the error rate signal. Moreover, we plan to enhance our software by developing a mechanism that will be helpful in the music therapy treatment, and also assist the music therapists in treating the patients suffering from disorders like anxiety, acute depression, and trauma.

REFERENCES

[1] F. ABDAT, C. MAAOUI and A. PRUSKI, "Human-computer interaction using emotion recognition from facial expression," UKSim 5th European Symposium on Computer Modeling and Simulation, 2011, pp. 196-201.

[2] Chim Chwee WONG, Emy Salfarina ALIAS and Junichi KISHIGAMI, "Playlist environmental analysis for the serendipity-based data mining," 2013 International Conference on Informatics, Electronics and Vision, May 2013.

[3] Rakesh Kumar, Rajesh Kumar and Seema, "Gabor Wavelet Based Features Extraction for RGB Objects Recognition Using Fuzzy Classifier," International Journal of Application or Innovation in Engineering & Management, Vol. 2, Issue 8, August 2013, pp. 122-127.

[4] P. Viola and M. J. Jones, "Robust real-time object detection," International Journal of Computer Vision, Vol. 57, No. 2, pp. 137–154, 2004.

[5] Hui Peng and Yao Wang, "WMIC-based technology server network management software design," 2010 Second Pacific-Asia Conference on Circuits, Communications and System, August 2010.

[6] Derrick H. Nguyen and Bernard Widrow, "Neural networks for selflearning control systems," IEEE Control Systems Magazine, Vol. 10, Issue 3, April 1990.

[7] Qingfeng Liu, Ajit Puthenputhussery and Chengjun Liu, "Novel general KNN classifier and general nearest mean classifier for visual classification," IEEE International Conference on Image Processing, September 2015.

[8] Michael J. Lyons, Shigeru Akemastu, Miyuki Kamachi, Jiro Gyoba, "Coding Facial Expressions with Gabor Wavelets", 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200-205, 1998.

[9] P. N. Bellhumer, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection", IEEE Transactions on Pattern Analysis and Machine Intelligence, Special Issue on Face Recognition, 17(7):711--720, 1997.

[10] P. Ekman and W. V. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement," Palo Alto, California, USA: Consulting Psychologists Press, 1978.

[11] T. Li and M. Ogihara, "Music genre classification with taxonomy," In Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, pages 197–200, Philadelphia, USA, 2005.

[12] A.G. Ramakrishnan, S. Kumar Raja and H.V. Raghu Ram, "Neural network-based segmentation of textures using Gabor features," Proc. 12th IEEE Workshop on Neural Networks for Signal Processing, pp. 365 - 374, 2002.

[13] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski, "Classifying facial action," IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 21, October 1999, pp. 974-989.

[14] S. S. Ge, C. Wang, C. C. Hang, "Facial expression imitation in human robot interaction," Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, August 2008.

[15] Carlos N. Silla, Alessandro L. Koerich and Celso A. A. Kaestner, "Feature selection in automatic music genre classification," Tenth IEEE International Symposium on Multimedia, October 2008, pp. 39-44.

[16] Andrew Ryan, Jeffery F. Cohn, Simon Lucey, Jason Saragih, Patrick Lucey, & Fernando De la Torre and Adam Rossi, "Automated facial expression recognition system," IEEE Journal, 2009, pp. 172-177.

[17] Changbo Hu, Ya Chang, R. Feris, M. Turk, "Manifold based analysis of facial expression," IEEE Conference on Computer Vision and Pattern Recognition Workshop, June 2004.

[18] T. Kanade, J. F. Cohn, Y. Tian, "Comprehensive database for facial expression analysis," Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2000, pp. 46-53.

[19] Z. Zhang, M. J. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometrybased and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in IEEE FG, April 1998.

[20] M. Pantic, L. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art," IEEE Transactions On Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, 2000.

[21] P. Menezes, J.C. Barreto, and J. Dias, "Face tracking based on Haar-like features and eigenfaces," 5th IFAC Symposium on Intelligent Autonomous Vehicles, Lisbon, Portugal, July 5-7, 2004.