Ingénierie des Systèmes d'Information
Vol. 28, No. 1, February, 2023, pp. 225-229
Journal homepage: http://iieta.org/journals/isi

Enriching Song Recommendation Through Facial Expression Using Deep Learning

Shalaka Prasad Deore

Department of Computer Engineering, M.E.S. College of Engineering, S.P. Pune University, Pune 411001, India

Corresponding Author Email: shalaka.deore@mescoepune.org

https://doi.org/10.18280/isi.280126

Received: 5 November 2022
Accepted: 20 January 2023

Keywords:
convolutional neural network, expression, mood, decision making, mood identification, song recommendation

ABSTRACT

The music recommendation systems are highly linked with the emotional response of the user as the majority of the music is based on the mood of the listener. A large number of researches have been performed for the detection of emotion through the use of a variety of different techniques. These approaches have been helpful in achieving the emotion of the subject using various devices and other hardware which can be highly expensive with very low rates of accuracy. Whereas the detection of expression of the subject can be useful in determining the mood or the emotion with a considerable degree of accuracy. Therefore, to achieve the effective identification of emotion of an individual for effective music recommendation has been proposed in this research paper. The presented approach utilizes image normalization and Convolutional Neural Networks (CNN) which are trained on a dataset consisting of a number of different emotional responses. This trained model is then used to determine the mood of the individual and recommend music based on the detected mood. The experimental evaluation of the approach is performed to determine the accuracy of the emotion recognition which has resulted in highly accurate results. We achieved 62.88% testing accuracy with MSE and RMSE values of 8.5 and 2.9 respectively. The obtained results are promising and show that the fuzzy classification technique optimizes the outcomes.

1. INTRODUCTION

The expressions on one's face and variations in facial patterns provide information about the subject's emotional state and aid in the regulation of a good discourse with the topic. The subject's overall mood is deduced from these expressions, leading to a deeper comprehension of the subject. Facial expressions play a vital part in nonverbal communication and human relationships. Face expression analysis entails evaluating and visually identifying changes in facial characteristics as well as various facial gestures.

Understanding the issue is the most significant component of mood detection. If a subject is not understood, their problem cannot be handled. Being able to discern a person's mood helps one understand the issue better. Using this strategy, a doctor may readily have a deeper understanding of the patient. The identification of emotional response of a human being is one of the most common and day to day tasks that are performed with reasonable accuracy by all healthy individuals. These emotional responses are easily detected as humans have been evolved to communicate efficiently with non-verbal gestures. This has enabled us to achieve effective conversation and collaboration with one another to achieve completion of the complex tasks that can be easily fulfilled to fuel growth and development of the entire human race as a whole. The recognition of emotions has been one of the defining factors of the primates which have been crucial for immense development that we see nowadays.

The process of emotion recognition even though quite easily performed by human beings and other primates, can be quite difficult to identify by automated programs using computer vision methodologies. The expressions on one's face are important for understanding the individual, communicating with humans, and connecting with them. They also play an important part in a patient's medical rehabilitation and serve as a foundation for behavioral research. Mood detection based on the technique of taking face photographs main gives a highly practical method to non-invasive mood identification.

The lack of an effective approach for automatic identification of the emotion is highly complex endeavor that has been one of the most recurring problems in computer vision paradigm. The effective realization of the automatic emotion recognition can be applicable in a variety of different scenarios that can improve the accessibility and the implementation of a plethora of different use cases. The recommendation engines can be improved through the realization of the emotion recognition as the emotional responses are linked to the behavior and the listening patterns of the individual.

The Convolutional Neural Network analyzes the region of interest and calculates the convolution rate. Various features are calculated using filters. Then these features are passed through futher layers to extract more specific features. The softmax function converts a fully-connected layer's output from n-units to a probability distribution, which is then selects a response based on the subject's facial expression. This expression is then passed to the music recommendation module, which suggests suitable music to the user based on the emotional response detected. This study focuses on identifying the user's facial expression at different time and assisting suitable music depending on mood.

This research article dedicates section 2 for the evaluation

of the past works under the name literature survey. And the implemented technique is broadly described under the section proposed methodology which is numbered as 3. Section 4 discusses the obtained results. And finally, section 5 concludes this paper along with the scope for future enhancements.

2. LITERATURE SURVEY

The literature review is focuses on various ways of detecting mood based on facial expression using various machine learning approaches. Researchers are using many machine learning algorithms to detect mood efficiently. The following section elaborates the work done in this area.

Mao et al. [1] investigated a fresh problem of song suggestion depending on competency. They used a vocalist profile to model a singer's vocal ability, taking into account voice pitch, intensity, and quality. They suggested a supervised learning technique for training a speech quality assessment function, with the goal of computing voice quality at query time. Incompetence modeling, a simplified version of the singer profile is also provided to decrease the recording effort. They also presented a song model that allowed for vocalist matching. They were able to create a learning-to-rank strategy for music recommendation using human-annotated ranking datasets using the suggested models. The proposed approach's efficacy and benefits over two baseline techniques were proved in the trials.

Jao and Yang [2] described methods for generating exemplar-based representations for music and showed that utilizing the representation to train basic linear SVMs may result in high accuracy rates in tag-based music retrieval. The suggested techniques employ unlabeled data as exemplars in an overcomplete dictionary to compute feature representations, and then use labeled training data to construct a discriminative classifier for music auto-tagging depending on the calculated feature representations. To prevent losing short-time audio information due to the temporal integral and the redundancy of employing feature vectors from a restricted number of clips, the dictionary is made up of frame-level feature vectors randomly picked from a vast and diverse range of unlabeled music clips. The authors utilize this dictionary to build feature representations for both the training and test samples and learn a classifier using techniques like the support vector machine.

Rosa et al. [3] presented a personal music recommendation system that depends on a new lexicon-depended enhanced sentiment metric called eSM, which uses the Sentimeter-Br2 measure in conjunction with a unique correction factor depending on the user's profile. The authors show how a lexicon-based sentiment intensity meter combined with a correction factor may enhance the performance of a music recommendation system while utilizing a low-complexity approach that benefits consumer electronic devices. The correction factor is dependent on a person's attributes that may be easily derived from social media. Therefore, the new metric provides a more accurate sentiment value. To increase the user's quality of experience, the suggested recommendation system also examined ergonomic considerations of usability.

Sun [4] suggested a multilayer attention representation-based recommendation technique to distinguish the differences in music preferences among users. The system learns the embedded representations of songs from a multidimensional perspective and mines the preference correlations between users and songs using information such as user characteristics and song content. It primarily addresses the following problem: to mine different users' differential preferences for multidimensional features of the same song, an embedded representation depended on attention mechanism is proposed to learn song representations through user-based attention networks, and then build song-based attention networks to learn user preference representations depend on the learned song representations. A temporal relationship recommendation algorithm that depends on the attention network is proposed to learn temporal dependencies from listening behaviors and improve the accuracy of song recommendations to distinguish the degree of contribution of different historical behaviors to users' decisions.

Chin et al. [5] developed a new method for predicting the Probability Density Function (PDF) of music emotion throughout the emotion (VA) space. The concept is dependent on the assumption that the emotion distribution of an unknown music piece can be acquired by combining the PDFs of observed training pieces using the same set of combination coefficients that can be utilized to rebuild the unknown piece in audio space. Specifically, they demonstrated that the k-Nearest Neighbor (kNN) approach can be utilized to compute the combination coefficients and subsequently reliably forecast the emotion PDF in trials using the NTUMIR and MediaEval2013 datasets. The authors also show how the suggested technique may be utilized to retrieve music depending on emotions.

Lin et al. [6] offered an HK-ANN model for short-term music suggestions that use heterogeneous data for embedding. To begin, the authors incorporate graphical data, textual data, and visual data using TransR, PV-DM, and VAE, respectively. The sum of these embedding findings is a high-dimensional representation of the entity, which encompasses the majority of the heterogeneous information found on online music platforms. Then, to acquire short-term preferences for the user listening to music, the authors present an Recurrent Neural Network (RNN) model with an attention mechanism. Their algorithm has a better recommendation impact than the existing mainstream music recommendation model, according to the findings. More crucially, because most real-world data follow a long-tailed distribution, the general recommendation system will mostly propose popular goods to consumers.

For music mood categorization, a new approach called OMPGW is suggested by Mo and Niu [7] which give an adaptive time-varying description of music signals with improved spatial and temporal precision. The suggested method is dependent on a combination of three signal processing methods and is used to extract audio features. Spectral centroid, spectral roll-off, spectral flux, spectral bandwidth, spectral contrast, spectral flatness measure, spectral contrast, sub-band power, frequency cepstrum coefficient, and coefficient histogram are among the 10 ATF characteristics derived using the OMPGW approach. The ATF characteristics surpass all other algorithms, demonstrating the superiority of the suggested method. Experiments using five mood-annotated datasets resulted in the effectiveness of the recommended strategy being clarified.

Chang et al. [8] examined a Personalized Music Recommendation System (PMRS) that uses a CNN approach and a Collaborative Filtering (CF) recommendation algorithm. The CNN method categorizes music depending on the audio signal that is present in the song. The CNN extracts hidden information from auditory signals in music and categorizes the music accordingly. In PMRS, the authors offer a CF method

for extracting the user's history from the log file as well as the CNN output for recommending music in various music genres. To explain how the PMRS algorithm works, they created an Android music application. The algorithm is trained using the publicly accessible million song dataset.

Guo et al. [9] introduce a session-aware recommendation model that integrates RNN and CNN, to develop the user's long-term preferences as well as sequential intent in the current session. The findings of the experiments show that using historical information can help increase the accuracy of sequential-related suggestions. Using a CNN and RNN to process the long-term historical relationships and the short-term consecutive interactions, respectively, author proposed a Joint Neural Network (JNN) in this research for session-aware suggestions.

Dridi et al. [10] offered a recommendation technique in which the algorithm infers the context in which music tracks should be chosen. To address the CARS problems, the authors first conducted a literature study of several contextual recommendation techniques concentrating on scenario inference. To increase the quality of music recommendations, they highlight the reason and necessity of detecting the present contextual situation from interactive contextual factors in a recommender system. Their research focuses on fuzzy rules that include associated contextual aspects to infer the user's present state. Finally, depending on the detected contextual circumstance, relevant music recommendations are created. The suggested recommender outperformed typical contextual recommendation algorithms, according to data.

Zhu et al. [11] presented an approach for Recommender Systems that depends on the Collaborative Filtering technique. The BitSet optimization procedure and a unique similarity measure are used in this strategy. The design approach of the suggested solution is novel: the authors are dealing with a hybrid method that combines memory-based and model-based approaches. In this way, they gain positive outcomes in several different areas that don't generally go together, such as accuracy and prediction time with memory-based techniques, and recommended explanations and updated results with model-based approaches. The suggested approach's significance is determined by the balance and quality of its outcomes: it is a technique that achieves good accuracy, forecast time, and time to create and update its model.

To enhance the personalization and efficiency of recommendation, a course instructor recommendation system (FCTR-LFM) depended on fuzzy clustering and latent factor model (LFM) was developed by Yao and Deng [12]. The major task consists of defining a series of ways to achieve quantitative results of teacher characteristics, course features, and teaching performance under the guidance of pedagogy standards. The experimental dataset will be built using these results. A high-dimensional sparse evaluation matrix is utilized to convey the data. A fuzzy clustering model for teachers is created, allowing instructors to automatically cluster depending on their features and solving the problems of sparsity reduction in the assessment matrix and cold beginning of teachers. The evaluation matrix is decomposed into the product of two low-dimensional matrices including implicit components using the enhanced LFM.

Rodrigo [13] presented the key principles and ideas of fuzzy logic and its applications, with a focus on the fuzzy approximation theorem and fuzzy inputs as latent spaces. Second, he provides an updated study of fuzzy logic's applicability in musical applications. The Fuzzy Logic Control Toolkit (FLCTK), a suite of tools for creating musical material in the MaxMSP real-time sound synthesis environment, is presented third. Fourth, he shows how applications in sound synthesis, algorithmic composition, and many-to-many musical mappings may be used. Finally, he goes through some of the compositional components of Incerta, an acousmatic multichannel piece created with the FLCTK in MaxMSP.

The literature survey reveals that the mood detection based on facial expression is very important research area. It is also very helpful in medical area. Patient's expression assists doctor in evaluating his/her mental health on each visit. The method can be used in a variety of situations, including therapy and psychoanalysis to ascertain the patient's mood.

In this study, according to the mood music is recommaned. Reecommandation system will identify moods like happy, sad, fearful, angry, disgusted, nutral, surprised etc.

3. PROPOSED SYSTEM FOR FACIAL EXPRESSION EVALUATION

The presented technique is for achieving the recognition of facial expression for the purpose of enabling appropriate music recommendation. In the proposed method based on facial expression the mood of the user will be recognized and accordingly music is recommended by the system. The architectural diagram of music recommendation system consists of number of steps: Preprocessing, Image normalization, CNN model training and then fuzzy classification as depicted in Figure 1.

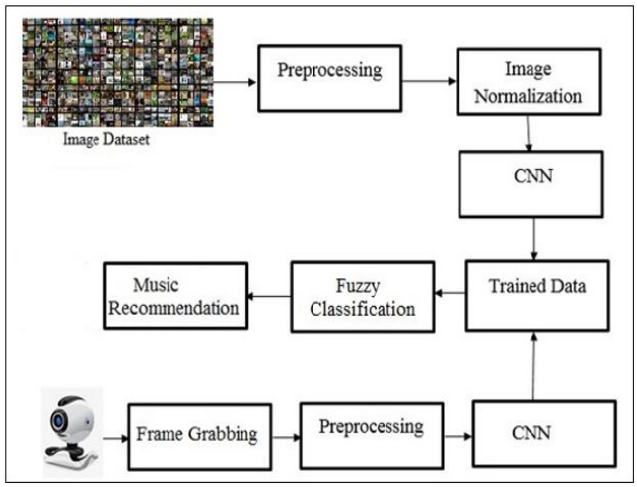

Figure 1. Block diagram of music recommendation system through facial expression

The presented approach achieves the prescribed goals through the use of the following steps.

Step 1: Preprocessing and image Normalization – Before the commencement of the training, the facial expression images with different expressions such as, angry, disgusted, fearful, happy, neutral, sad, and surprised are resized to the dimension of 48 × 48 as the resultant height and width.

In this step of the presented approach, an ImageDataGenerator object is constructed utilizing keras and tensorflow libraries with the rescale ratio of 1./255 for the image assessment for both the training and testing datasets. This procedure is performed for the images of all the facial expressions. ImageDataGenerator class includes the locations

of the training and testing directories, the setting of the image dimensions to 48 × 48, the batch size to 64, the activation of the grayscale colour mode, and to categorical class.

Step 2: Convolutional Neural Network – The detection of the facial expression is actually performed in this module of the proposed methodology. This is one of the most essential modules that take the original image as well as the skin object as an input to this step of the methodology. A dataset consisting of the facial expressions. The training of the CNN approach through this dataset achieves a model file with an extension of .h5. The dataset being provided as an input from the above URL is already split into training and testing folders. These folders are again segregated into folders specific to a particular expression, namely, angry, disgusted, fearful, happy, neutral, sad, and surprised. These facial expression images are fed into the CNN model as part of the training process. The supplied facial images are first downsized to a length and width of 48 x 48 pixels. Upon those images, the model is trained for 50 epochs with a batch size of 64. In the python environment, the TensorFlow and Keras libraries are utilized to enable the individual elements of the CNN model. The specification of the model is depicted in Table 1.

Table 1. CNN model specification

Layer	Activation
CONV 2D 32x3x3	RELU
CONV 2D 64x3x3	RELU
MaxPooling2D	
Dropout 0.25	
CONV 2D 28x3x3	RELU
MaxPooling2D	
Dropout 0.25	
Dense 1024	RELU
Dropout 0.25	
Dense 7	Softmax
Optimizer	Adam

The deep Convolutional Neural Network obtained by this architecture is then employed as a resultant model .h5 file to execute for the 50 epochs. The testing images obtained in the dataset as an input from the very first stage of the proposed technique is then fed into the trained model.

Step 3: Testing and Fuzzy Classification – The testing of the system for the facial expression is performed in this step of the methodology. The OpenCV library is utilized to initialize the integrated webcam of the laptop and the user's face is captured as a test image. This image is then subjected to the Haar Cascade classifier which performs the detection of the facial region in the image. Once the facial region is detected, the image is subjected to the trained model in the form of the .h5 file to obtain the best matched expressions.

The best matched expressions derived from the trained .h5 file through the Convolutional Neural Network are continuously performed and added into a list. The different detected expressions are counted in the list. The derived best matched expressions from the captured image are searched for the maximum count, whereas the minimum count is set as 1. The difference between the maximum and minimum count is calculated, which is then divided into 5 divisions as VERY LOW, LOW, MEDIUM, HIGH, and VERY HIGH and considered as the fuzzy crisp values.

The expressions are then evaluated for their presence in the VERY HIGH categories based on a predetermined threshold. Once an expression is detected, it is then utilized for the playing of the music. The system retrieves the appropriate music file randomly from the expression specific preselected folders of .mp3 files and starts to play the media.

4. RESULTS AND DISCUSSIONS

A Java and Python programming languages was used to produce the suggested methodology for the goal of identifying the user's mood utilizing the Convolution neural network. The Sypder and NetBeans IDEs were used to create the proposed technique. The development computer was equipped with an Intel Core i5 processor, 8 GB of RAM, and 1 TB of storage.

The integrated web cam of the notebook or an additional camera device is both employed to capture the individual's facial pictures. The Convolutional Neural Network is the primary model that must be reviewed in order to accurately identify the user's mood. The performance assessment is necessary to recognize any inconsistencies in the model's implementation. The evaluation procedure is outlined below.

4.1 Performance evaluation using Root Mean Square approach

A number of experiments were carried out to determine the error obtained by the proposed methodology, the mechanism for facial expression identification using Convolutional Neural Networks. The inaccuracy attained by the methods for appropriate emotion identification of the user may be used to determine the performance criteria.

The RMSE, or Root Mean Square Error, is used to determine the error achieved by the provided technique. The existence of inaccuracy in the proposed technique for mood identification using facial expression and CNN is indicative of the stipulated approach's computational efficiency. The RMSE method simplifies the assessment of errors between two continuously correlated metrics. The accuracy of mood detection and the inaccuracy of mood detection are the metrics evaluated in this technique. These numbers are calculated, and the error is calculated using Eq. (1).

$$RMSE = \sqrt{\sum_{i=1}^{n} ((x1, i - x2, i)^2 / n)} \quad (1)$$

where, $\sum (x_1 - x_2)^2$ is differences squared for the summation in between the expected no. of mood identifications and the obtained no. of mood identifications and n is number of trails.

Ten different users who completed 10 tests on the machine with various facial expressions were used to measure these two factors. Table 2 provides a summary of these calculation findings.

The findings of the empirical assessment of the technique have made it easier to visualize the error rate graphically, as shown in Figure 2. The graph shows the comparison of mean square error between expected and actual mood recognition based on facial expression during testing phase. The graph shows how accurately the system can anticipate users' moods based on their facial expressions with a minimum level of error. This is due to the very precise deployment of the Convolutional Neural Network, which significantly improves detection performance. The MSE and RMSE values of 8.5 and 2.91, respectively, show that the Fuzzy classification strategy optimizes the results. Our training accuracy is 97.64%, while

real-time testing accuracy is 62.88%. This assessment demonstrates the precise and accurate implementation of the mood detection technique for music selection using face recognition.

Table 2. Mean square error measurement

Sample No.	No. of expected mood identification	No. of obtained mood identification	MSE
1	10	9	1
2	10	8	4
3	10	6	16
4	10	7	9
5	10	5	25
6	10	6	16
7	10	9	1
8	10	7	9
9	10	8	4
10	10	10	0

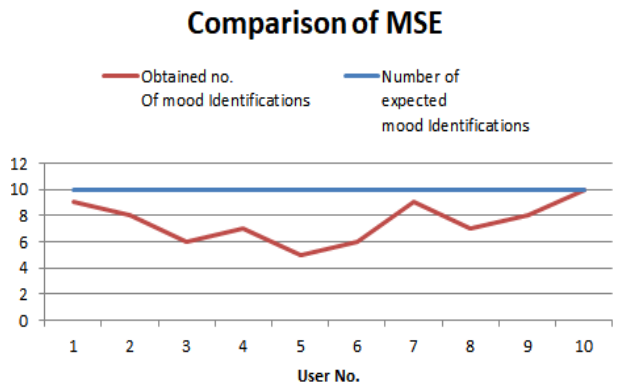

Figure 2. Comparison of MSE in between Expected v/s obtained number of mood identifications

5. CONCLUSIONS

The research framework for the objective of mood identification through identification of face expression has been came to the realization by the use of Convolutional Neural Networks and Fuzzy Classification. The frames displaying the individual's face are captured using the live video from a web camera. After appropriate preprocessing, these images are efficiently retrieved and used for assessment. The Haar Features are utilized to identify face regions from the preprocessed frames. These images are then passed along to the following phase for face detection. The model is being used face for mood identification, resulting in a masked black-and-white image. The face detected is colored as white and the other parts of the image are converted to black to form a binary image. The CNN method includes this binary picture as an input to do facial expression recognition. The Fuzzy classification is used to classify and to obtain accurate mood detection, which will then be utilized to propose and play the appropriate music. The performance of proposed model for mood detection was proven by the experimental assessment. We achived 62.88% accuracy on real-time testing data and training accuaracy is 97.64%.

Future research options for this facial emotion recognition method can be implemented using various deep learnig models. We are also planning to implement compact convolution network by reducing some training parameters.

REFERENCES

[1] Mao, K., Shou, L., Fan, J., Chen, G., Kankanhalli, M.S. (2015). Competence-based song recommendation: Matching songs to one's singing skill. IEEE Transactions on Multimedia, 17(3): 396-408. https://doi.org/10.1109/TMM.2015.2392562

[2] Jao, P., Yang, Y. (2015). Music annotation and retrieval using unlabeled exemplars: Correlation and sparse codes. IEEE Signal Processing Letters, 22(10): 1771-1775. https://doi.org/10.1109/LSP.2015.2433061

[3] Rosa, R.L., Rodríguez. D.Z., Bressan, G. (2015). Music recommendation system based on user's sentiments extracted from social networks. In IEEE International Conference on Consumer Electronics (ICCE), pp. 383-384. https://doi.org/10.1109/ICCE.2015.7066455

[4] Sun, J. (2022). Variational fuzzy neural network algorithm for music intelligence marketing strategy optimization. Computational Intelligence and Neuroscience, 2022: 1-10. https://doi.org/10.1155/2022/9051058

[5] Chin, Y., Wang, J., Wang, J., Yang, Y. (2018). Predicting the probability density function of music emotion using emotion space mapping. IEEE Transactions on Affective Computing, 9(4): 541-549. https://doi.org/10.1109/TAFFC.2016.2628794

[6] Lin, Q., Niu, Y., Zhu, Y., Lu, H., Mushonga, K.Z., Niu, Z. (2018). Heterogeneous knowledge-based attentive neural networks for short-term music recommendations. IEEE Access, 6: 58990-59000. https://doi.org/10.1109/ACCESS.2018.2874959

[7] Mo, S., Niu, J. (2019). A novel method based on OMPGW method for feature extraction in automatic music mood classification. IEEE Transactions on Affective Computing, 10(3): 313-324. https://doi.org/10.1109/TAFFC.2017.2724515

[8] Chang, S.H., Abdul, A., Chen, J., Liao, H.Y. (2018). A personalized music recommendation system using convolutional neural networks approach. In IEEE International Conference on Applied System Invention (ICASI), pp. 47-49. https://doi.org/10.1109/ICASI.2018.8394293

[9] Guo, Y., Zhang, D., Ling, Y., Chen, H. (2020). A joint neural network for session-aware recommendation. IEEE Access, 8: 74205-74215. https://doi.org/10.1109/ACCESS.2020.2984287

[10] Dridi, R., Zammali, S., Arour, K. (2017). Fuzzy rule-based situational music retrieval and recommendation. In IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), pp. 549-556. https://doi.org/10.1109/AICCSA.2017.154

[11] Zhu, B., Hurtado, R., Bobadilla, J., Ortega, F. (2018). An efficient recommender system method based on the numerical relevances and the non-numerical structures of the ratings. IEEE Access, 6: 49935-49954. https://doi.org/10.1109/ACCESS.2018.2868464

[12] Yao, D., Deng, X. (2020). Teaching teacher recommendation method based on fuzzy clustering and latent factor model. IEEE Access, 8: 210868-210885. https://doi.org/10.1109/ACCESS.2020.3039011

[13] Rodrigo, C. (2020). Creating music with fuzzy logic. Frontiers in Artificial Intelligence, 3: 2020. https://doi.org/10.3389/frai.2020.00059