Multimedia Tools and Applications (2020) 79:765–783
https://doi.org/10.1007/s11042-019-08192-x

Recognition of emotion in music based on deep convolutional neural network

Rajib Sarkar[1,2] · Sombuddha Choudhury[1] · Saikat Dutta[1] · Aneek Roy[1] · Sanjoy Kumar Saha[1]

Received: 10 November 2018 / Revised: 1 August 2019 / Accepted: 6 September 2019 /
Published online: 13 September 2019

Abstract
In the domain of music information retrieval, emotion based classification is an active area of research. Emotion being a perceptual and subjective concept, the task is quite challenging. It is very difficult to design signal based descriptors to represent emotions. In this work deep leaning network is proposed and experiment is done with benchmark datasets namely, *Soundtracks*, *Bi-Modal* and *MER_taffc*. Experiment has also been done with hand crafted descriptor consisting of different time domain and spectral features, linear predictive coding and MFCC based features. Different classifiers like, neural network, support vector machine and random forest are tried. Although the combined feature set with neural network provides an optimal result for the datasets, but in general the performance of such approaches is limited. It is difficult to obtain a consistent feature set that works across the classifier and datasets. To get rid of the issue of feature design, deep learning based approach is followed. A convolutional neural network built around VGGNet and a novel post-processing technique are proposed. Proposed methodology provides substantial improvement of performance for the datasets. Comparison with other reported works on three different datasets also establishes the superiority of the proposed methodology. The improvement in performance has been substantiated by Z test.

Keywords Music emotion recognition · Convolutional neural network · Deep learning · Audio features

1 Introduction

Every piece of music is associated with an emotion and accordingly it generates an intuitive feeling to the listener. Identification of inherent emotion present in a music is an active area of research [23, 26, 62]. Despite the use of sophisticated techniques, identification of emotional category of musical excerpts is quite challenging. This is mainly due to the

✉ Rajib Sarkar
rjbskar@gmail.com

Extended author information available on the last page of the article.

subjectiveness of emotion. The perception of emotion may vary from person to person. Moreover, the conveyed emotion depends not only on the structural features of music but also on the state (gender, age, personality etc.) and contextual aspects (like occasion and place) of the listener. This makes the task of emotion based classification of music further difficult.

Automatic classification of music signal according to different attributes like singer, genre, emotion is an important task. It helps in organizing the repository in a structured manner and also enables effective retrieval of desired music. With the rapid growth in the size of digital music libraries, manual classification is impossible. Hence, an automated system is in demand. Considerable efforts have been put on genre or singer based classification. But, emotion being related to state of the mind is an important criteria of retrieval and demands more attention. A listener may like to make the choice according to his/her mood. Thus, music emotion recognition (MER) becomes useful as it helps to group the songs according to their emotion automatically. Such categorization can act as a fundamental step for developing emotion based music recommendation system and also can be utilized in the applications like, music therapy [52] and cognitive analysis. This observation has motivated us to focus on emotion based classification.

One basic approach for classification is to compute the descriptors from the audio signals and then feed them to certain classifier [16, 18, 19, 43, 64]. But, success is limited for such systems as it is difficult to represent emotion by means of low level of features. In this context, deep learning has drawn attention. It has already achieved significant outcome in various tasks of computer vision [29, 55] and natural language processing [4]. In recent times deep learning approaches are being tried for speech emotion recognition [3, 21, 22, 38, 58]. A very few attempts [10, 36] are reported for music emotion recognition. Keeping the complexity of representing emotion in mind and inspired by the success of deep learning in image, video and speech, we have considered deep learning based approach in our work.

In this work, the problem at hand is to classify the music signal according to emotion. Four broad classes of emotion like happy, anger, sad and neutral have been considered. These classes conform to the four quadrants of the model suggested by Thayer [56] and Russell [48]. In this direction, a convolutional deep learning network is proposed that helps us to extract the meaningful features. Moreover, the burden of designing the low level descriptors is removed. Performance of the proposed system is evaluated on three popular music emotion datasets.

The contribution of the work lies in customizing VGGNet which is used in image classification problem. The network has been modified and made lighter by reducing number of layers. The network classifies the audio segments in the clips in to emotional categories. Finally, a simple but novel post-processing technique has been applied on the labelled segments to determine the emotional category of the audio clip as a whole. Rest of the paper is organized as follows. Survey of past work is presented in Section 2. Section 3 elaborates the proposed methodology. Experimental results and concluding remarks are put in Sections 4 and 5 respectively.

2 Past work

Music emotion recognition (MER) has drawn the attention of the researchers over a decade. Still it remains as an active area of research [10, 36, 63, 65]. It is observed that two major steps are involved in the process: designing the suitable features to describe the music signal and thereafter identifying the emotion. Features may be conventional hand crafted ones as considered by most of the works or learnt features which has become the trend with the advent of deep learning. Using the features regression based approach can be followed to map the music into emotion plane suggested by the model of Thayer [56] and Russell [48]. The alternative approach is to rely on the classifier. In this section, we present brief survey on the features and emotion identification approaches.

A wide variety of hand crafted features have been used by the researchers. The patterns inherent in a music signal provide the perception of emotion [30]. Features are used to summarize the patterns. Energy or the power of a music clip is frequently used [18, 19, 33, 49, 65] as it has very correlation with arousal [15]. A music clip with fast tempo is often correlated with positive valence and slow tempo is correlated with negative valence [15]. Hence, use of tempo is also very common [24, 37, 49, 53, 60, 61]. Timbral features, captured in different forms are also utilized by the researchers. Such features include Mel-frequency cepstral coefficients (MFCC) [22, 34, 35, 43], Daubechies wavelets coefficient histogram (DWCH) [37, 60, 61]. Zero crossing rate (ZCR) [35, 39, 65] and pitch [2, 43, 65] are also useful. Variants of Spectral features [34, 39, 65] like spectral rolloff, spectral flux as well as tonality [24, 61] are also considered in various works. Panda et al. [44] extracted rhythmic, dynamics, melodic, harmonic and tonality based features.

As it is not an easy task to design hand crafted features for a given goal, in recent time considerable efforts have been put to learn the features using deep network. Researchers experimented with deep learning techniques to perform [66], video data [25], facial images [67] etc. For acoustic audio data, most of the works are on speech emotion recognition [2, 12, 22, 27, 58]. It is still worth to follow those to understand the applicability of deep learning in the context of music signal. Few efforts [10, 36, 50] are directed towards music also. Convolutional Neural Network(CNN) has been tried by number of researchers [36, 38, 58]. Most commonly, a CNN is fed with spectrograms generated from audio signals. A series of convolution and pooling operation is performed on it to build the feature vector. Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) [10, 58] has been considered. For RNN, input is the raw audio signal and LSTM divides into number of frames.

To recognize the emotion, regression based approach has also been followed. Emotion in music can be represented as two orthogonal components- *Arousal* and *Valence*. *Arousal* of a music represents energy, activation or intensity whereas *valence* denotes how pleasant a music is. Several two-dimensional models have been proposed of which Russell's [48] and Thayer's [56] are widely used. Figure 1 is a simple representation of circumplex model

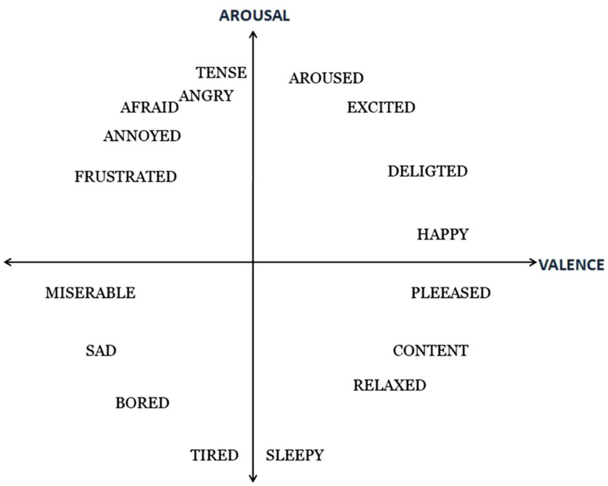

Fig. 1 Two Dimensional Emotion Plane: Valence vs. Arousal

proposed by Russell where X and Y axis denote *valence* and *arousal* respectively and it shows the position of different emotional classes in the plane. In this approach, music clips in the training set are annotated with *valence* and *arousal* values and it is used to prepare the 2D emotion plane. Regression model is formed to predict *arousal* and *valence* by considering the low level features as the observed values. Separate regression models can be trained for *arousal* and *valence* [60]. The regressed value for both are used to find the position of the song in the 2D emotion plane describing the emotion. Researchers have worked with different regression models for predicting *valence* and *arousal* values. Yang et al. [61] experimented with three different regression algorithms namely, multiple linear regression (MLR), support vector regression (SVR) and AdaBoost.RT (BoostR) with a feature set consisting of Spectral Contrast, DWCH (Daubechies wavelets coefficient histogram) and features obtained from PsySound [7] and MARSYAS [59]. Seo et al. [53] extracted acoustic features like average height, width of the wavelengths, peak average, beats per minutes (BPM) etc. They have used SVR to predict the emotion of Korean pop(k-pop) genre songs. SVR is also used by Han et al. [24]. They have used Scale, Average Energy, Harmonics and Rhythm as musical features. Gaussian process regression (GPR) [39] is also applied on multiple sets of features extracted by MARSYAS. As the annotations are collected through surveys, the issue of inconsistency remains while training the models.

In classifier based approach music clips are first represented by a set of features. Thereafter, feature vector is fed as input to the classifier for emotion recognition. Commonly used classifiers include support vector machine (SVM) [18, 19, 35], artificial neural network (ANN), radial basis function ANN (RBF-ANN) [43], Gaussian mixture model (GMM) [33, 68], Random Forest [65]. Researchers have experimented with different parameter and kernel setups for the classifiers. In some cases [43], Principal component analysis (PCA) and linear discriminant analysis (LDA) have been used for reduction of feature dimension.

It is observed that variety of features and classifiers/regression models have been considered by the researchers. But success of all such systems are quite limited. Hence, emotion based categorization still remains an active area of research.

3 Proposed methodology

In general, for classification problem, designing a uniform set of features that works across various datasets and classifiers is very critical. Emotion being very much subjective and psychological issue, it is further challenging. Limitations of hand picked low level features (mostly designed intuitively) affects the performance of a classifier. It has motivated us to apply deep learning network to design a set of features that will work more consistently for different datasets. The audio signal, may be pre-processed is fed to the deep network to learn the complex structural factors of music contributing to emotion.

Proposed methodology consists of three stages. At first, the audio signal is pre-processed to represent it into a concise but meaningful form which is fed to our convolutional neural network. A post processing is applied on the prediction output of the network. Pre-processing steps, proposed network architecture and the post-processing steps are elaborated in the following sections.

3.1 Pre-processing

The raw audio signal goes through a sequence of steps before being fed to the network. Each clip is normalized so that sample amplitudes are restricted within $[-1, 1]$. The music clip is divided into number of segments – each of 5 seconds duration [5, 42, 57, 62]. These small segments are used as the unit to perceive the emotion. It makes the task more challenging.

A two dimensional spectrogram [46] is computed for the segment and used as the input for the proposed network. Past study indicates that spectral features play important role in identifying the emotion. Spectrogram is our choice for input as it summarizes spectral information in a concise form. Moreover, convolutional neural networks (CNN) have shown promising performance on image data. Spectrogram being a pictorial representation, it can be utilized as the input for the networks similar to those used in image and vision problem.

To obtain the spectrogram, the audio segment is divided into number of frames with equal size with an overlap among the consecutive frames. In our work, a frame consists of 1024 samples. The spectrogram is obtained by taking short-time Fourier transform on the frames. Thus, it reflects time-frequency spectrum of the signal. The horizontal and vertical axes denote time (frame number) and frequency respectively. An element of the spectrogram shows the energy of a frequency component at an instance. Thus, energy variation of various frequency components over time is captured in the spectrogram. The frequency scale is converted from linear scale to mel-scale as it resembles human auditory system. To reduce the dimension, the mel-scale is divided into 128 bins. The logarithm of the values are considered to dampen the effect of large magnitude. Log values are scaled by using standardization procedure *i.e.* mean subtraction and division by the standard deviation. In our work, the spectrogram is formed using 196 frames. Thus, the dimension of spectrogram becomes 196×128 and it is fed to the network. Thus, the pre-processing steps can be summarized as follows.

- Amplitude of each music clip is normalized with in $[-1, 1]$.
- Each clip is divided into number of segments of 5 seconds duration.
- A two dimensional log magnitude mel-scale spectrogram is computed for the segment and used as the input for the proposed Deep Learning network.

3.2 Proposed network architecture

Convolutional neural network (CNN) is biologically inspired architectures characterized by their local receptive structures, sparse connectivity and shared weights. It has been successfully applied in image processing [54] tasks and also in speech recognition [1].

Two dimensional convolution has been applied along dimensions of time and frequency on the input spectrogram. Every layer of convolution has a fixed number of filters which convolve with the inputs to the corresponding layer and produces feature maps. We denote the m-th feature map of the k-th layer as h^k_m. Corresponding input and bias for the k-th layer are x^k and b^k respectively. For the m-th feature map of k-th layer weight is W^k_m. Elements of h^k is obtained as follows:

$$h^k_{ijm} = \sigma((W^k_{ijm} * x^k) + b^k)$$

where σ is some non-linearity function and * denotes convolution operation.

Unlike conventional ANNs, not all neurons in a layer are connected to all neurons in the next layer. The neurons in a layer respond to activation that falls within its own receptive area. As layers are stacked, the receptive areas of the neurons become increasingly global. This helps capture both short and long term dependencies which are extremely significant in case of audio. Again, in CNN, the filters are replicated across a layer enabling the sharing of parameters. This ensures that same features are detected regardless of their position contributing to translational invariance.

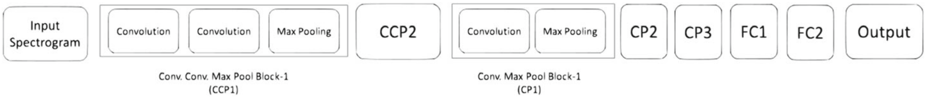

Fig. 2 Block diagram of the proposed convolutional neural network. CCP, CP and FC stand for *Convoluton-Convolution-Pooling*, *Convolution-Pooling* and *Fully Connected* respectively

Convolution layers are typically followed by pooling layers where the convolved data is down sampled usually by considering the maximum or average values for every small subsection of the matrix. Pooling helps to reduce the number of parameters in the model, thereby reducing over fitting concerns. It also helps to achieve translational invariance.

The proposed architecture is built around VGGNet [54]. In the proposed model, we have considered fewer layers. It alleviates the problem of over-fitting on the small sized training datasets. Figure 2 shows the schematic diagram of the proposed network. The first two blocks of the network are referred as CCP blocks. One such block consists of two convolution layers followed by a max pooling layer. It is then followed by three blocks (referred as CP), each consisting of alternating layers of convolution and max pooling. Finally, there are three fully connected (FC) layers. The last one is with the same dimension as the number of output classes. The detailed architecture is given in Table 1.

Convolution is performed along both time and frequency axes using small square filters of size 3×3. A fixed stride length of 1 is used for all the convolution layers. The number of filters is progressively increased in the later blocks of the network. We are motivated to use small kernel dimensions for convolution to reduce the number of trainable parameters.

Table 1 Architecture of the proposed convolutional neural network (CNN)

Data shape	Layer type	Description
$196 \times 128 \times 1$	Input	Log mel spectrogram
$196 \times 128 \times 64$	Conv	Kernel: 3×3 Stride: 1×1
$196 \times 128 \times 64$	Conv	Kernel: 3×3 Stride: 1×1
$98 \times 64 \times 64$	Max Pool	Kernel: 2×2 Stride:2×2
$98 \times 64 \times 64$	Conv	Kernel: 3×3 Stride: 1×1
$98 \times 64 \times 64$	Conv	Kernel: 3×3 Stride: 1×1
$49 \times 32 \times 64$	Max Pool	Kernel: 2×2 Stride: 2×2
$49 \times 32 \times 128$	Conv	Kernel: 3×3 Stride: 1×1
$16 \times 10 \times 128$	Max Pool	Kernel: 3×3 Stride: 3×3
$16 \times 10 \times 128$	Dropout	Keep prob. = 0.75
$16 \times 10 \times 256$	Conv	Kernel: 3×3 Stride: 1×1
$5 \times 3 \times 256$	Max Pool	Kernel: 3×3 Stride: 3×3
$5 \times 3 \times 256$	Dropout	Keep prob. = 0.75
$5 \times 3 \times 256$	Conv	Kernel: 3×3 Stride: 1×1
$1 \times 1 \times 256$	Max Pool	Kernel: 3×3 Stride: 3×3
$1 \times 1 \times 256$	Dropout	Keep prob. = 0.75
256	Fully Connected	Flattened to 1D tensor with 256 neurons
256	Dropout	Keep prob. = 0.5
256	Fully Connected	256 neurons
256	Dropout	Keep prob. = 0.5
4	Softmax	4 output classes

Instead of alternating convolution and max pooling layers, in the CCP blocks we have used two convolution layers one after the other. This is to realize larger sized filters at a lower cost. Max pooling layer is used to down sample the data. The kernel size for pooling is 2×2 in the CCP blocks and 3×3 for the CP layers. Once again the increase in size for the later blocks reduces the number of weights in the flattening layer. It is observed that further increase in the depth or width of the layers did not improve the performance of the proposed model. The output of the last pooling layer is flattened and then fed to the fully connected layer.

L2 regularization is applied to the weights of the fully connected layers. Adding the regularization component will drive the values of the weight matrix down. This will effectively

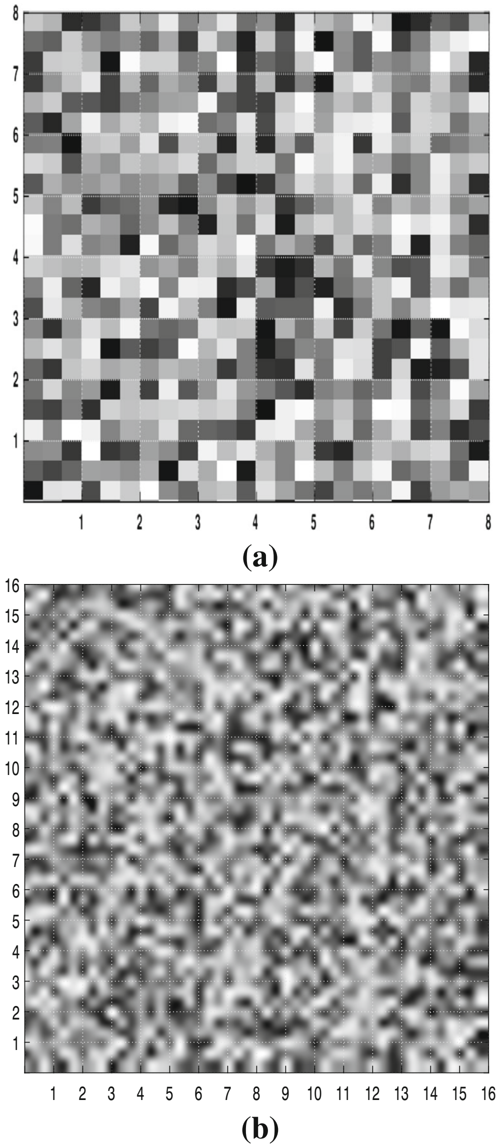

Fig. 3 Visualization of filters. **a** first convolution Layer (64 filters of kernel size 3X3). **b** seventh convolution Layer (256 filters of kernel size 3X3)

decorrelate the neural network and reduces over-fitting. The last three CP blocks and the FC layers are followed by a dropout layer to further prevent over-fitting on the training data. ReLu activation [41] is non-saturating and allows faster training. Hence it is used for convolution and first FC layer instead of sigmoid or tanh activation. The activation function is defined as $ReLu(x) = max(0, x)$ where x is an input to a neuron. For the last FC layer, Softmax activation is used and it is represented as $\sigma(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$ where z is a multidimensional vector having as many dimensions as the number of output classes. Thus, the tasks carried out different layers can be summarized as follows.

- Convolution is performed on the input spectrogram using small square filters of size 3×3.
- Pooling layer down samples the convolved spectrogram.
- L2 regularization is applied to the fully connected layers to prevent over-fitting.
- Dropout layers added to further prevent over-fitting on the training data.
- Convolution, pooling, fully connected and drop out layers appear in the network in accordance with the architecture summarized in Table 1.
- Finally, Softmax activation is used to predict the emotion at segment level.

Figure 3 shows the filters corresponding to the first and last convolution layer. There are 64 and 256 filters of size 3×3 in those layers respectively. Initially, from the input spectrogram the filters capture the variation of signal energy with frequency or time or both. It can be thought of as equivalent to determining the edges of different orientations in case of image. Dominating frequency components at different instances are highlighted by max pooling. At later stages, convolution leads to further abstract representation. Figure 4 shows

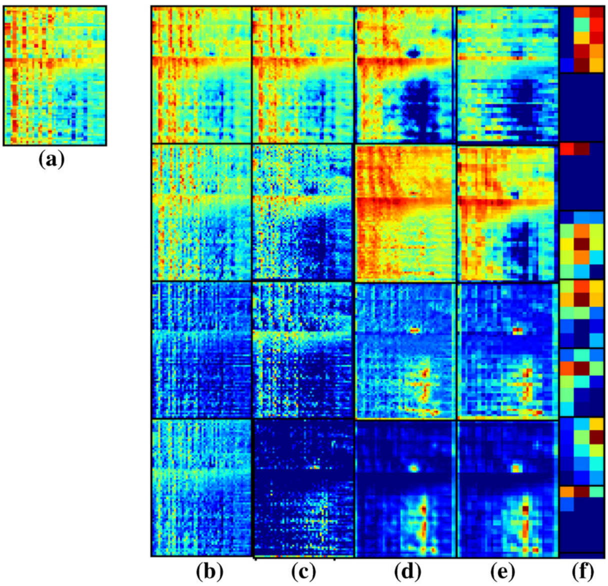

Fig. 4 An input spectrogram and output after different convolution layer: **a** input spectrogram and few filtered output after **b** first convolution layer, (c) second convolution layer, **d** fifth convolution layer, **e** sixth convolution layer, and **f** seventh convolution layer

one input spectrogram and output after various convolution layers. In the spectrograms blue denotes minimum energy and red corresponds to maximum. For each layer few sample spectrograms have been shown. It is noted that in the initial stages, local details are visible and gradually those are summarized. Figure 5 shows the output after last convolution layer corresponding to music clips of four different emotions. For better visualization, instead of 256 spectrograms of size 5 × 3, first 24 have been shown for each clip. It is also observed that they differ considerably for different emotions.

3.3 Post-processing

The network estimates a class label for every segment of each test clip. Different segments may be identified as different emotional class. We used a combination of voting and run-length based technique to estimate the label for the whole clip. A clip consists of number of segments and different segment may be predicted as different class. Voting strength for a clip belonging to a particular class is determined by the number of segments in the clip predicted as that class. More the segments predicted as a class, higher is the corresponding voting strength. But it does not take care of the position of occurrence of the segments in the clip. A sequence of segments with same label may also set the emotion of the clip. To capture this aspect run-length strength is introduced. A run is formed by the consecutive segments with same label and number of segments in the run is the run-length. Finally, the opinion is formed by the weighted combination of two strengths. The post-processing steps for a clip are detailed as follows.

- Let N_s be the number of segments in the clip.
- Let N_{c_i} be the number of segments predicted as i-th class.
- Voting strength for i-th class, $VS_i = \frac{N_{c_i}}{N_s}$.
- A clip may have multiple runs labelled as i-th class and $\{RL_{i1}, RL_{i2} \ldots\}$ denote the run-lengths for i-th class in the clip.
- Run-length strength for i-th class, $RLS_i = \frac{L_i}{N_s}$. Where, $L_i = max\{RL_{ij}\}$.
- Score that the clip belongs to i-th class, $S_i = w_1 \times VS_i + w_2 \times RLS_i$. Where w_1 and w_2 are two weights and $w_1 + w_2 = 1$.
- Finally, a clip is labeled as class k if $S_k = max\{S_i\}$.

Associating run-length information helps to capture the effect of persistence of an emotion. A listener is able to conceive emotions having prolonged spans (even if they are fewer)

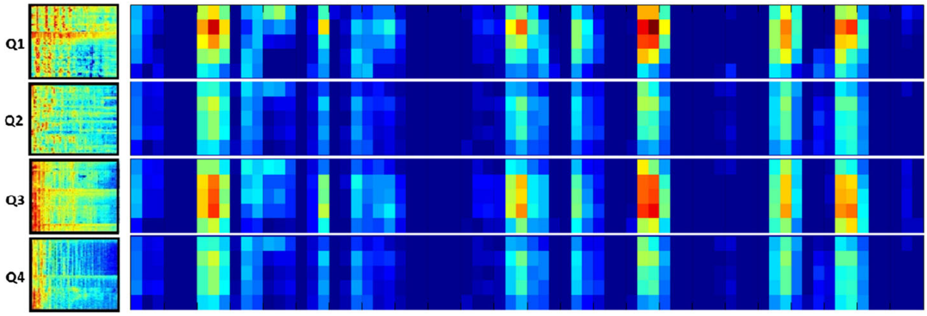

Fig. 5 Input spectrogram for four emotions and the output after seventh convolution layer where Q1, Q2, Q3 and Q4 denote happy, anger, sad and tender respectively

better than those having multiple short spans [5, 13]. Hence, w_2 should be more than w_1 to emphasize run-length. Over emphasis on w_2 can have detrimental effect in case the segments are categorized randomly giving rise to low run. We have experimented with a set of values ranging from 0.4 to 0.8. Best result was obtained for 0.7. However, the performance was also very close for 0.6.

4 Experimental results

Experiments are performed on three different benchmark datasets following both the approaches namely *feature based* and *deep learning based approach*. The model in Fig. 1 shows the position of different emotions in a two dimensional plane. In our work, instead of dealing with so many classes, we have considered four broad categories. These are *happy*, *anger*, *sad* and *tender/neutral*. It may be noted that the four categories correspond to the first, second, third and fourth quadrant of the Russell's model.

4.1 Datasets

Soundtrack [14], and Bi-Modal [37] are the datasets used in our work. In all the datasets, sampling rate for audio clips are 22.05 KHz. The details of the datasets are as follows.

Soundtracks The dataset [14] consists of 360 audio-clips collected from background tracks of movies with duration around 30 seconds. Each clip is annotated with different emotion class like anger, sad, happy and tender. A clip can have multiple tags with confidence value. In our experiment label with maximum confidence is considered for matching. Number of audio clips in anger, happy, sad and tender emotion categories are 156, 58, 68 and 78 respectively.

Bi-Modal This dataset [37] consists of 162 songs. Each song clip is of 30 seconds duration. Both, audio signal and lyrics (textual) data (hence, Bi-Modal) of the songs are available. In our work, we have considered the audio signal part and ignored the lyrics. The clips are annotated with the four quadrants as shown in Fig. 1. Number of songs in quadrant 1, 2, 3 and 4 are 52, 45, 31 and 34 respectively. It may be noted that the quadrants correspond to *happy*, *anger*, *sad* and *tender* respectively.

4.2 Hand crafted feature based approach

We have first worked with hand crafted features to study their performance. In order to compute the low level features, audio signal is divided into number of frames each consisting of n (taken as 512 in our work) samples and there is half overlap between the successive frames. Various time domain features like *short term energy (STE)* and *zero crossing rate (ZCR)* [31] which can reflect the arousal and frequency content respectively. Spectral features [31] like *spectral flatness*, *spectral crest factor*, *Spectral centroid*, *Spectral rolloff* and *Spectral flux* have been considered. Thirteen *linear prediction cepstral coefficients (LPCC)* [47] have been included in the feature set to represent the production model of the vocal tract. First thirteen *Mel Frequency Cepstral Coefficients (MFCC)* [32] have been considered to take care of hearing perception of the listener. All the features (*time domain, spectral, LPCC and MFCC*) are computed over the frames. Finally, mean and standard deviation of the individual features over all the frames in the music clip are taken as the clip level features. Thus, 66-dimensional feature vector is formed to represent a clip. All the Features are extracted from the audio signal using the toolbox MARSYAS [59].

Music clips are represented by the extracted feature set and then supervised approach is followed for classification. The classifier is trained with a training dataset and thereafter the trained model is used for test data. We have used three different types of classifiers in this regard – a large-margin classifier (Support Vector Machine(SVM) [11]), a Decision tree based classifier (Random Forest(RF) [6]) and a perceptron based classifier (Neural Network(NN) [40]). All are implemented using Scikit-learn library [45].

We performed experiments with different combination of time-domain, spectral features, MFCC and LPCC based features. Random Forest is trained with a total number of ten trees in the forest and for splitting a node Gini-index [20] is used as impurity measure. In SVM, Radial Basis Function (RBF) kernel is used and the value of regularization parameter is taken as one. For Neural Networks, LBFGS weight-optimization method along with adaptive learning rate and Sigmoid activation function is used. For every experiment (combination of dataset, feature set and classifier), five fold cross validation is applied and average accuracy is reported. Table 2 summarizes the result. It is very difficult to obtain a feature combination that works best across the classifier and dataset. However, in most of the cases when all the features are combined provides better result. In general, the success of various feature-classifier combination is quite limited. It may be noted that for BiModal and Soundtracks dataset F1-score and classification accuracy are used as performance metric respectively as the same have been used by other researchers working with those datasets.

4.3 Deep learning based approach

As discussed in Section 3, music clip is divided into number of segments of five seconds duration. A song/music has a dominating emotional category. But, it may not remain constant over the whole clip. To address the issue, segments are overlapped heavily (four seconds in our case). Thus, the presence of segments with dominating emotion will be emphasized in comparison to deviated ones. Moreover, the substantial overlap will increase the number of segments that helps the network and makes the post-processing meaningful even for clips of small duration. It may be noted that a clip (*i.e.* all the segments) as a whole is either used for training data or as test data.

The proposed network has a total of 1,203,140 trainable parameters. Training data is split into mini batches with batch size of 64 and training is done by minimizing categorical cross entropy loss [17, 51] between predictions and targets. Adam optimizer [28] is used with a learning rate of 0.001. Dropout technique is implemented where a random fraction

Table 2 Classification performance for different combinations of hand crafted feature sets and classifiers

	Bimodal (F1-score in %)			SoundTracks (Acc. in %)		
Features	SVM	RF	NN	SVM	RF	NN
A	44.78	42.02	46.48	43.61	41.46	41.51
B	47.12	50.77	54.66	50.00	48.06	52.68
A + B	47.71	50.88	56.62	51.38	48.27	51.80
A + B + C	53.23	52.94	62.74	53.61	49.10	54.31
A + B + C + D	54.26	52.54	63.45	53.77	49.71	55.41

Feature Sets: A = time domain features; B = spectral features; C = MFCC; D = LPCC

Table 3 Precision, Recall and F-1 score (in %) for Soundtracks dataset

Class	Precision	Recall	F-1 score
quadrant 1	58.20	71.23	63.61
quadrant 2	54.37	50.68	51.46
quadrant 3	82.25	82.65	82.28
quadrant 4	60.02	42.91	49.32

of neurons are switched off to prevent overfitting of the training set. L2 regularization is also applied to the weights of the fully connected layers. Weight initialization is done using truncated normal initializer. All the codes for training the model were written in Python using Keras [9] library. Experiment is carried out on Nvidia Quadro M5000 GPU with 8 GB of memory.

Class wise precision, recall and F-1 score for soundtracks dataset are shown in Table 3. Table 4 shows the performance for Bi-modal. It is observed that significant confusion arises between sad and tender (quadrant 3 and 4 of 2-D Russell Plane [48]), happy and anger (quadrant 1 and 2) classes. This may be accredited to the fact that both sad and tender classes belong to the low arousal category of 2D Russell plane. Happy and anger belong to the high arousal category. It indicates that proposed model is stronger in discriminating emotions based on arousal and relatively poor for valence.

By observing the experimental outcomes, it is well understood that designing the features to represent the emotion is quite difficult. More difficult is to obtain a consistent set of features that provides optimal result for any classifier and for different datasets. Classification accuracy is also limited for the conventional approach based on hand picked features and classifier. In this context, proposed convolutional neural network improves the performance substantially for the datasets.

Performance of the proposed deep learning based methodology is compared with the work of Saari et al. [49]. They have worked with soundtracks dataset. A set of 66 frame level audio features (extracted with MIRtoolbox) has been considered. Wrapper-selection has been employed and best performance is achieved with 4 randomly selected features. As they have worked with four fold cross validation, for comparison we have also followed the same. Table 5 shows that proposed methodology provides better result. It may be noted that, hand crafted feature based experiment with all the features combined together and neural network as the classifier (as shown in Table 2) provides better accuracy. Proposed deep learning based approach improves the result further.

Malherio et al. [37] have worked with Bi-modal dataset. As we have ignored textual data, performance is compared with audio based work of Malherio et al. [37]. They have used loudness, pitch, timbral, rhythmic, spectral contrast, and Daubechies wavelets coefficient histogram (DWCH) as acoustic domain features and SVM classifier. Table 6 shows the comparative results. As Malhero et al. has followed ten fold cross validation, for comparison

Table 4 Precision, Recall and F-1 score (in %) for Bi-Modal dataset

Class	Precision	Recall	F1-score
quadrant 1	80.46	81.59	80.98
quadrant 2	92.07	74.04	81.79
quadrant 3	72.97	68.52	68.82
quadrant 4	74.20	86.85	79.70

Table 5 Comparison of performance for Soundtracks dataset

Methodology	Accuracy (in %)
k-NN BE of Saari et al. [49]	56.5 ± 2.8
SVM BE of Saari et al. [49]	54.3 ± 1.9
Proposed deep learning based approach	67.71 ± 3.63

Table 6 Comparison of performance for Bi-Modal dataset

Methodology	F1-score (in %)
Malherio et al. [37] (using only audio features)	72.60
Proposed deep learning based approach	77.82 ± 4.06

Table 7 Precision, Recall and F-1 score (in %) for MER_taffc dataset

Class	Precision	Recall	F1-score
quadrant 1	79.98	72.69	76.16
quadrant 2	81.80	81.83	81.82
quadrant 3	95.44	95.44	95.44
quadrant 4	74.99	81.80	78.25

Table 8 Comparison of performance for MER_taffc dataset

Methodology	F1-score (in %)
Panda et al. [44]	76.40 ± 0.04
Proposed deep learning based approach	82.95 ± 1.42

Table 9 $p-$values for statistical two proportion Z test with best reported results

	Panda et al. [44]	Malherio et al. [37]	Saari et al. [49]
Proposed Approach	1.8×10^{-6}	0.061	2.7×10^{-6}

we have also followed the same. It may be noted that the performance of hand crafted feature based experiment of ours (as shown in Table 2) is inferior to the work of Malherio et al. [37]. But, deep learning based approach performs better.

We have also considered the work of Panda et al. [44] for comparison. They have used rhythm, dynamics, melody, harmony, tonal color based acoustic domain features and SVM classifier. They have worked on MER_taffc [44] dataset. It consists of 900 songs. Each song clip is of 30 seconds duration. The clips are annotated with the four quadrants as shown in Fig. 1. Each quadrant has exactly 225 number of song clips. The details of the dataset preparation are reported in [44]. Proposed deep learning based methodology has been applied on the same dataset and detailed result is shown in Table 7. Comparative result is shown in Table 8 and it is observed that proposed methodology provides better result.

We have conducted statistical two proportion Z test [8] to compare the performance of proposed deep learning based approach with other works. The $p-$values are shown in Table 9. Based on the p-values it is concluded that improved performance obtained in case of proposed methodology in comparison to the works of Saari et al. [49] and Panda et al. [44] is statistically significant under 5% α level. With respect to the work of Malherio et al. [37], improvement of proposed work is significant for α level greater than 6.1%. It may be noted that the dataset used in this specific case is the smallest of the three and that may affect our performance. However, in general performance of the proposed methodology is statistically significant.

5 Conclusion

In this work, we have followed deep learning based approach for music emotion recognition and experiment is carried out on three benchmark datasets. Experiment has also been done with handcrafted features. Different time domain and spectral features are chosen based on the past efforts of the researchers. LPCC and MFCC based features are also included as those correspond to the aspects of vocal production and human perception respectively. Although the combined feature set provides a moderate result for different benchmark datasets, but the performance varies for different classifier. To avoid the difficulty of designing proper features, deep learning based approach is considered. Proposed convolutional neural network (CNN) is the modified version of VGGNet with comparatively less number of layers. It works with the audio segment of very small duration, even of five seconds to recognize the emotion. A novel post-processing technique has been proposed that works on the class label of the segments to determine the clip level emotional category. Experimental results show that proposed network improves the recognition accuracy considerably. Comparison of performance with three different systems reflect the superiority of the proposed methodology and it has been substantiated by Z test. It may be noted that with larger dataset performance may improve further. In future, efforts may be directed to improve the performance in case of low arousal. In order to emphasize the time series nature of audio data, a combined CNN-LSTM network may be considered for music emotion recognition in future. Also it will be worth to explore transfer learning.

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no conflict of interest.

References

1. Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 22(10):1533–1545
2. Albornoz E, Sänchez-Gutiërrez M, Martinez F, Rufiner H, Goddard J (2014) Spoken emotion recognition using deep learning. In: Iberoamerican congress on pattern recognition, pp 104–111
3. Badshah AM, Rahim N, Ullah N, Ahmad J, Muhammad K, Lee MY, Kwon S, Baik SW (2019) Deep features-based speech emotion recognition for smart affective services. Multimed Tools Appl 78(5):5571–5589
4. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv:1409.0473
5. Bigand E, Vieillard S, Madurell F, Marozeau J, Dacquet A (2005) Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. Cogn Emot 19(8):1113–1139
6. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
7. Cabrera D et al (1999) Psysound: a computer program for psychoacoustical analysis. In: Australian acoustical society conference, vol 24, pp 47–54
8. Casella G, Berger RL (2002) Statistical inference, vol 2. CA, Duxbury Pacific Grove
9. Chollet F (2015) Keras. https://github.com/fchollet/keras
10. Coutinho E, Trigeorgis G, Zafeiriou S, Schuller BW (2015) Automatically estimating emotion in music with deep long-short term memory recurrent neural networks. In: Mediaeval
11. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
12. Cummins N, Amiriparian S, Hagerer G, Batliner A, Steidl S, Schuller BW (2017) An image-based deep spectrum feature representation for the recognition of emotional speech. In: International conference on multimedia, pp 478–484
13. Droit-Volet S, Ramos D, Bueno L, Bigand E (2013) music, emotion, and time perception: the influence of subjective emotional valence and arousal? Front Psychol 4:417
14. Eerola T, Vuoskoski JK (2011) A comparison of the discrete and dimensional models of emotion in music. Psychol Music 39(1):18–49
15. Gabrielsson A, Lindström E (2001) The influence of musical structure on emotional expression. Oxford University Press, Oxford
16. Gharavian D, Bejani M, Sheikhan M (2017) Audio-visual emotion recognition using fcbf feature selection method and particle swarm optimization for fuzzy artmap neural networks. Multimed Tools Appl 76(2):2331–2352
17. Goldberg Y (2017) Neural network methods for natural language processing. Synth Lect Hum Lang Technol 10(1):1–309
18. Han BJ, Rho S, Jun S, Hwang E (2010) Music emotion classification and context-based music recommendation. Multimed Tools Appl 47(3):433–460
19. Hassan A, Damper R, Niranjan M (2013) On acoustic emotion recognition: compensating for covariate shift. IEEE Trans Audio Speech Lang Process 21(7):1458–1468
20. Hastie T, Tibshirani R, Friedman J (2008) The Elements of Statistical Learning, 2 edn., chap. Random Forests. Springer, pp 592
21. Huang Z, Dong M, Mao Q, Zhan Y (2014) Speech emotion recognition using cnn. In: ACM International conference on multimedia, pp 801–804
22. Huang Z, Xue W, Mao Q, Zhan Y (2017) Unsupervised domain adaptation for speech emotion recognition using pcanet. Multimed Tools Appl 76(5):6785–6799
23. Huq A, Bello JP, Rowe R (2010) Automated music emotion recognition: a systematic evaluation. J Music Res 39(3):227–244
24. Jun Han B, Rho S, Dannenberg RB, Hwang E (2009) Smers: Music emotion recognition using support vector regression. In: International society for music information retrieval, pp 651–656
25. Kahou SE, Pal C, Bouthillier X, Froumenty P, Gülċehre Ċ, Memisevic R, Vincent P, Courville A, Bengio Y, Ferrari RC et al (2013) Combining modality specific deep neural networks for emotion recognition in video. In: International conference on multimodal interaction, pp 543–550
26. Kim Y, Schmidt EM, Migneco R, Morton BG, Richardson P, Scott J, Speck JA, Turnbull D (2010) Music emotion recognition: a state of the art review. In: International society for music information retrieval, pp 255–266
27. Kim Y, Lee H, Provost EM (2013) Deep learning for robust feature generation in audiovisual emotion recognition. In: International conference on acoustics, speech and signal processing, pp 3687–3691
28. Kingma D, Ba J (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
29. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105

30. Krumhansl CL (2002) Music: a link between cognition and emotion. Curr Direct Psychol Sci 11(2):45–50
31. Lerch A (2012) An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics, 1st edn. Wiley-IEEE Press, New York
32. Logan B (2000) Mel frequency cepstral coefficients for music modeling. In: International society for music information retrieval, pp 138–147
33. Lu L, Liu D, Zhang H (2006) Automatic mood detection and tracking of music audio signals. IEEE Trans Audio Speech Lang Process 14(1):5–18
34. Lu Q, Chen X, Yang D, Wang J (2010) Boosting for multi-modal music emotion. In: International society for music information and retrieval conference, pp 105–105
35. Lin YC, Yang YH, Chen HH (2011) Exploiting online music tags for music emotion classification. ACM Trans Multimed Comput Commun Appl 7S(1):26:1–26:16
36. Liu X, Chen Q, Wu X, Liu Y, Liu Y (2017) Cnn based music emotion classification. arXiv:1704.05665
37. Malheiro R, Panda R, Gomes P, Paiva R (2016) Bi-modal music emotion recognition: Novel lyrical features and dataset. In: International workshop on music and machine learning
38. Mao Q, Dong M, Huang Z, Zhan Y (2014) Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans Multimed 16(8):2203–2213
39. Markov K, Iwata M, Matsui T (2013) Music emotion recognition using gaussian processes. In: Mediaeval
40. Minsky M, Papert S (1969) Perceptrons. MIT Press, Cambridge
41. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: International conference on machine learning, pp 807–814
42. Nordström H, Laukka P (2019) The time course of emotion recognition in speech and music. J Acoust Soc Amer 145(5):3058–3074
43. Ooi CS, Seng KP, Ang LM, Chew LW (2014) A new approach of audio emotion recognition. Expert Syst Appl 41(13):5858–5869
44. Panda R, Malheiro RM, Paiva RP (2018) Novel audio features for music emotion recognition. IEEE Transactions on Affective Computing
45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830
46. Rabiner LR, Schafer RW (2007) Introduction to digital speech processing. Found Trends Signal Process 1(1):1–194
47. Rao KS, Reddy VR, Maity S (2015) Language identification using spectral and prosodic features. Springer, Berlin
48. Russell J (1980) A circumplex model of affect. J Person Soc Psychol 39(6):1161–1178
49. Saari P, Eerola T, Lartillot O (2011) Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. IEEE Trans Audio Speech Lang Process 19(6):1802–1812
50. Schmidt EM, Kim Y (2011) Learning emotion-based acoustic features with deep belief networks. In: IEEE Workshop on applications of signal processing to audio and acoustics, pp 65–68
51. Sadowski P (2016) Notes on backpropagation. homepage: https://www.ics.uci.edu/˜pjsadows/notes.pdf (online)
52. Sanyal S, Banerjee A, Sengupta R, Ghosh D (2016) Chaotic brain, musical mind-a non-linear neurocognitive physics based study. Journal of Neurology and Neuroscience
53. Seo YS, Huh JH (2019) Automatic emotion-based music classification for supporting intelligent iot applications. Electronics 8(2):164
54. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
55. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Computer vision and pattern recognition, pp 1–9
56. Thayer RE (1990) The biopsychology of mood and arousal. Oxford University Press, Oxford
57. Thammasan N, Fukui K, Numao M (2016) Application of deep belief networks in eeg-based dynamic music-emotion recognition. In: International joint conference on neural networks, pp 881–888
58. Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, Zafeiriou S (2016) Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: International conference on acoustics, speech and signal processing, pp 5200–5204
59. Tzanetakis G, Cook P (1999) Marsyas: a framework for audio analysis. Organised Sound 4(3):169–175
60. Yang YH, Lin YC, Su YF, Chen HH (2007) Music emotion classification: a regression approach. In: International conference on multimedia and expo, pp 208–211
61. Yang YH, Lin YC, Su YF, Chen HH (2008) A regression approach to music emotion recognition. IEEE Trans Audio Speech Lang Process 16(2):448–457

62. Yang YH, Chen HH (2012) Machine recognition of music emotion: a review. ACM Trans Intell Syst Technol 3(3):40:1–40:30
63. Yang X, Dong Y, Li J (2018) Review of data features-based music emotion recognition methods. Multimedi Syst 24(4):365–389
64. Yeh CH, Tseng WY, Chen CY, Lin YD, Tsai YR, Bi HI, Lin YC, Lin HY (2014) Popular music representation: chorus detection & emotion recognition. Multimed Tools Appl 73(3):2103–2128
65. Zhang F, Meng H, Li M (2016) Emotion extraction and recognition from music. In: International conference on natural computation, fuzzy systems and knowledge discovery, pp 1728–1733
66. Zheng WL, Lu BL (2015) Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. IEEE Trans Auton Ment Dev 7(3):162–175
67. Zeng N, Zhang H, Song B, Liu W, Li Y, Dobaie AM (2018) Facial expression recognition via learning deep sparse autoencoders. Neurocomputing 273:643–649
68. Zao L, Cavalcante D, Coelho R (2014) Time-frequency feature and ams-gmm mask for acoustic emotion classification. IEEE Signal Process Lett 21(5):620–624

Rajib Sarkar obtained his Masters in computer application degree from Jadavpur University, India in 2013. Currently, he is working as Assistant Professor in Computer Science department of Derozio Memorial College, West Bengal State University, India. His research interest includes audio signal processing, pattern recognition and deep learning.

Sombuddha Choudhury did his B.E. in Computer Science and Engineering from Jadavpur University, India in 2017. His interest is in the domain of machine learning and deep learning.

Saikat Dutta did his bachelor degree in computer science and engineering from Jadavpur University, India in 2018. Currently, he is pursuing graduate program at Indian Institute of Technology, Chennai. His areas of interest are machine learning and pattern recognition.

Aneek Roy completed his bachelor program from Computer Science and Engineering department of Jadavpur University, India in 2018. Soft computing and computer vision are his areas of interest.

Sanjoy Kumar Saha obtained his Bachelor and Master in engineering degree in Electronics and Tele-Communication from Jadavpur University, India in 1990 and 1992 respectively. He obtained PhD from Bengal Engineering and Science University, Shibpur (now IIEST, Shibpur), India in 2006. Currently working as Professor in Computer Science and Engineering Department of Jadavpur University, India. His research area includes signal processing, pattern recognition and information retrieval.

Affiliations

Rajib Sarkar[1,2] · **Sombuddha Choudhury**[1] · **Saikat Dutta**[1] · **Aneek Roy**[1] · **Sanjoy Kumar Saha**[1]

Sombuddha Choudhury
sombuddha.choudhury@gmail.com

Saikat Dutta
saikat.dutta779@gmail.com

Aneek Roy
aneek.roy5@gmail.com

Sanjoy Kumar Saha
sks_ju@yahoo.in

1 Department of Computer Science and Engg., Jadavpur University, Kolkata, 700032, India

2 Computer Science Department, Derozio Memorial College, Kolkata, 700136, India