ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(14), DOI: 10.17485/ijst/2019/v12i14/142669, April 2019*

Performance Analysis of Different Kernels of Support Vector Machine and Self-organizing Maps for Classifying Musical and Non-musical Personal Videos

Pratap Sanap and Shaila D. Apte*

Electronics Department, Rajarshi Shahu College of Engineering, Pune – 411033, Maharashtra, India;
pratapsanap@rediffmail.com,sdapte@rediffmail.com

Abstract

Objectives: This study aims at detecting the video with or without music. The present article will support the contextualization process of the videos. **Method**: Videos with musical piece are detected using temporal and spectral features. Its focus is also on the comparative study of two classifiers i.e. Self-organizing maps and Support Vector Machines. **Findings**: Music detection in audio track of video was mainly carried out for professional videos. The proposed work focuses on personalised videos which are very challenging due to limitations in the environmental conditions as well as procedures of recording. No work is found on the use of musical portion detection especially in the personal video recordings. The successful detection of music in audio track of videos is used for prioritization in the contextualization process. Accuracy of detecting musical video is 85% with combination of temporal and spectral features. The limitation in the accuracy is due to non-professional recordings of the personal videos. **Application**: With recent advancement in hardware technology, the high-end HD cameras are available in every handheld device. The video capture is significantly increased especially in the personal videos. The use of images and videos are becoming essential part of everyone's routine from students for studies to elderly individuals for safety and entertainment. It is forecasted that the 80% of internet traffic will be occupied by the videos. Hence video analysis has become very crucial to detect unwanted video portions to avoid the transportation of these videos. It is a vital task to design a model to detect the videos having musical portion for prioritization for video editing and contextualization process.

Keywords: Audio-visual Stream Analysis, Detection of Music Video. Multimedia Signal Processing

1. Introduction

Video capture has increased across diverse domains such as entertainment, security, health-care and home automation. Presently a lot of focus is on automation of video analysis. The proposed methodology focuses on classifying videos having music in it. Such estimations can be very useful to navigate via the videos and extract meaningful information. Furthermore, the proposed method can be used in estimating the boundaries of musical portion within audio stream. The boundary detection accuracy further deteriorates if the video segmentation is carried out on the videos recorded at personal or family function. This article addresses issue of detection of the musical portion of personal videos using supervised

*Author for correspondence

approach by combining features calculated using Mel-frequency Cepstral Coefficients, Spectral and Temporal parameters of the video signal.

Primarily, all the audio analysis systems are defined on the principle of audio finger printing. The methodologies defined so far analyse the audio signal and pre-process it for feature extraction. It is then used for identification of musical portion. Database for training the present system is generated by extracting the features from the training samples.

A reference database contains audio feature vector to train the model. Then an unknown test sample is identified by comparing its features with those of the reference database. The main challenge of such systems is to design a robust features vector and to propose a matching method that can accurately classify irrespective the size of the reference database. Videos recorded with personal devices have following undesired characteristics that require special attention while classifying musical videos.

- Instrumental breakdowns can be introduced on purpose by the video producer scripts.
- Music pieces can be played successively without any pauses between them.
- Outdoor Recordings, atmosphere noise contribution.
- Incomplete recordings.

The audio feature vector is represented with pitch estimation, chroma representation and MFCC parameters. They are mainly categorised in temporal and spectral features.

The paper is organized as follows. Section 2 deals with the literature review. Section 3 describes the methodology of detecting musical portion of the video. Section 4 explains the process of features extraction. Section 5 explains classification techniques. Section 6 gives details of the database. Section 7 describes result and analysis and Section 8 pronounces the discussions and conclusion.

2. Literature Survey

Considering the importance of classifying musical videos, several approaches are developed. In this section, different methodologies and processes are reviewed. The following literature review analyses the work on musical video estimation. These are referred in the present work to elaborate the enhanced algorithm those are developed.

In[1] carried out the research to detect sound abnormality even after the recording has camera blind spot. In this paper, the author proposed the use of embedded system to upgrade current CCTV system. Time and frequency domain features are used by the authors to detect abnormal sound such as human screaming and glass breaking. Decision tree is used for classification. The method of abnormal sound detection indicates average of 88% accuracy.

In[2] proposed a framework to detect sports video for personal video recorders. Creating small but appropriate classes assisted the authors to optimize the computational complexity of GMM. The analysis shows that summarization algorithms also work for music video contents.

In[3] proposed a novel method of video summarization. The proposed schematic uses the moving object details by considering region of interest of moving objects extracted from background model and object motion in subsequent frames. While designing the methodology, authors have also considered the memory requirements for analysing the videos.

In[4] explored the property of having significant temporal and emotion information in the physical layer of the musical signal and is utilised for retrieval. The retrieval methodology is based on time changing musical emotions. 3D musical emotion model-Resonance-Arousal-Valence (RAV) is used. To model music and emotion over the time, Multiple Dynamic Textures (MDT) model is pro-

posed by the author. Further, Kalmal filtering is used to calculate the model parameters.

In[5] proposed music-driven summarization system for home videos based on contents of the videos. Many audio and video features are used for analysing and synchronizing input audios and videos. The synchronization is directed by matching rhythm of the video with that of the audio. Four profiles for synchronizing video with audio are proposed, which provide significant flexibilities in conducting the synchronization process.

In[6] proposed video summarization and exploration methodology to extract features of each frame of the video. The priority is given to the users to interactively explore those features of each frame. These visual effects are mapped to represent shots of the video. It uses a self-organizing map to detect and remove the redundant frames. Using clustering algorithms, all the shots are connected to the network structure. The proposed method defines expert system to find relation between nodes to derive the similarity index.

In[7] proposed ASR research work is to detect the input signal with near 100% accuracy and its operational steps on real-time data feed. Author also provided summary about recent methods on ASR engines, explaining their advantages and disadvantages and derived an efficiency of the proposed technique using computational analysis.

In[8] proposed an expert system with knowledge-based analysis to identify different audio segments in videos. Audio content analysis is the primary step for event detection and such related applications. Author proposed to extract Mel-Frequency Cepstral coefficient (MFCCs) features. Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) are used for classification purpose. The results are not so encouraging due to environmental conditions. Hence it is proposed to use Support Vector Machines (SVMs), Neural Networks and Random Forest for classification and segmentation of video.

3. Methodology

Literature review indicates that most of the work is done on audio fingerprint for independent musical note detection and audio signal analysis for user verification, etc. However, there is an opportunity to develop a framework which can detect musical piece of the videos. Videos have an inherent property of smooth transition between adjacent frames. The audio signal from the video can be analysed in prioritization of the video segmentation process.

The proposed innovative algorithm for detection of musical audio frames in video is described in this section. Video file header has information about the specifications and storage format of audio and image format. The developed application reads the header information and extracts the audio chunks from the video file. The audio chunks are composed sequentially to create audio metadata from the video signal. The entire audio signal is windowed with window size of 30 ms. The window size of 30 ms is selected because over this period, the musical audio parameters can be assumed to be constant and linear time invariant model can be used for analysis. The audio frame is analysed for rhythmic parameters and the metadata is generated. The classification is expected to detect the videos with and without musical portion. Considering the binary (Musical and non-musical) nature of the data, classification is carried with Support Vector Machine (SVM) and Self-Organizing Maps (SOM) efficiently. At the end, the article also provides comparative analysis of the two classifiers mentioned above. The generated metadata is further utilized for indexing the video for editing and contextualization.

Since this is supervised classification, the methodology is divided in two sections i.e. training and testing. In total 500 samples of videos are used. The labelled data of 300 video samples is used for training and data of 200 video samples is used for validation. The database details are provided in the section 6 below. The method-

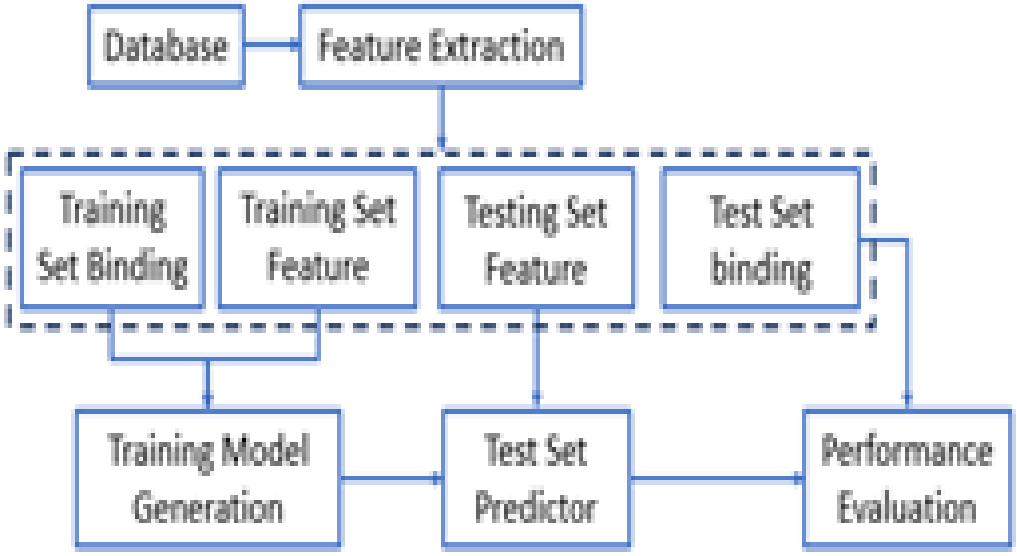

Figure 1. Training and validation methodology.

ology adopted for the present work is demonstrated in the schematic shown in Figure 1. It demonstrates overall methodology of supervised machine learning approach for detection of musical videos. In training phase, the labelled database of 150 samples of non-musical videos and 150 samples of musical videos is used. All the samples of respective labels are given to feature extraction module. The feature details and extraction methodology are described in the following sections. The extracted featured with respective labels are given to the training module. SVM and SOM models are generated for classification. The characteristics and configuration of classifies are described below.

Music or Non-Music Video detection: Figure 2 explains the steps used for detecting videos with music. Temporal, Spectral and MFCC features are used with SVM and SOM classifiers for classifying musical or Non-musical videos. The audio signal is broken into windows (frames) of 30 ms. These windows are non-overlapping and rectangular. For each window 3 temporal, 3 spectral features and 13 MFCC features are extracted. The Temporal and Spectral features include Entropy of the

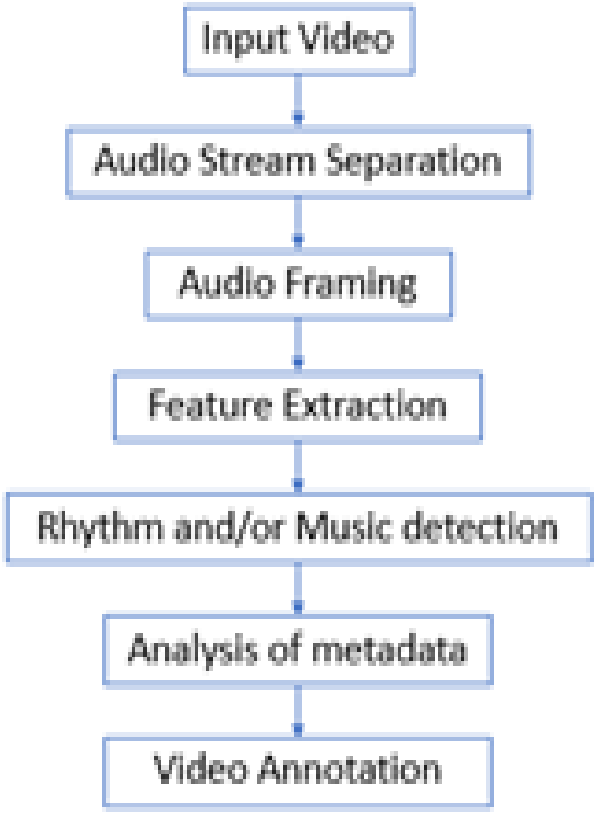

Figure 2. Musical video detector methodology.

window, Energy on the window and Zero Crossing Rate of the window. Along with the temporal features, spectral features such as Spectral Roll-off, Spectral Centroid and Spectral Flux are calculated for each window. Above mentioned step defines 6 feature sequences for each window of the audio signal. In the sequel, for each of the 6 feature sequences, statistical parameters are evaluated. This step leads to 6 single statistic values (one for each feature selected). Those 6 values are the final feature values that characterize the input audio signal.

4. Feature Extraction

The entire audio from video is extracted and given as an input to feature extraction module. For calculating temporal and Spectral features, the audio file is divided in the short time window of 30 ms and feature vector of each such window is calculated. The feature extraction module is used to calculate 6 features for each window. The size of the resulted feature vector is N*M, where N is the number of windows and M is the number of features. Standard deviation of the feature vector is calculated to form one feature vector for each video file. Similarly, 13 MFCC features are calculated for each video file. The final feature vector is of size K*19 is generated, where K is the number of video files with 19 features for each file is generated. The generated feature vector is given as an input to the training module.

4.1 Temporal Features

Entropy: Entropy gives the amount of abrupt changes in the energy of the input signal. To compute entropy, divide the signal in subframes (j), compute its energy and divide it by the total energy. Equation 1, 2 and 3 demonstrates the method to calculate the entropy.

$$e_j = \frac{EsubFramej}{EshortFramei} \dots\dots\dots\dots\dots\dots\dots (1)$$

Where

$$EshortFrame_i = \sum_{k=1}^{K} EsubFrame_k \dots\dots\dots (2)$$

At a final step, the entropy, H(i), of the sequence is computed according to the equation:

$$H(i) = \sum_{k=1}^{K} e_k . log2(e_k) \dots\dots\dots\dots\dots (3)$$

Short time energy: Short term energy is computed as taking square of sum of the signal values and then it is normalized by the window length.

$$STE = \frac{1}{N}\sum_{n=1}^{N} (x(n))^2 \dots\dots\dots (4)$$

Zero Crossing Rate: ZCR is the measure of calculating frequency of the signal in time domain. It is directly related to the number of times the waveform crosses the dc line. Variance of the zero-crossing rate is calculated to strengthen extracted features. ZCR calculation is defined in the equation below.

$$ZCR = \frac{1}{2N}(\sum_{n=1}^{N} |S(x(n)) - S(x(n-1))|) \dots (5)$$

Where x is the time-domain signal, S is the signum function and N is the size of frame. The signum function implementation can be defined as:

$$S(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases} \dots\dots\dots\dots\dots\dots (6)$$

4.2 Spectral Features

Spectral roll off: Spectral roll-off is defined as the percentile of the power spectrum. Generally, 85 or 95 of percentile is used for the music detection. Spectral roll

off measures the spectral shape and it gives higher values for high frequencies. This helps to detect the Musical videos.

$$SRP = f(n) \text{ where } f(n) = \frac{f_s}{K} N \ldots\ldots\ldots (7)$$

Where N is the largest bin that fulfils Equation.

$$\sum_{k=0}^{N} |X(k)|^2 \leq TH. \sum_{k=0}^{K-1} |X(k)|^2 \ldots\ldots\ldots\ldots\ldots (8)$$

Where X (k) are the magnitude components, k frequency index and f (K) spectral roll-off point with 100 ∗ TH % of the energy. TH is a threshold between 0 and 1. Experimentally for detecting the musical video the threshold is 0.85 and 0.9

Spectral flux: Spectral flux is a measure of how rapidly power spectrum of the audio signal is altering; it is computed by comparing the power spectrum for current window against the power spectrum of the previous window. Since Non-musical video has changing spectrum within window, Spectral flux will be differentiating these variations to classify musical and non-musical videos.

$$F(x,t;v) = \oint_{\Omega} I(x,t;n,v) n \, dw(n) \ldots\ldots\ldots (9)$$

Spectral centroid: It defines the "center of mass" of the frequency spectrum. It defines the brightness of the audio signal. The Spectral centroid is expressed using Equation number 10.

$$Centriod = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \ldots\ldots\ldots\ldots (10)$$

Where x(n) represents weighted magnitude, of bin number n and f(n) represents the center frequency of the respective bin. Spectral centroid is a very important feature for detecting mixed sounds. Whish can help in classifying the musical and non-musical videos. Spectral Centroid also models the sound sharpness which is very important in Musical detector.

MFCC**:** MFCCs' are the most important parameters for audio processing. The property of MFCC to work on noisy and mixed signal is beneficial for classifying musical and non-musical videos. The methodology of calculating MFCC feature is defined below.

Pre-emphasis**:** Pre-emphasized refers to emphasizing the signal spectra. The DC offset is removed, a low order FIR filter is applied to input signal to flatten the spectrum. This makes it less susceptible to find precision effects later in the audio processing.

$$H(z) = 1 - az^{-1} \; 0.9 < a < 1 \ldots\ldots\ldots\ldots (11)$$

The value of 'a' is typically 0.95. The pre-emphasis filter boosts the audio signal by 20 dB/decade.

Framing**:** The framing refers to decomposing a signal in to frames of 30 ms duration. The spectral analysis is carried out on each frame. This allows us to find spectral variation along the signal.

Windowing**:** Every frame of 30 ms is passed via a smooth window function to reduce the effect of spectral leakage. Equation below refer the method of applying window on the signal

$$y(n) = x(n)\, w(n) \ldots\ldots\ldots\ldots (12)$$

Hamming window is commonly used. For a constant transition width of 8π/N, Hamming window gives smallest side lobe amplitude. Thus it avoids the spectral leakages. Equation for hamming window is given below:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi}{N-1}\right) \ldots\ldots\ldots\ldots (13)$$

Spectral Estimation**:** The spectral estimation involves following steps. Step 1- computes DFT of each windowed

frame. Step 2 extracts the magnitude of the spectral coefficients.

The frequency is warped according to the characteristics of human ear defined by the Mel filters. These Mel filters are defined on the overlapped critical bands. The Mel frequency is computed as:

$$f_M = 2595 * \log(1 + \frac{f}{7}) \ldots\ldots\ldots\ldots (14)$$

Where f_M is the Mel frequency corresponding to the linear frequency f. The filter bank energy is obtained after Mel filtering

$$E_i^X = \sum_{k=1}^{N} |X(k)|^2 \Psi i\,(k) \ldots\ldots\ldots\ldots\ldots (15)$$

Where |X(k)| is the amplitude spectrum, k is the frequency index Ψ_i, are the i^{th} Mel band pass filter, $1 \leq i \leq M$, and M is number of Mel-scaled triangular band-pass filters. E_i^X is the filter bank energy.

Natural logarithm: Natural logarithm of the magnitude spectrum is evaluated.

Discrete Cosine Transform: The DCT of the log of magnitude spectrum is computed which transforms the signal in cepstral domain. DCT can be defined as:

$$C_i^x = \sum_{i=1}^{M} lpg(E_i^x)C \mid i\frac{(2i-1)\pi}{2M} \mid \ldots\ldots\ldots (16)$$

The first 20 cepstral coefficients are used. This contains most of the information of the audio signal.

5. Classification Techniques

SVM and SOM are the two classifiers used in this article. By nature, SVM is a binary classifier and the proposed methodology is expected to classify between two classes i.e. musical and non-musical video files. SOM is a kind of Artificial Neural Network (ANN) and is an unsupervised learning technique. The objective of the article is to evaluate the performance of supervised and unsupervised classification methods. Hence these two classifiers are used.

5.1 Support Vector Machines (SVM)

The objective of the Support Vector Machine algorithm is to find a hyperplane in an N-dimensional space (N - the number of features) that distinctly classifies the data points. Being binary classifies in nature SVM is very effective in this article. Proposed method analyses the performance of 5 different SVM kernels (Linear, Quadratic, Polynomial, Gaussian Radial Basis, Multilayer Perceptron kernel) with Spectral and Temporal Features and MFCC Features. The SVM can be observed as a kernel apparatus. As a result, it can alter the nature of SVM functions by using a diverse kernel. Kernel methods return the inner product between two points in an appropriate feature space. Hence it really works well on the concept of resemblance, with considerable computational cost even in multi-dimensional parameters.

5.2 SVM Kernel Functions

5.2.1 Linear

Linear kernels are very useful when working with large data vectors. The linear spline separates two classes appropriately may not be useful for complex datasets having non-linear separations. Linear kernel training is much faster than other SMV kernels.

$$f(x) = B(0) + sum(ai * (x,xi)) \ldots\ldots\ldots (17)$$

The above linear kernel equation is calculating the inner products of input vector (x) with all support vectors in training data. Coefficients B0 and ai (for each input) are estimated from training data.

5.2.2 Quadratic

Are the nonlinear kernels and can be considered as an extension of linear kernels to separate two classes.

5.2.3 Polynomial

It is popular very in signal processing applications. The Equation is given below:

$$k(X_i, X_j) = (X_i * X_j + 1)^d \quad \ldots\ldots\ldots (18)$$

Where d is the degree of the polynomial.

5.2.4 Gaussian Radial Basis

It is a general-purpose kernel; used when there is no prior knowledge about the data. The Equation is given below:

$$k(X_i, X_j) = exp(-\gamma \, ||X_i - X_j||)^2 \quad \ldots\ldots\ldots (19)$$

Where

$$\gamma = {}^{1}/_{2\sigma^2}$$

5.2.5 Multilayer Perceptron

Multilayer SVM architecture contains multiple support vector classifiers in the output layer. To deal with multiple classes, the use of classification dataset for each classifier $\{(x_1, y_1^c), \ldots, (x_l, y_l^c)\}$ where xi are input vectors and $y_1^c \in \{-1, 1\}$ are the target outputs that denote if the example xi belongs to class c or not. All classifiers M_c share the same hidden-layer of regression SVMs. M_c determines its output on an example x as follows:

$$gc(f(x|\theta)) = \sum_{i-1}^{l} y_i^c \alpha_i^c K_2\left(f(x_i|\,\theta), f(x|\,\theta)\right) b_c \ldots (20)$$

5.2.6 Self-organizing Maps (SOM)

SOM is the interesting class of unsupervised learning system that is based on competitive learning, Here, the output neurons compete amongst themselves to be activated. This will activate one at a time. This activated neuron is called a winner - neuron. It takes simply the winning neuron. Such competition can be better implemented by having lateral inhibition connections using negative feedback paths between the neurons. The forces neurons to organize themselves. Hence, such a network is called a Self-organizing Map (SOM).

The SOM will transform an incoming signal pattern of an arbitrary dimension into a one or two-dimensional discrete map. It performs this transformation adaptively in a topological order. SOM is usually set by placing neurons at the nodes of a one or two-dimensional lattice structure. Higher dimensional maps are not so common. The neurons are selectively tuned to various input patterns (stimuli). They are tuned to classes of input patterns during the competitive learning. The locations of the neurons after tuning (i.e. the winning neurons) become ordered. A meaningful coordinate system for the input features is created on the lattice. The SOM is thus said to form the required topographic map of the input patterns.

6. Database Summary

As the present work is focusing on personalized videos, there is security and privacy issue hence standard database of personalized videos is not available. The database is created in house after taking subjective opinion of experts. 5 experts participated in the subjective assessment test. The experts are undergraduate and graduate students with understanding of multimedia contents of personal videos. Experts viewed and formulated the subjective matrices on the video quality.

The goal of the database is to investigate the observers' subjective opinion on the musical portion in the videos, provide benchmark for objective assessment models and

facilitate future contextualization applications. The video database used for the present work is downloaded from social media and is also captured using personal devices. The database consists of 500 video files recorded for minimum of 2 minutes recording. The videos used for analysis are of 30 frames/second with a spatial resolution at 848 × 480 with audio sampling frequency of 44.1k, Bit-Rate of 203 KBPS and Stereo Channels. The sampling frequency for audio is kept 44.1 kHz to grab high frequency details in the musical portion. Appropriate care is taken to extract audio signal irrespective of the encoding format.

7. Result and Analysis

This section describes the results obtained for labelled 200 videos database for detecting musical or non-musical videos. The result section is divided in two Sections. The accuracy is calculated by taking the ratio of correctly classified data samples with total number of testing samples of each class. The equation for accuracy calculation is mentioned below.

$$accuracy = \frac{Positives\ Samples}{Total\ testing\ samples} * 100 \ldots (21)$$

Table 1. Comparative analysis of SVM kernels

Features	SVM Kernels	Accuracy (%)
Spectral and Temporal	Linear	80
	Quadratic	75
	Polynomial	75
	Gaussian Radial Basis	81
	Multilayer Perceptron	74
MFCC	Linear	74
	Quadratic	75
	Polynomial	75
	Gaussian Radial Basis	81
	Multilayer Perceptron	78
Combined	Linear	80
	Quadratic	79
	Polynomial	77
	Gaussian Radial Basis	83
	Multilayer Perceptron	78

Table2. Results obtained by SOM

Features	Classifier	Accuracy (%)
Temporal and Spectral	SOM	83
MFCC		81
Combined		85

Table 3. Precision and recall values calculated for musical and non-musical video database

Classifier	TP	FP	FN	Recall	Precision
SOM	170	20	10	0.94	0.89
SVM	162	22	16	0.91	0.88

Section 1: Table 1 displays the comparative analysis of different SVM kernels for different features. Also displays the result obtained by combining the features. Spectral and temporal features give better results as compared to MFCC for linear kernel. The results for MRCC are also comparable. The music and non-music are classified using frequency contents above 3 KHz as compared to speech and other audio contents. The rhythm is also captured by these features.

Section 2: Table 2 displays the comparative analysis of SOM for different features. Also, the table below displays the result obtained by combining the features.

Results obtained after combining the features are little promising than the results obtained for individual analysis.

Section 3: Table 3 displays the Recall and Precision values are calculated with Temporal, Spectral and MFCC features. TP is denoted are True Positives, FP is denoted as false positives and FN is denoted as False Negatives. Precision and recall are calculated using the equations mentioned below.

$$Precision = \frac{TP}{TP + FP} \quad \ldots\ldots\ldots\ldots\ldots (22)$$

$$Recall = \frac{TP}{TP + FN} \quad \ldots\ldots\ldots\ldots\ldots (23)$$

<Insert Tables 1 to 3 here>

8. Discussions and Conclusion

We have presented an effective, robust system for analysing and classifying musical and non-musical videos. Tables of result show that the percentage of musical and non-musical video detection values is greater than 80%. Radial Basis Functions (RBFs) are set of functions which

have values from a specific distance from a seed point. Gaussian Kernels also have diagonal covariance matrix with constant variance, which is very useful for the proposed algorithm. Significant care has taken to avoid over fitting while using Gaussian RBF Kernels. Hence the algorithms provided 81% accuracy using SMV's Gaussian Radial Basis with the features combined. The algorithm provided 85% of accuracy using SOM with combined (temporal and Spectral) features. The article provides significant knowledge that can be used for indexing and prioritising the video contextualization process.

9. References

1. Teng TT, Sze LT, Yeng OL. Abnormal sound analytical surveillance system using microcontroller. IEEE 12th International Colloquium on Signal Processing and its Applications; 2016. p. 162–6. https://doi.org/10.1109/CSPA.2016.7515824.
2. Otsuka I, Radhakrishnan R, Siracusa M, Divakaran A, Mishima H. An enhanced video summarization system using audio features for a personal video recorder. IEEE Transactions on Consumer Electronics. 2006; 52(1):168–72. https://doi.org/10.1109/TCE.2006.1605043.
3. Salehin MM, Paul M. An efficient method for video summarization using moving object information. 18th International Conference on Computer and Information Technology; 2015. p. 237–42. https://doi.org/10.1109/ICCITechn.2015.7488075.
4. Deng JJ, Leung CHC. Dynamic time warping for music retrieval using time series modeling of musical emotions. IEEE Transactions on Affective Computing. 2015; 6(2):137–51. https://doi.org/10.1109/TAFFC.2015.2404352.
5. Huang CH, Wu CH, Kuo JK, Wu JL. A musical-driven video summarization system using content-aware mechanisms. IEEE International Symposium on Circuits and Systems. 2005; 3:2711–4..
6. Qian Y, Kyan M. Interactive user oriented visual attention-based video summarization and exploration framework. IEEE 27th Canadian Conference on Electrical and Computer Engineering; 2014. p. 1–5. https://doi.org/10.1109/CCECE.2014.6901095.
7. Sharma U, Maheshkar S, Mishra AN. Study of robust feature extraction techniques for speech recognition system. International Conference on Futuristic Trends on Computational Analysis and Knowledge Management; 2015. p. 654–8. https://doi.org/10.1109/ABLAZE.2015.7154944.
8. Raghuram MA, Chavan NR, Koolagudi SG, Ramteke PB. Efficient audio segmentation in soccer videos. IEEE Canadian Conference on Electrical and Computer Engineering; 2016. https://doi.org/10.1109/CCECE.2016.7726616.