International Journal Of Computer Science And Applications Vol. 1, No. 2, August 2008
ISSN 0974-1003

Audio Based Movies Characterization Using Neural Network

Sanjay Jain
Institute of Technology & Management(ITM), Gwalior
0751-2235158,91-9425724658
sanjayjainitm@gmail.com

R.S. Jadon
Madhav Institute of Tech. & Science(MITS), Gwalior
0751-2409365,91-9425122675
rsj_mits@yahoo.com

ABSTRACT

In this paper we propose a neural net learning based method for characterization of movies using audio information. We have first extracted the audio streams from the movie clips and then computable audio features are extracted from the audio streams. We than use neural net based classifier to classify the movie clips using these audio features. The features extracted from the clips are volume root mean square, volume standard deviation, volume dynamic range, zero crossing rate and salience ratio. We have demonstrated the effectiveness of our approach for characterizing the movie clips into action and non-action.

Categories and Subject Descriptors

Audio data processing, Soft computing.

Keywords

Audio features, Audio stream, Learning, Movie clips, Neural Network.

1 INTRODUCTION

Recent advances in multimedia compression technology, the significant increase in computer performance and the growth of Internet, have led to the widespread use and availability of digital video. The availability of audio-visual data in the digital format is increasing day by day. This data includes documents, audio-visual presentation, home made video and professionally created contents such as TV shows and movies. Movies constitute a large portion of the entertainment industry. Every year around 4500 movies are released all over the world which spans over 9,000 hours or half a tera byte of data and 33,000 television stations broadcast for twenty-four hours a day and produce eight million hours per year, amounting to 24,000 terabytes of data [5]. With the digital technology getting inexpensive and popular, there has been a tremendous increase in the volume and availability of movies through cable and Internet such as video on demand. Currently several web sites host movies and provide users with the facility to browse and watch online movies.

Therefore, automatic genre classification of movies is an important task. Movies are always preceded by previews (trailers) and promotional videos because of the commercial nature of movie productions. Directors often follow rules pertaining to the specific genre of a movie. Such rules are referred as Film Grammar or Cinematic Principles in the film literature. By following these principles, camera movements, sound effects, and lighting can create mood and atmosphere, induce emotional reactions, and convey information to the viewers. Although, different directors use these principles differently, movies of the same genre have a lot of features in common. For example, most of the action movies have similar shots and sound effects. People have tried to use the video part for classification by using the optical flow, panning, zooming values etc. but the audio based classification has an edge over the latter.

Recently several researchers have started to investigate the potential of analyzing the accompanying audio signal for video scene classification [1,4,8,9]. Obviously, audio information alone may not be sufficient for understanding the scene content, and in general, both audio and visual information should be analyzed. However, because audio-based analysis requires significantly less computation, it can be used in a preprocessing stage before more comprehensive analysis involving visual information. In this paper, we focus on audio analysis for scene understanding.

Audio in fact tells a lot about mood of the clip, the music component, the noise, fast or slowness of the pace and the human brain too can classify just on the base of audio. Our aim is to analyze audio cues from the movie clips and make an educated guess about its genre.

Rest of the paper is organized as follows: In section 2 we review the movie genres and grammar of film. Section 3, describes method for characterization of movie clips. In section 4, we describe extraction process of various audio features. Section 5 describes the neural net based learning system for characterizing the movie clips. Experimental results are given in section 6. Section 7 concludes the paper.

2 OVERVIEW OF MOVIE GENRES

Movie genres are various forms or identifiable types, categories, classifications or groups of movies that have similar, familiar or instantly recognizable patterns, filmic techniques or conventions that include one or more of the following: settings (prob), content, themes, plot, narrative events, motifs, styles, structures, situations, icons, characters (or characterizations), and stars.

The main movie genres are Action & non-action. Non-action

genre contains Adventure, Comedy, Crime, Drama, Animation, Science Fiction, Horror, Thriller, and War etc. On the other hand action contains Fire, Explosion [6, 7]. Film uses certain conventions often used by the filmmakers are known as film grammar [2, 7].

3 MOVIE CHARACTERIZATION METHOD

Our method first extracts the audio stream from the movie using the software Virtual Dub. We then perform the sampling of audio signal and converting it in to text format. Matlab's **wavread** and **dlmwrite** function are used. These functions sample the audio signal at 44.1 Khz and stored in the text files. After that we perform the audio features extraction from the sampled text files. Once we extracted the audio features from various clips. We perform the training of the Neural Network by inputting the extracted features and the actual category of the movie clips for all the sample sequences. The Network is then simulated for new samples by inputting their extracted features. Finally, we characterizes the movie clips into action, non-action.

4 AUDIO FEATURES EXTRACTION

In the present work we have extracted the audio-based features. The features extracted from audio streams are volume root mean square, volume standard deviation, volume dynamic range, zero crossing rate and salience ratio. These features are also reported by Liu [10].Based on these features the identifiable genres are action and non-action.

4.1 Volume Standard Deviation (VSD)

The standard deviation of the distribution of volume (of each frame) is calculated, and is treated as VSD. Thus, it is a clip level feature rather than a frame level feature.

$$\sigma = \sqrt{\left[\sum (A-\bar{A})^2/(n-1)\right]}$$

Where σ means standard deviation and $\bar{A}$ is the mean. Table 1 shows Volume Standard Deviation of some action & non-action movie clips. The result shows that the action movie clips has higher values of VSD.

4.2 Volume Root Mean Square (VRMS)

VRMS of the n^{th} frame is calculated, by the following formula:

$$V(n) = \sqrt{1/N \sum S_n^2(i)}$$

where $S_n(i)$ is the i^{th} sample in the n^{th} frame audio signal, and N is the total number of samples in the frame. Figure 4.1 clearly shows the difference of action and non-action clip based on volume Root Mean Square.

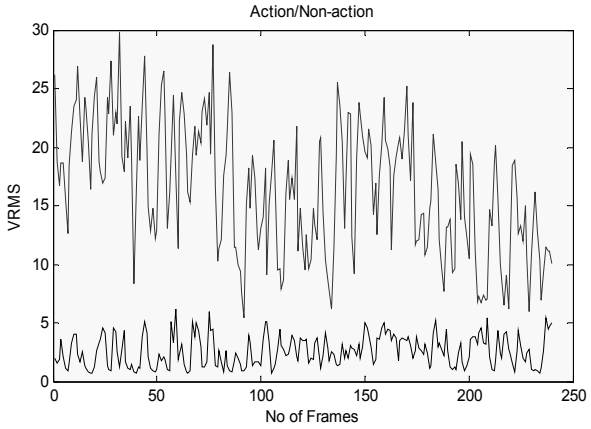

Figure 1. Plot the VRMS of action movie clip "The Bourne Identity" and the non action clip "Tea with Mussolini"

4.3 VDR (Volume Dynamic Range)

In audios with action in the background, volume of the frame does not change much, while in non-action audios, there are silent periods between the speeches, and hence VDR is expected to be higher. Table 1 shows the values of Volume Dynamic Range of some action and non-action movie clips. VDR is calculated as

$$VDR = [(MAX(v) - MIN(v)) / MAX(v)]$$

where MIN(v) and MAX(v) represent the minimum and maximum volume within a clip respectively.

4.4 Zero Crossing Rate (ZCR)

This feature is helpful in distinguishing audios with non action(pure speech), and action speech with some background disturbances (noise). Generally, non-action (pure speech) audio samples have a lower ZCR.

$$Z(n)=1/(2*N)\sum|sgn[S(m)] - sgn[S(m-1)]|$$

where,

$$sgn[S(m)] = \begin{cases} 1 & S(m) >= 0 \\ -1 & S(m) < 0 \end{cases}$$

and N is the number of samples in the clip.

4.5 Silence Ratio

Volume Root Mean Square(VRMS) and Volume Standard Deviation(VSD) are helpful to calculate Silence Ratio (SR) of a frame. Silence ratio is not calculated for each frame, but rather, Volume Root Mean Square and Volume Standard Deviation of each frame is used in calculating SR of a clip. Thus, it is a clip level feature rather than being a frame level feature. The silence ratio is calculated as follows:

$$SR(n) = sr/n$$

Where 'sr' is initially zero and is incremented by one if VRMS is less than the half of VSD in each frame and 'n' is the total number of frames in the clip. In action clips there are always

some noise in the background, which results in a low silence ratio. On the other hand silence ratio is much higher in non-action clips. Table 1 show the values of silence ration feature of some action & non-action clips.

Table 1. Shows the values of different features of action and non-action clips

Movie Clips	Genre	VSD	VDR	Silence Ratio
Bad Boys	Action	2.21	0.85	.00000
Terminator	Action	2.33	0.70	.00000
The Bourne Identity	Action	5.47	0.81	.00000
Spiderman	Action	4.71	0.85	.00000
Blue Streak	Action	1.87	0.93	.00833
American Pie2	Non-Action	1.89	0.86	.00000
Noting Hill	Non-Action	0.69	0.90	.00833
Tea With Mussolini	Non-Action	1.31	0.89	.00416
Princess Diary	Non-Action	1.89	0.98	.01666
Austin	Non-Action	1.55	0.95	.15000

5 NEURAL NET BASED LEARNING

Learning is a major functional characteristic of neural networks. We have used Multi-Layer Perceptron (MLP) architecture trained using the back propagation algorithm. A MLP is composed of layers of processing units that are interconnected through weighted connections. These layers are input, output and Intermediate layer called hidden layer. The network is trained using back propagation with three major phases. In the first phase an input vector is presented to the network, which is five extracted feature values in our case, which leads via the forward pass to the activation of the network as a whole. This generates a difference (error) between the output of the network and the desired output. In the next phase error is computed for the output unit and propagates this factor successively back through the network (error backward pass). In the final phase we compute the changes for the connection weights by feeding the summed squared errors from the output layer back through the hidden layers to the input layer. This process is continued until the connection weights in the network have been adjusted so that the network output has converged, to an acceptable level, with the desired output. The trained network is then given the new data and processing and flow of information through the activated network should lead to the assignment of the input data to the output class [3]. We have two categories in the output: action and non-action.

6 EXPERIMENTAL RESULT

We have implemented this system on windows platform in Java. Initially a training set comprising of previews from various movies is populated. We generated 40 samples in the training set from variety of action and non-action movie clips to train the network. We have tested the system for over 60 clips of 10 seconds each from 20 movies to simulate the network for the characterizations. The output of the trained network for different samples is shown in the Table 2. This data set is obtained from various sources, which includes Movie VCD, Promos etc. Virtual Dub is used for extracting the audio streams from the movie clips. MATLAB's is used for sampling the audio streams and convert it into text files. Core Java is used for all other tasks. The neural network methodology was implemented by using MATLAB's Neural Networking Toolbox.

7 CONCLUSIONS

In this paper we have presented a learning based scheme for classifying movies using volume root mean square, volume standard deviation, volume dynamic range, zero crossing rate and salience ratio Currently only two categories action/non-action are included, but with additional audio-visual features this scheme can be extended for characterizing other movie classes also.

8 REFERENCES

[1] C. Saraceno and R. Leonardi, 1997. Audio as a Support to Scene Change Detection and Characterization of Video Sequences, Proc. of ICASSP'97, Vol. 4, pp.2597-2600.

[2] Daniel Arijon,1976. Grammar of the Film Language. Hasting House Publishers, NY.

[3] D. E. Rumelhart, G. E. Hinton, and J. L. McClelland,1986. A general framework for parallel distributed processing. In Parallel Distributed Processing, The MIT Press, Cambridge, MA pp.45-76.

[4] J. Nam and A. H. Tewfik, 1997. Combined Audio and Visual Streams Analysis for Video Sequence Segmentation, Proc. of ICASSP'97, Vol. 3, pp.2665-2668.

[5] Peter Lyman and Hal R. Varian, 2000. School of Information Management and Systems at the University of California at Berkeley.

[6] Presents a comprehensive list of movie genres with their overviews, (http://us.imdb.com/Sections/Genres/).

[7] Sanjay Jain and R.S. Jadon, 2006. Features Extraction for Movie Genres Characterization, in Proceeding of WCVGIP-06.

[8] S. Pfeiffer, S. Fischer and W. Effelsberg, 1996. Automatic Audio Content Analysis, Proc. ACM Multimedia, pp.21-30.

[9] Y. Wang, J. Huang, Z. Liu, and T. Chen, 1997. "Multimedia Content Classification using Motion and Audio Information," Proc. of IEEE ISCAS' 97, Vol. 2, pp.1488-1491.

[10] Z. Liu, Y. Wang and T. Chen, 1998. Audio Feature Extraction and Analysis for Scene Segmentation and Classification, to appear in Journal of VLSI Signal Processing System.

ISSN 0974-1003

Table 2. Shows the output of the trained network for different samples

Movie Clips	No. of Clips	Actual Genre	Result	Remarks
American Pie2	04	Non action	Action	wrong
Austin	06	Non action	Non action	correct
Bad Boys	02	Action	Action	correct
Blue Streak	06	Action	Non action	wrong
Noting Hill	06	Non-action	Non action	correct
Princess Diary	04	Non action	Non action	correct
Spiderman	02	Action	Action	correct
Tea With Mussolini	07	Non action	Non action	correct
Terminator	04	Action	Action	correct
The Bourne Identity	08	Action	Action	correct