Middle-East Journal of Scientific Research 24 (S2): 226-231, 2016
ISSN 1990-9233
© IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.S2.155

Automatic Video Genre Classification Using Audio and Video Features

M. Divya and C. Balasubramanian

Dept of CSE, Mepco Schlenk Engineering College, Sivakasi, India

Abstract: Video genre identification is the process of identifying the genre of the video. Today, video analysis is a key issue in digital video application including video retrieval and video annotation. In our approach videos are initially classified into commercials, news, movies, sports, music, cartoons and documentaries by computing the visual features and audio features. Video features such as color moment, edge histogram, wavelet feature and local binary pattern which are extracted from each key frame of the video. Audio feature such as Mel Frequency Cepstral Coefficient is extracted from the audio signal of the video. Classification is performed using Support Vector machine. In our work total of 150 videos have been used for classifying the system.

Key word: Video · Genre · Classification

INTRODUCTION

As digital video libraries are becoming realistic, supported by several technological innovations including MPEG video compression, vast and high speed disk arrays, high speed local/wide area networks, etc. In order to make efficient use of the video data it should be labeled or indexed. Due to the availability of large digital video libraries, it is necessary to classify and categorize video content automatically so that users can easily search or choose the video. Video genre classification is a significant task for video database management, such as annotation, searching and indexing. Most classification algorithm uses both audio and visual cues to identify the video genre. There are large numbers of approach to content based classification of video data. These approaches could be broadly divided into three groups: text based approach, audio-based approach and visual-based approach. These approaches extract features from text, audio and visual. Text based categorization based on viewable text or automatic transcript from the video. The weighted voting method is used for automatic news video story categorization based on the closed caption text in [1]. First extract the set of keyword from the closed caption text and the categorization is done by calculating the likelihood score for each category. Linear time complexity is achieved for category prediction .Still it has an issue that is selecting the keyword. In some category first N keyword may not be related to its category.

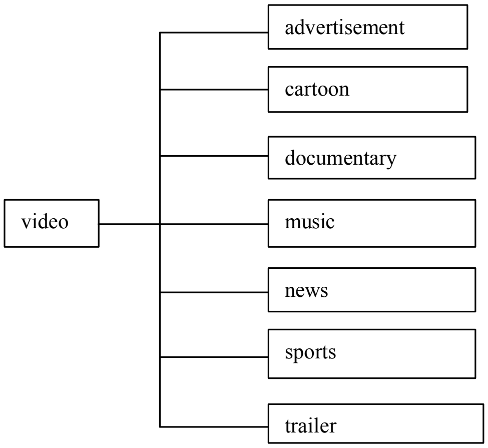


Fig. 1: Categories of videos involved in the proposed system

Fischer *et al.* [2] investigated automatic recognition of film genre by 3-level classification. At first level, some features of a video including motion energy, color histogram, spectrum and waveform of the video are extracted. At second level, attributes of videos are extracted from the syntactic characters of a video. At final level, temporal variation pattern of each style attribute is compared with the typical profile of video genre. Based on visual features, movies are broadly classified into four categories: Comedies, Action, Dramas or Horror films. Features like average shot length, color variance, motion content and lighting key are computed. In this work, M ean SHIFT Classification is used. In [3],

Corresponding Author: M. Divya, PG student, Dept of CSE, Mepco Schlenk Engineering College, Sivakasi, India.

author uses the data driven learning based approach to classify a video programme into a set of five pre-defined genre including sports, cartoons, news, commercials and music. They uses the features from short term spectral estimates of audio signals and motion dynamics of video signals and use Gaussian Mixture Model as a classifier. The classification accuracy is achieved is approximately 74%, 73% and 87% respectively. Others use the audio patterns to classify TV program genre [4]. In which audio patterns are extracted to classify TV programs such as news and talk shows. It did not use the inter-relationship between the information given by the audio. In this paper, we use audio and visual feature to classify a video into commercials, news, movies, sports, music, cartoons and documentaries. We extract the video features and audio features that are used to distinguish between the different classes. Visual information is described using color moments, edge histogram, wavelet and local binary pattern. Mel Frequency Cepstral Coefficients are used for audio processing. Support vector machine have been used for the classification task. The rest of the paper is organized as follows: section II presents the data which are used for evaluation and section III presents the proposed video genre classification system and next session describes the experimental results and last session concludes the video genre classification based on audio and video.

Task and Corpus: Experiments are conducted on corpus composed of videos which belongs to one of the seven categories: documentaries, music, movies (trailers), commercials, sports, news and cartoons. The database contains 150 videos. Each category has 20 videos. Videos are relatively short: from 2 to 5 minutes long.

Proposed System: The overview of genre classification system is shown in Fig 2. The input videos have been preprocessed. Next to it, visual features are extracted from the key frames which are extracted from the preprocessing. Visual features include color moments, edge histogram, wavelet feature and local binary patterns. Then the 13 MFCC coefficients are computed from the audio signals of video. Once the feature is extracted, SVM classifier is used for the classification task. A. Video Preprocessing This section represents the video processing which is required for further processing. In preprocessing the video has been split into several shots/frames. Then key frames are extracted from the frames obtained from the video.

Shot Separation: To detect the shot, first the frames are extracted from the given video. From the detected frames, the histogram has to be evaluated for the consecutive frames. Based on the histogram values the difference is calculated between every consecutive frame.

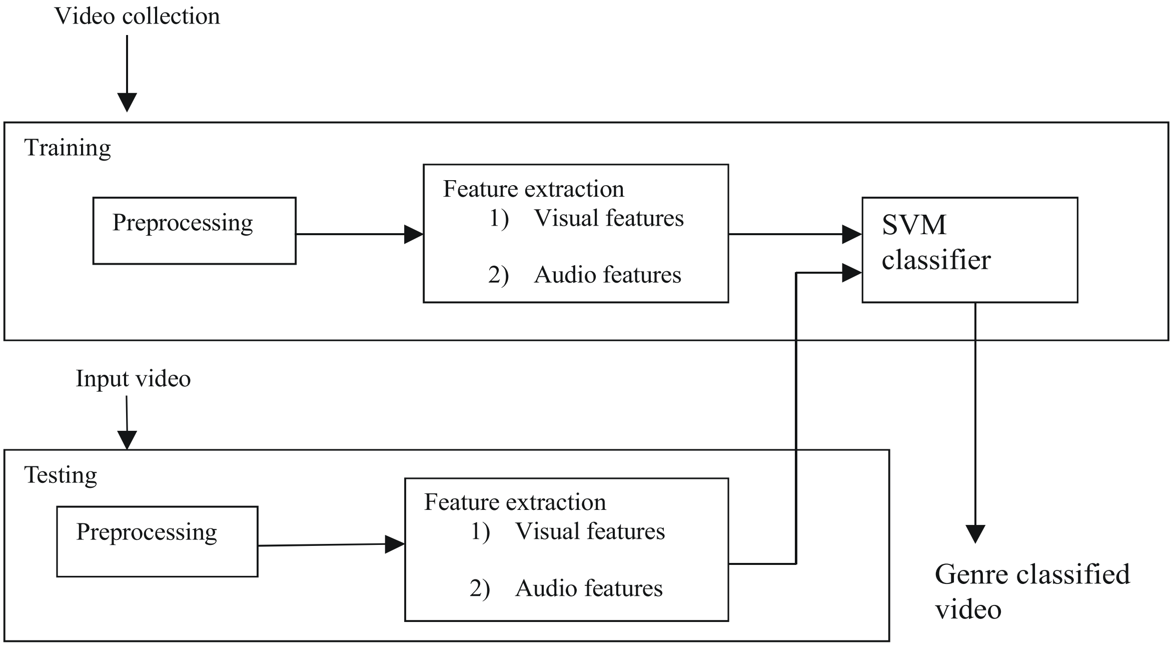


Fig. 2: Genre classification system

The mean and standard deviation is estimated to fix the threshold value. Comparing the threshold value with the calculated difference between consecutive frames, the shot is detected. Fig 3 shows the flow chart to detect the shot from the video.

Key Frame Extraction: We use the entropy difference method which is used in [5] to identify the key frames from each shot. More similar frames and duplicate key frames are removed by finding the similarity between the frames.

Feature Extraction: There are several features which can be extracted from the key frames. In our work, visual features and audio features are extracted. Visual features include color moments, edge histogram, wavelet features and local binary pattern. Audio features include computing the Mel Frequency cepstral Coefficients for the audio signal of the video.

Color Moments: Color moments are used to recognize images based on their features of colors. There is an assumption that distribution of color in an image can be interpreted as a probability distribution. So that we have considered the first three moments (mean, standard deviation and skewness). Once calculated, these moments provide a measurement for color similarity between images. Three central moments are calculated for each channel (red, green, blue) of the color image. First we split the image into 5×5 sub-image and for each region we compute these moments for the three color channel to produce 25×9=225 dimensional vector. Let N be the quantized color and Pi be the number of pixels of the ith color then the mean, standard deviation and skewness are calculated by

$$E_i = \sum_{N}^{j=1} \frac{1}{N} p_{ij} \tag{1}$$

$$\sigma_i = \sqrt{\frac{1}{N} \sum_{N}^{j=1} \left(p_{ij} - E_i\right)^2} \tag{2}$$

$$s_i = \sqrt[3]{\frac{1}{N} \sum_{N}^{j=1} \left(p_{ij} - E_i\right)^3} \tag{3}$$

Edge Histogram: Edge in the image is considered as a significant feature to represent the content of the image. Edge descriptor uses the 4 directional (vertical, horizontal and 45 degree and 135 degree directional edges) and non-directional edge. First we split the image into 4×4 sub image and find 5 edge filters for each sub images to produce 16×5=80 dimensional vector. After the edge extraction from the image block, we count the total number of edges for each edge type in sub images. For which we have used 5 sobel edge filters. Then the sub image is convolved with the filter coefficients.

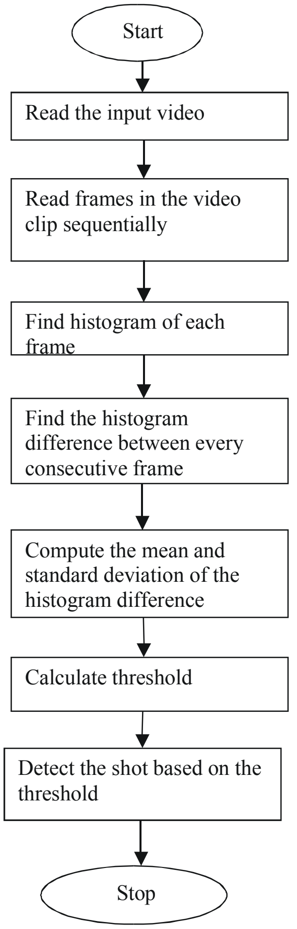
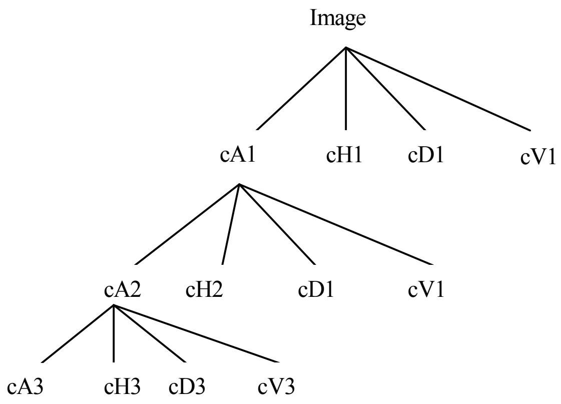

Start
Read the input video
Read frames in the video clip sequentially
Find histogram of each frame
Find the histogram difference between every consecutive frame
Compute the mean and standard deviation of the histogram difference
Calculate threshold
Detect the shot based on the threshold
Stop

Fig. 3: Flow diagram for shot detection

Wavelet Features: It provides the representation of the image texture from the energies of wavelet sub bands. First we split the image into 3×3 sub images then apply haar wavelet to get 9 wavelet coefficients for each region thus provide 9×9=81 dimensional vector. Using haar wavelet, three level decomposition is achieved as shown in Fig 4. In first level decomposition one approximation coefficient and three detail coefficients are obtained from the image. Then the approximation coefficient is further decomposed into 1 approximation and 3 detail coefficients. Finally, at the third level decomposition we have obtained the 9 detailed coefficients.



Local Binary Pattern: Local binary pattern looks at points surrounding a central pixel and check whether the surrounding points are greater than or less than the central point. If the surrounding points are greater than the central point then put 1 otherwise 0. Here we have used the 8 neighborhood to calculate the local binary pattern.

$$LBP_{P,R} = \sum_{p=0}^{P-1} s\left(g_p - g_c\right)2^p \qquad (4)$$

$$s(x) = \begin{cases} 1, ifx \geq 0 \\ 0, otherwise \end{cases} \qquad (5)$$

Where
P- pixel neighborhood
R- radius

Mel Frequency Cepstral Coefficients:MFCC is the most commonly used feature extraction technique in ASR system which is used in for automatic speech recognition. First speech signal is divided into overlapped frames. Then the features are extracted from each frame. Several operations such as pre-emphasis, framing, windowing and Mel Spectral analysis are performed at the input signal of the video. We have derived 13 MFCC from each frame. MFCC is calculated by

$$\text{Mel(f)} = \frac{1000}{\log(2)} \log\left(1 + \frac{1}{1000}\right) \qquad (6)$$

SVM Classification: Support vector machine is trained to distinguish the features of one category video from all other categories. There are two phase: training phase and testing phase. In training phase, features are extracted and tags are given by the user for classification. Then the extracted features are trained and model is created. In testing phase, after feature extraction of the given image, extracted features are compared with the model which will decide the class. Based on the features of the input it will be classified into corresponding classes. In our work, multiclass classification is achieved using one-against all approach.

Algorithm for Training:

```
Let training video set = $V_{train}$
Where Vtrain= { V1, V2…….Vn}
For i=1 to n
  Extract frame F= { F1, F2,……Fj}
  For i=1 to j
    Find key frame based on entropy difference
    method (key[t] )
  End
  For i=1 to t
    Feature_vector=features( key[i])
    Label= class label given by user
  End
end
```

Algorithm for Testing:

```
Let test video= Vtest
Extract frame F= { F1, F2,……Fj}
  For i=1 to j
    Find key frame based on entropy difference
    method (key[t])
  End
  For i=1 to t
    Feature_vector = features( key[i])
    Label = SVMclassify(key[i])
  End
```

EXPERIMENTAL RESULTS

The experiments are carried out on the video dataset comprising of commercials, news, movies, sports, music, cartoons and documentaries. For each category 20 videos are collected from YouTube. For each video genre, half of the videos were used for training and remaining videos are used for testing. Multiclass SVM was constructed for each class using one against all approach.

Visual feature such as color moments, edge histogram, wavelet feature and local binary patterns are extracted. Those obtained feature vector are given as an input to SVM classifier. Table 1 shows the classification result of multiclass SVM for considered seven video genres.

Table 1: Classification results of SVM for visual feature

advertisement		Cartoon	
Correctly classified	Incorrectly classified	Correctly classified	Incorrectly classified
15	5	14	6
13	7	12	2
16	4	13	7
17	3	16	4
documentary		music	
Correctly classified	Incorrectly classified	Correctly classified	Incorrectly classified
13	7	11	9
12	8	6	12
16	4	15	5
19	1	16	4
News		sports	
Correctly classified	Incorrectly classified	Correctly classified	Incorrectly classified
15	5	14	6
19	1	12	8
18	2	17	3
10	10	7	13

trailer	
Correctly classified	Incorrectly classified
12	8
10	10
6	14
17	3

Audio features such as 13 Mel Frequency Cepstral Coefficients are extracted from the audio signal of video. Table 2 shows the classification results of multiclass SVM for considered seven video genres.

Table 2: Classification results of SVM for audio feature

advertisement		Cartoon	
Correctly classified	Incorrectly classified	Correctly classified	Incorrectly classified
14	6	18	2
9	11	16	4
13	7	11	9
16	4	12	8
documentary		music	
Correctly classified	Incorrectly classified	Correctly classified	Incorrectly classified
15	5	6	12
13	7	16	4
11	9	8	12
14	6	17	3
News		sports	
Correctly classified	Incorrectly classified	Correctly classified	Incorrectly classified
18	2	14	6
14	6	13	7
10	10	16	4
12	8	8	12

trailer	
Correctly classified	Incorrectly classified
13	7
9	11
12	8
8	12

From the observation news, documentaries and carton videos provide better classification result.

CONCLUSIONS AND FUTURE WORK

In this work videos were collected and preprocessed. After preprocessing audio and visual features are extracted. Once features are extracted multiclass SVM is constructed to classify the video into seven categories: commercials, news, movies, sports, music, cartoons and documentaries. In future the system could be extended to more number of genres and sub categories like cricket, basketball which are all come under sports.

ACKNOWLEDGEMENT

We wish to thank the Management and the Principal of Mepco Schlenk Engineering College, for the support in carrying out this research work. We are extremely thankful and indebted to our beloved head of the Department Dr. K Muneeswaran for sharing his expertise and for his valuable guidance and encouragement extended to us.

REFERENCES

1. Xu, L.Q., M. Roach and J. Mason, Classification of non-edited broadcast video using holistic low-level features, in Proc. Int. WorkshopDistrib. Comput. (IWDC), 2002.
2. Weiyu Zhu, C.T. and S.P. Liou, 2001. Automatic news video segmentation and categorization based on closed-captioned text, in Proc. Int. Conf.Multimedia Expo (ICME), 2001.
3. Jasinchi, R. and J. Louie, Automatic TV program genre classificationbased on audio patterns, in Proc. Euromicro Conf., 2001.
4. Rasheed, Z., Y. Sheikh, M. Shah, 2005. On the use of computable features for film classification, IEEE Trans. Circuits Syst. Video Technol., 15(1): 52-56.
5. Moncrieff, S., S. Venkatesh and C. Dorai, 2003. Horror film genre typing and scene labeling via audio analysis, in Proc. Int. Conf. Multimedia and Expo (ICME).