A Review of Video Classification Techniques

Mittal C. Darji[1], Dipti Mathpal[2]

Assistant Professor, Information Technology Department, G.H. Patel College of Engineering & Technology, Gujarat, India
Trainee Assistant Professor, Information Technology Department, G.H. Patel College of Engineering & Technology, Gujarat, India

---***---

Abstract - *Video classification literature has been reviewed and techniques for the same are provided here in this paper. Classification process in general requires features based on which one can distinguish among the categories. These features are mainly taken from text, audio or visual content of the video. Based on that mainly three classification techniques are there as discussed here. Based on the application user has to select the method and features. Pros and cons of each method are mentioned in this paper with suitable applications.*

Keyword- Video classification, Text based classification, Audio based classification, Video based classification, features

1. INTRODUCTION

The amount of video achieves that we have are increasing tremendously day by day. Use of internet and latest technologies are making it easy to share videos. This is leading to lots of duplication too. Finding out the type of videos you want to see is a very difficult task. Such a time consuming and tedious job must be made automatic. This automation task is called as video classification by researchers.

Video classification has been used to classify videos into categories like sports, comedy, news, dance, horror etc. Some researchers have also classified a single video into parts of different categories. All these classifications require the characteristics which differ for each category. These characteristics are called features.

Features can be extracted from any of the three components: Text, Audio and Video [1]. Researchers have used all the three in various ways for fulfilling their purposes of classification. This paper has summarized the methods and features used over the time.

Rest of the paper is organized as follows: In section II we will describe the text based method. In section III we will see how audio based approach is used. Section IV contains the video based methods. Comparison of all these methods is described in section V. We will conclude in last section number VI.

2. TEXT BASED CLASSIFICATION

In this method, we produce text from video and analyze it for classification. Text can be: 1) visible text on screen 2) text extracted from the speech [2]. In first category, the text visible on screen is extracted. For example, the score board of game, number on jersey of player, captions written on the screen etc. Such text can be extracted using Optical Character Recognition (OCR). In second category, the text is extracted from speech using speech recognition. This method is mainly used in providing subtitles or closed captions. Closed captions are mostly used to provide other types of sound such as a sound of animal or music. Subtitles are placed on screen to provide understanding in a familiar language.

This text based research can also be used in document text classification and areas like handwritten text to digital document conversion, signature verification, handwriting matching etc. However, the problem is that such text is in a large amount and hence is difficult to deal with. Also, OCR is having a higher rate of errors. Text extracted from OCR will mostly contain a higher amount of spelling mistakes and omissions. A commonly used method while working with text is to represent the text using feature vector in bag-of-words model. This model uses the number of occurrence of any word. But this model does not contain the information about the order of these words in document.

3. AUDIO BASED CLASSIFICATION

This approach is more used than text based in research and it is because audio processing requires lesser computational recourses and time. Storage of audio and its features requires lesser space than the video and text. To process audio, signal is sampled on a particular rate and from each sample certain features are extracted for review. These sampling windows can be overlapped in some cases. Suitable features from sampled signal are extracted based on the application requirement. Features of audio can be broadly classified in either physical features or perceptual features [3].

3.1 Physical Features

These are also called as tie domain features as they are directly measured from frequency values of the signal [6]. These are also called as low level features of signal.

Amplitude values of sampled signal are directly used to compute these feature values. Such features are: Zero Crossing Rate (ZCR), Short Time Energy (STE), Spectral Roll-off, Spectrum Centroid, Spectral Flux, Fundamental Frequency, Mel-Frequency Cepstral Coefficient (MFCC) etc.

3.2 Perceptual Features

Psychological acoustic model is proposed that measures the perceptual features of sound based on the human perceptual system for sound [4]. Human understands the sound based on his perceptual towards what he heard. These are the features of sound that defines it hearing characteristics. Such features are: Loudness, Pitch and Timbre. Out of them loudness and pitch are mostly used. Timbre is used to differentiate between the sounds that displays similar values of loudness and pitch but are actually different.

From the videos, audio signals are extracted and from audio signals features are extracted. These features contain variation in values based on the category of audio. For example, a female voice is having higher pitch value than male voice. Music is having higher continuous amplitude than speech. Music also contains higher ZCR than speech due to a frequent variation in amplitude. These kinds of observations are used by researchers to classify videos.

4. VIDEO BASED CLASSIFICATION

Most of the researchers have used this method as human perceives most of the information based the vision. Also some researchers have combined these visual aspects with audio and text whenever required. Visual features are mostly extracted from the frames of video or from the shots of video. Basic construction of a video is like: fundamental part of video is a frame. Hence video can be called as a collection of frames. More than one frames shot from a single camera action is called a shot. Scene is one or more shots from the video. Most of the researchers have used shots based approach as it is the most natural and understandable aspect to segment video. But the problem that is faced in this is that it is difficult to get the exact boundaries of shots through the automatic methods [7]. Many a times either the boundaries we get are overlapped or they don't provide correct separation.

Visual features are mostly color based, motion based or based on length of shots. These features need to provide the information of lights, action, background or pace of video. Visual features are mostly as mentioned below:

4.1 Color-Based Features

Video frame is composed of pixels and each pixel is represented by a set of values from a color space [8]. To represent colors various color models have been proposed and from those, RGB (Red Green Blue) and HSV (Hue Saturation Value) are widely used. RGB model represents the amount of each color in particular pixel. HSV model represents hue that is wavelength of color, saturation that is how pure is the color and value that is lightness or brightness of color.

Distribution of color in a frame is often represented by color histogram. It represents the number of pixel in a frame for each possible color. Mostly histograms are used to compare two frames. But the disadvantage is that we cannot find the exact pixel with particular color. Another problem can be, the frames may have different lightning conditions and hence for comparison preprocessing is required.

4.2 Shot-Based Features

For this we need to first separate the shots from video. Various boundary detection techniques are used but they may not give always correct detections. Boundaries can be hard cut, faded or dissolve. Hard cut shows an abrupt change in color intensities. Hard cut are the shots in which one shot ends abruptly and another begins [9]. Faded shot may fade out slowly or may fade in slowly. Dissolve is a gradual transition from one shot to another where last shot fades out and next shot fades in. these shot transition types can also be used as a feature. Simplest method for finding shots is to take color histogram difference of frames [10]. RGB or HSV both can be used for this.

4.3 Object-Based Features

This approach is not much used as it is difficult to detect and identify objects from video frames. In this approach, first the objects are identified and then features are detected from them. For example, faces can be detected from video and then features like skin tone, texture, size, position etc are extracted [11] [12].

5. COMPARISON

Each of the three methods of classification is very well explored and used by researchers. One has to select any one based on the suitability to application. The table below explains the suitability of each method in detail.

TABLE 1: Comparison of three classification approaches

Classification Method	Feature Type	Advantages / Disadvantages	Application
Text Based Classification	OCR Closed Captions Speech Recognition	Computationally expensive Higher dimensionality Higher error rate	Reading score board Providing subtitles Reading headlines form news video
Audio Based Classification	Physical Features Perceptual Features	Shorter in length and size Computationally cheaper Difficult to differentiate similar sounds	Classifying movie into dialogs and songs Classifying videos into horror, action, comedy Classifying video into speech, music, environmental sound
Video Based Classification	Color-Based Features Shot-Based Features Object-Based Features	Larger size Computationally expensive Preprocessing is required Difficult to identify shots, not accurate	Object tracking Video summarization Separating news video and sight scenes Classification of different sports videos

6. CONCLUSION

Various video classification literatures have been reviewed and it is observed that mainly three approaches are used: 1) Text 2) Audio and 3) Video. Various features are reviewed in each of them. Also, many of the researchers have used combination of these features too based on the requirement of application. A lot of researchers have put efforts in this area but still this is an emerging area which requires ideas to be used in terms of performance, resource utilization and practical implementation.

REFERENCES

[1] D. Brezeale and D. J. Cook, "Automatic Video Classification: A Survey of the Literature", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, 2007.

[2] M. C. Darji, Dr. N. M. Patel, Z. H. Shah, "A REVIEW ONAUDIO FEATURES BASED EXTRACTION OF SONGS FROM MOVIES", International Journal of Advance Engineering and Research Development (IJAERD) e-ISSN: 2348 - 4470 , print-ISSN:2348-6406.

[3] Burred J J, Lerch A (2004) Hierarchical Automatic Audio Signal Classification. J Audio Engineering Society 52(7/8):724-739

[4] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search, and retrieval of audio," IEEE MultiMedia, vol. 3, no. 3, pp. 27–36, 1996.

[5] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," Journal of VLSI Signal Processing Systems, vol. 20, no. 1-2, pp. 61–79, 1998.

[6] U. Srinivasan, S. Pfeiffer, S. Nepal, M. Lee, L. Gu, and S. Barrass, "A survey of mpeg-1 audio, video and semantic analysis techniques," Multimedia Tools and Applications, vol. 27, no. 1, pp. 105–141, 2005.

[7] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," in In SPIE Conference on Storage and Retrieval for Image and Video Databases VII, vol. 3656, 1999, pp. 290–301.

[8] C. Poynton, A Technical Introduction to Digital Video. New York, NY: John Wiley & Sons, 1996.

[9] Y. Abdeljaoued, T. Ebrahimi, C. Christopoulos, and I. M. Ivars, "A new algorithm for shot boundary detection," in

Proceedings of the 10th European Signal Processing Conference, 2000, pp. 151–154.

[10] H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," Multimedia Systems, vol. 1, pp. 10–28, 1993.

[11] P. Wang, R. Cai, and S.-Q. Yang, "A hybrid approach to news video classification multimodal features," in Proceedings of the Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing and the Fourth Pacific Rim Conference on Multimedia, vol. 2, 2003, pp. 787–791.

[12] X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, and S. Li, "Automatic video genre categorization using hierarchical SVM," in Proceedings of IEEE International Conference on Image Processing (ICIP), 2006, pp. 2905–2908.

[13] S. M. Doudpota, S. Guha,"Mining Movies to Extract Song Sequences", ACM 978-1-4503-0841-0 MDMKDD'11, August 21, 2011.

[14] D. Brezeale, D. Cook, "Automatic Video Classification: A Survey of the Literature", IEEE Transactions in Volume:38, Issue: 3, 2008.

[15] C. V. Jawahar, B. Chennupati, B. Paluri, N. Jammalamadaka, "Video Retrieval Based on Textual Queries", Proceedings of the Thirteenth International Conference on Advanced Computing and Communications, Coimbatore, December 2005.

[16] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, "Speech/Music discrimination for multimedia applications", Acoustics, Speech, and Signal Processing, ICASSP, Proceedings, IEEE International Conference in vol. 6 2000.