A survey on video classification techniques

Nirav Bhatt

Post Graduate Student
Computer Science & Engineering Department
B.H.Gardi College of Engineering & Technology Rajkot, Gujarat, India

***Abstract*- There are many videos available in this present day. To help the viewers to find video of their interest, that started to work on methods of automatic video classification. In this paper we survey the video classification literature. We find three main approaches for classification as: text, audio and visual. And then compared all of that approaches. We conclude with ideas for further research.**

***Index Terms*- Video, Video classification techniques, Text based approach, Audio based approach, Visual based approach**

1. INTRODUCTION

Today people have access to enormous amount of video, both on Internet and on Television. The amount of video that a viewer has to choose from is now so large that it is infeasible for a human to go through it all to find video of their interest. Viewer will narrow their choices by looking for video within specific categories or genre. Because of the very large amount of video to categorize, research has begun for classifying video automatically.

For performing automatic classification of video, a large number of approaches have been attempted. After review process of methods we can categorize this approaches in to three groups: text-based approaches, audio-based approaches and visual-based approaches.

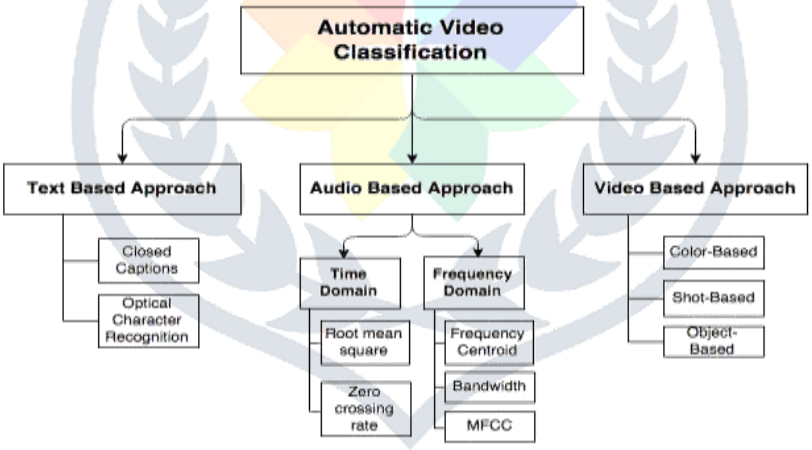

Fig 1: Automatic video Classification approaches

2. TEXT BASED APPROACH

Least common approach in video classification is text only approach. Text produced from a video falls into two categories. Firstly, viewable text could be text on objects that are filmed (scene text), a cricketer's name on a jersey or the name on a name plate of house or it could be text placed on bottom of the screen such as the score for a sports event or subtitles. Secondly it could be text placed on screen, it could be text on objects that are filmed (scene text), Text features are produced from this viewable text by identifying text objects followed by the use of optical character recognition (OCR) [1] to convert these objects to usable text.

Closed captioning will also use for placed text on the television screen with open captioning or subtitling. Both servers the same purpose, however in open captioning the text is actually part of the video and it would need to be extracted using text detection methods and OCR. Subtitles are meant for people who can hear the audio of a video but they can't understand it because

audio is in another language or because it is unclear; that why, subtitles typically won't include references to non-dialog sounds.

One advantage of text-based approaches is that they can utilize the large body of research conducted on document text classification [2]. Main advantage is that the relationship between the words and specific genre is very easy for humans to understand. For example, few people would be surprised to find the words 'six', 'no-ball', and 'out' in a transcript from a cricket game.

Difficulty in using text-based features is that the text derived from OCR of on-screen text has fairly high error rates [3]. While the closed captions for a movie is approaches a given quantity as limit.

3. AUDIO BASED APPROACH

Audio-only approaches are more frequent in the video classification literature than text-only approaches. Advantage of audio-based approaches is that they require only small number of computational resources than any other methods. And if we want to store the features, audio features will require lesser space.

We can derive the features from the time domain and the frequency domain. Brief description for some commonly used audio features is as follows.

- **Time-Domain Features:**
  The root mean square (RMS) of the signal energy approximates the human perception of the loudness or volume of a sound [4]. Zero crossing rate (ZCR) is the number of signal amplitude sign changes in the current frame. Higher zero crossing rates are resulted from higher frequencies. Music has lower variability of ZCR that in speech. If the ZCR and loudness are below thresholds, then it's possible to have this frame silent. The silence ratio is the proportion of a frame with amplitude values below some threshold. Music normally has lower silence ration than speech. Commercial channels has lower silence ratio than news cannels.

- **Frequency-Domain Features:**
  The energy distribution is the signal distribution across frequency components. The Frequency centroid, which approximates brightness, is the midpoint of the spectral energy distribution and provides a measure of where the frequency components are concentrated [5]. Normally brightness is lower in speech than in music, whose frequency is normally above 7 kHz.

  Bandwidth is a measure of the frequency range of a signal [6]. Many type of sounds have more narrow frequency ranges than the other sound. Speech typically has a lower bandwidth than music.

  Mel-frequency cepstral coefficients (MFCC) are produced by taking the logarithm of the spectral components and then placing them into bins based upon the Mel frequency scale, which is perception-based. This is followed by applying the discrete cosine transform (DCT) [7]. The DCT has good energy compaction, that is, after transforming a set of values, most of the information needed to reconstruct those coefficients is concentrated in a few of the new coefficients. By keeping those coefficients values in which most of the energy is concentrated, the dimensionality can be reduced while still allowing approximations of the original values to be produced.

4. VIDEO BASED APPROACH

Usually we called a video, a collection of images, which are known as frames. Number of frames within a single camera action is called a shot. A scene is one shot or number of shots. For example, a conversation between two people in a movie may be filmed such that only one person is shown at a time. Time to time the camera appears to stop and move to another person represents a shot change, but the group of shots that represent the entire conversation is a scene.

Visual-based approaches choose shots since a shot is a way to segment a video and each segment represents "conversation between people" or "Bike riding scene". Shot represents by a one frame, called it the keyframe. Keyframe is the proceeding frame of a shot. Movies that focus on action tend to have shots of shorter duration than those that focus on character development [8]. One problem with using shot-based methods is that the methods for automatically identifying shot boundaries don't always perform well [9]. Scene identification is very difficult. So there are some video classification approaches to do so.

- Color-Based Features:
  A video frame is composed of a set of dots known as pixels and the color of each pixel is represented by a set of values from a color space [10]. Many colour spaces exist for representing the colours in a frame. Two most popular spaces are hue saturation- value (HSV) and the red-green-blue (RGB) colour spaces. In the RGB colour space, the colour of each pixel is

represented by combination of the individual colours red, green and blue in some amount. In the HSV colour space, colours are represented by hue (i.e., the wavelength of the colour percept), saturation (i.e., the amount of white light present in the colour), and value (also known as the brightness, value is the intensity of the colour) [11]. The distribution of colours in a video frame is represented using a color histogram, that is, a count of how many pixels in the frame exist for each possible color. Color histograms are used for comparing two frames, which may have similar counts for similar frames. We can't determine the positions of pixels with clearly defined colours using colour histogram. To overcome these we can divide a frame into many regions and then apply a color histogram to each of the region of frame to capture spatial information.

- Shot-Based Features:

  Detection of shot is necessary to use it. Its very difficult task to automate, in part because there are several ways of making transitions from one shot to the next. Shot transitions are comes under following categories: hard cuts, fades, and dissolves. Hard cuts are those in which one shot abruptly stops and another begins [12]. There are two types of fades: a fade-out that consists of a shot gradually fading out of existence to a monochrome frame and a fade-in occurs when a shot gradually fades into existence from a monochrome frame. One shot fading out while another shot fades in is comes under a dissolve; both shots features we can see in this process.

  One of the simplest methods for detecting shots is to take the difference of the color histograms of consecutive frames, with the assumption that the difference in color histograms of frames within the same shot will be smaller than the difference between frames of different shots [13]. This approach is easy to implement but has a many problems. One problem is to decide the threshold difference which must use declare a change in shots. Shots that contain a little of motion require a lesser threshold value than those with lot motion. Also, the threshold value is likely to be different for different videos and even within the same video no particular value may correctly identify all shot changes [14]. High threshold value lead to miss some shot changes. Too low threshold value lead to identify shot changes that don't exist.

  Truong et al. [15] detect shot changes with shot transitions of the types hard cut, fade-in, fade-out, and dissolve [16]. Global threshold can detect the hard cuts to identify potential cuts, and then adaptive threshold will be used for applying a sliding window to these frames. We can detect fade-ins and fade-outs by identifying monochrome frames and then we have to check if the first derivative of the luminance mean is relatively constant or not. We can also detect a dissolves by taking the first order difference of the luminance variance curve which falls within a range calculated from the luminance variances of the shots preceding and succeeding the dissolve.

- Object-Based Features:

  It is uncommon feature. Object-based features are difficulty in detecting and identifying objects and the computational requirements to do so. When they are used, they tend to focus on identifying specific types of objects, such as faces [17] [18]. After object detection we can easily derive features from them like, dominant color, texture, size, and trajectory.

5. COMPARISON OF FEATURES

Table 1: Comparison Of Features

Feature Type	Pros/Cons
Text Features **Closed-Captions** **OCR**	High accuracy may achieve while not produced in real-time, high dimensionality. Its expensive, It can extract videotext; which is not present in a dialog.
Audio Features	Require lesser computational resources than visual features, difficulty in distinguishing multiple sounds.
Visual Features **Color-Based** **Shot-Based** **Object-Based**	 Simple to implement, not refined representation. Difficult in identifying shots automatically, its may not be accurate. Difficult, limited number of objects.

6. CONCLUSIONS

We have reviewed the literature on video classification. We found a various approaches for video classification. We can categorize these approaches in to three groups: text-based approaches, audio-based approaches and visual-based approaches.

Audio based approach is better for classify the video, as it require few computation recourses. We conclude that we can combine as many as techniques achieve better classification results on video. Still we have an opportunity to classify videos in different ways. We can classify video by first segmenting it and then apply thresolding on it, and in this way you can apply or develop new techniques.

We could apply classification techniques on movies. To classify the portion of the movie just like, songs, fight scene, comedy scene, so there is an opportunity for video classification here as well.

7. ACKNOWLEDGEMENT

I am deeply indebted & would like to express gratitude to my thesis guide Prof. Aspriha R. Das, B. H. Gardi College of Engineering & Technology for his great efforts and instructive comments in the dissertation work.

I would also like to extend my gratitude to Prof.Hemal Rajyaguru, Head of the Computer Science & Engineering Department, B. H. Gardi College of Engineering & Technology for his continuous encouragement and motivation.

I would also like to extend my gratitude to Prof. Vaseem Ghada, PG Coordinator, B. H. Gardi College of Engineering & Technology for his continuous support and cooperation.

I should express my thanks to my dear friends & my classmates for their help in this research; for their company during the research, for their help in developing the simulation environment.

I would like to express my special thanks to my family for their endless love and support throughout my life. Without them, life would not be that easy and beautiful.

References

[1] A. Hauptmann, R. Yan, Y. Qi, R. Jin, M. Christel, M. Derthick, M.Y. Chen, R. Baron, W.H. Lin, and T. D. Ng, "Video classification and retrieval with the informedia digital video library system," in Text Retrieval Conference (TREC02), 2002.

[2] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[3] A. G. Hauptmann, R. Jin, and T. D. Ng, "Multi-modal information retrieval from broadcast video using ocr and speech recognition," in JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, 2002, pp. 160–161.

[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] 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.

[6] G. Lu, "Indexing and retrieval of audio: A survey," Multimedia Tools Applications, vol. 15, no. 3, pp. 269–290, 2001.

[7] B. Logan, "Mel frequency cepstral coefficients for music modeling," in International Symposium on Music Information Retrieval, 2000.

[8] N. Vasconcelos and A. Lippman, "Statistical models of video structure for content analysis and characterization," IEEE Transactions on Image Processing, vol. 9, no. 1, 2000.

[9] 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.

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

[11] A. D. Bimbo, Visual Information Retrieval. San Francisco, CA: Morgan Kaufman, 1999.

[12] 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.

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

[14] R. Jadon, S. Chaudhury, and K. Biswas, "A fuzzy theoretic approach for video segmentation using syntactic features," Pattern Recognition Letters, vol. 22, no. 13, pp. 1359–1369, 2001.

[15] B. T. Truong, C. Dorai, and S. Venkatesh, "Automatic genre identification for content-based video categorization," Proc. 15th International Conference on Pattern Recognition, vol. IV, pp. 230–233, 2000.

[16] B. T. Truong and C. Dorai and S. Venkatesh, "New enhancements to cut, fade, and dissolve detection processes in video segmentation," in Proceedings of the eighth ACM international conference on Multimedia (MULTIMEDIA '00), 2000, pp. 219–227.

[17] 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.

[18] 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.