International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 6 Issue: 3 01 - 05

Video Classification:A Literature Survey

Pravina Baraiya[1]
Department of Information Technology,
Shantilal Shah Engineering College,
Bhavnagar, India
palakbaraiya@gmail.com

Asst. Prof. Disha Sanghani[2]
Department of Information Technology,
Shantilal Shah Engineering College,
Bhavnagar, India
dishasanghani83@yahoo.in

***Abstract*—**At present, so much videos are available from many resources. But viewers want video of their interest. So for users to find a video of interest work has started for video classification. Video Classification literature is presented in this paper. There are mainly three approaches by which process of video classification can be done. For video classification, features are derived from three different modalities: Audio, Text and Visual. From these features, classification has been done. At last, these different approaches are compared. Advantages and Dis-advantages of each approach/method are described in this paper with appropriate applications.

Keywords*-Video Classification; Audio-Based Approach; Text-Based Approach; Visual-Based Approach*

*****

I. INTRODUCTION

Today technology is developing day by day So that people have access to a large amount of data on the Internet and television. A number of videos are increasing day by day so it is difficult for the viewers to manually find the video of interest from these large sources of video. Viewers are looking for a video within particular categories. For categorizing a large amount of video data, research work has begun on automatically classifying video.

We mainly focus on reviewing various approaches to video classification. Various features from the video are taken for classifying the video. Video classification algorithm categorized various video by assigning the appropriate label to each video. (e.g. 'News Video', 'Cartoon Video' or 'Sports Video')

The rest of the paper is organized as follow. In section II, describe an approach that uses audio features for video classification. The approach that uses text features is described in section III. Section IV describes the approach that uses visual features. The comparison of various approaches for video classification is described in section V. In last section VI we provide conclusions.

II. GENERAL BACKGROUND

After studying literature survey of video classification, the approaches for video classification could be divided into four groups: Audio-based approaches, Text-based approaches, Visual-based approaches and fourth one are those that used various combination of audio, text and visual features. Features for video classification are drawn from these three different modalities: Audio, Text and Visual.

Generally, in video classification some author classifying the entire video while some author focused on classifying video segment such as identifying 'Sports Video', 'News Video' or 'Cartoon Video' [1]. Another algorithm classifying various sports video such as 'golf', 'Hockey' or 'football' [2].

III. AUDIO BASED APPROACH

Generally, audio clips are shorter in length and of small size. So if audio features need to be stored, it requires less space than other features. Another advantage of an audio feature is that they require less computational resource than visual features.

For generating features from an audio signal, the audio signal is sampled at a certain rate. And then these sampled signal grouped together into frames. According to literature, audio-based approaches are used more than text-based approaches for video classification. Features from audio can be obtained from either the time domain or the frequency domain. Some commonly used low-level audio features are described as follows:

1. Time- Domain Features:

The Root Mean Square (RMS) of a signal energy approximates loudness, which is calculated by taking series of windowed frames of sound and computing the square root of the sum of the squares of the windowed sample values [3]. The signal may be divided into various subbands and for each subbands energy measured separately. Different classes of sound fall into different subbands [4].

In the current frame, Zero Crossing Rate (ZCR) is the total number of sign changes for signal amplitude. Higher frequencies have higher zero crossing rate. For Silence frame generally, the loudness and Zero Crossing Rate are below thresholds. Normally music has less variability of the zero crossing rate.

2. Frequency-Domain Features:

Signal distribution across frequency components is known as the energy distribution. Generally, music has higher brightness than in speech.

Bandwidth is a measurement of the frequency range of a signal [5]. Typically speech has a lower bandwidth than music. Some type of sounds has more narrow frequency range than other sounds.

Mel-Frequency Cepstral Coefficients (MFCC) are generated from the logarithm of the spectral components and then placing them into bins based upon the Mel frequent scale, which is perception-based. This is followed by applying the Discrete Cosine Transform(DCT) [6].

Li Lu, Qingwei Zhao and Yonghong Yan, Kun Liu, in [7] stated a video classification for sports video in which author uses audio content as a feature. In a sports video, there are many different types of audio streams such as announcer's speech, advertisement, music, audience's voice and environment noise. In the proposed method, the author used announcer's speech as a feature to extract the keywords to classify different sports. The author uses a two-pass segmentation approach which uses a metric-based algorithm and adapted 128-mixture Gaussian Mixture Model (GMM) to perform the sports video classification.

IV. TEXT BASED APPROACH

Text-based approach text-only approach is less frequently used approach for video classification from video falls into two different categories. The first is viewable text which could be text on object or text placed on-screen [8] to obtain these text from video Optical Character Recognition (OCR) is used [9].

The second category of text is the transcript of the dialog. These texts are extracted from a speech by using Speech Recognition Method [10] sometimes these texts are provided in the form of closed captioning text or subtitles.

The main advantage of using text-based approach is that the relationship between specific word and the specific category is very easy to understand for humans. Text-based approach can utilize the large body of research on document text classification [11] which is another advantage of using text-based approach.

On-screen text which is obtained from Optical Character Recognition (OCR) has high error rates [12].

W. Zhu, C. Toklu, and S.-P. Liou in [13] stated the weighted voting method for automatic news video story categorization. News video story categorization. News video story categorization is done based on the closed captioned text. Generally, news videos were separated into stories using the separations in the closed captioned text. For classification, a set of keywords were extracted from closed captioned text to form a feature vector. The classification is accomplished by computing the likehood score for each category of the news video story and the database is updated incrementally in linear time. The author used various 425 news stories from CNN for experiment purpose. The author compared the performance of classification with Bayes Decision method and SNoW.

V. VISUAL BASED FEATURES

Normally humans easily understand the information which is based on the vision. So that most of the researchers have used this method for video classification. The video is composed of a set of frames in a sequence. And the shot is a collection of more than one frames taken from a single camera action. Visual features are extracted from these frames or from various shots of the video.

Some researchers have used a combination of these visual features with audio and text features according to requirement.

The Shot is a natural way of segmenting the video therefore many researchers use shot as a visual feature. But there is a problem with a shot based method which is that it is difficult to identifying shot boundaries [14].

The video contains so many frames so if it needs to be store it required large storage so for solving this problem researchers only used key frames. Key frames represent the shots. Some dimensionality reduction technique also used such example is the application of wavelet transform.

Visual-based features for video classification include the following features: Color-based, Shot-based and Object-based. Color is used as a proxy for light levels, the motion used to measure action while length if the shot is used to measure the pace of the video.

1. Color-Based Features:

A video is a set of frames and that video frame is made of pixels. Each pixel is represented by a set of values of various color spaces [15]. For representing a color in a video frame many color spaces exist. Two most popular color spaces are the red-green-blue (RGB) and Hue-saturation-value (HSV) color spaces.

In the red-green-blue (RGB) color space, each pixel in the video frame is represented by a various combination of the individual colors red, green and blue. While in hue-saturation-value (HSV) color space, colors of pixels are represented by hue, saturation, and values. Hue re-present the wavelength of color. Saturation represents the amount of white light present in the color and value represent the brightness and intensity of the color in pixel [16].

A color histogram is used to represent the color distribution in a frame. Which is mostly used to compare two video frames. Color histograms of video frames represent the number of pixel in a frame for each possible color.

There are two disadvantages of it. For particular color, we cannot find an exact pixel. The images which are represented in a frame may have been taken under different lighting conditions. So for comparison, it does not give perfect result hence various preprocessing on frames is required.

J. Hanna, F. Patlar, A. Akbulut, E. Mendi and C.Bayrak in [17] classify the sports video into three pre-defined categories: golf, Hockey, and Football. The author used color features to classify the sports video. Each frame of video is composed of pixels of different color. Color data from each pixel in RGB color space is gathered and averaged for each frame. Using red, green and blue saturation the speed of color change is calculated by subtracting each color's saturation from the saturation of the previous frame. Hidden Markov Model (HMM) is used for classification of sports video in this method. The speed of color change is used as an observation sequence in HMM for classifying the sports video.

2. Shot- Based Features:

For using these type of features for video classification, shots from video must be detected first. There are many types of transitions from one shot to the next shot. Which falls into following three categories: Hard cuts, Fades (fade-out and fade-in) and Dissolves.

In hard cuts, one shot abruptly stops and another shot begins [18]. A shot that gradually fading out of existence to a monochrome frame which is known as fade-out while a shot that gradually fades into existence from a monochrome frame is known as fade-in. one shot fading-out while another shot fades-in is known as a dissolve. For correctly identifying the shot changes, it is necessary to understand which types of shot transition occur in a video. For categorization, the shot transition type is very important feature [19].

3. Object-Based Features:

To use these type of feature, first objects are identified and then features are detected from video frames. But it is difficult to detect and identify the object from video frames so this type of approach is not most commonly used.

In this type of feature, first specific type of object is identified for example faces [20] [21] in the video frame and then features like dominant color, texture, size, and trajectory are derived from them.

VI. COMBINATION APPROACH

There are some disadvantages of using each approach individually so that many authors make use of combination is audio, text and visual features. The final feature vector is created by combining all features from three different modalities: audio, text and visual contents which are used as an input to classifiers for the video classification. While sometimes from each modality classifiers are trained then another classifier is used for making the categorization of video [22].

W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang in [23] uses Support Vector Machine for classifying the stream of news video into types of news stories. For detecting the video shots audio and visual features are used. The closed captions and any scene text detected using Optical Character Recognition (OCR).

K. Subashini, S. Palanivel, and V. Ramaligam in [24] classify the audio and video data into one of the pre-defined categories which are sports, news, advertisement, movies and serial. In this classification system author combine the audio and video features for video classification. Color histogram is used as visual features and Mel Frequency Capstral Coefficient is used as audio features. For classification Support Vector Machine (SVM) is used for both segmentation and classification.

VII. COMPARISION OF FEATURES

Classification Approach	**Comparison**	
	Features	***Advantages / Disadvantages***
Audio-Based Approach	Audio	It is difficult to detect multiple sounds at the same time. Audio clips are typically shorter in length and of small size. So it requires less computational resource.
Text-Based Approach	Close Captions OCR Text Speech Recognition	High Dimensionality. It is cheap to computationally extract. Computationally expensive High error rate
Visual-Based Approach	Color-based Features Shot-based Features Object-based Features	Easy to implement It is difficult to identify shot, it may not give the accurate result. It is computationally expensive. Limited to detect the number of objects.

VIII. VARIOUS CLASSIFIERS

In the literature review of video classification, we found that the there are many standard classifiers for classifying the video such as Bayesian, Support Vector Machine [20] and

Neural Networks. However Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) are two popular methods for classification.

Gaussian distribution is a probabilistic approach but it does not always model data well. So, as a solution to this problem using a linear combination of Gaussian distributions, which is known as a Gaussian Mixture Model [25]. Gaussian Mixture Model has been used for finding complex probability distribution as well clustering.

Another model for classification is the Hidden Markov Model (HMM) which is a probabilistic sequence model used for classification of sequential data. HMM is used to predict the sequence of the state changes, based on the sequence of the observation.

A video is a collection of features and there is a temporal relationship between these features. Many author use HMM to capture this temporal relationship.

A hidden Markov Model is a statistical model which builds upon the concept of a Markov Chain. Markov Chains, named after Andrey Markov, are the mathematical system that hops from one state to another. HMM represents a set of state and the probabilities of making a transition from one state to another state [26].

The job of this sequence model is to assign a label or class to each unit in a sequence, this mapping a sequence of observation to a sequence of labels. The sequence of units is given, it computes a probability distribution over the possible sequence of labels and chooses best labels sequence as output.

IX. CONCLUSION

We have reviewed the literature of video classification and from that, we observed that a large variety of approaches exists for video classification such as Audio-based approach, Text-based approach, and Visual-based approach. Each approach contains various features which we have reviewed in this paper. Researchers have used different approaches individually for video classification. Also, many of the researchers have used various combination of these approaches according to application requirement which gives better classification accuracy than individual approach.

A lot of research has been done in this area but still, it is an emerging area in terms of performance evaluation and resource utilization.

REFERENCES

[1] Narra. Dhana Lakshmi, Y.Madhavee Latha, A.Damodaram and K.Lakshmi Prasanna, “Implementation of Content Based Video Classification using Hidden Markov Model” in 7th International Advance Computing Conference, IEEE-2017

[2] Josh Hanna, Fatma Patlar, Akhan Akbulut, Engin Mendi And Coskun Bayrak, “HMM Based Classification of Sports Videos Using Color Feature ”, in 6th IEEE International Conference, IEEE-2012

[3] 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.K. Elissa, “Title of paper if known,” unpublished.

[4] P. Q. Dinh, C. Dorai, and S. Venkatesh, “Video genre categorization usingaudio wavelet coefficients,” in Fifth Asian Conference on Computer Vision, 2002.

[5] G. Lu, “Indexing and retrieval of audio: A survey,” Multimedia ToolsApplications, vol. 15, no. 3, pp. 269–290, 2001.

[6] B. Logan, “Mel frequency cepstral coefficients for music modeling,” inInternational Symposium on Music Information Retrieval, 2000.

[7] Li Lu, Qingwei Zhao and Yonghong Yan, Kun Liu, “A Study on Sports Video Classification Based on Audio Analysis and SpeechRecognition”, in 2010 International Conference on Audio, Language and Image Processing, IEEE-2010.

[8] V. Kobla, D. DeMenthon, and D. Doermann, “Identifying sports videosusing replay, text, and camera motion features,” in SPIE conference on Storage and Retrieval for Media Databases, 2000.

[9] 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.

[10] P. Wang, R. Cai, and S.-Q. Yang, “A hybrid approach to news video classificationmultimodal 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.

[11] F. Sebastiani, “Machine learning in automated text categorization,” ACMComputing Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[12] A. G. Hauptmann, R. Jin, and T. D. Ng, “Multi-modal informationretrieval 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.

[13] W. Zhu, C. Toklu, and S.-P. Liou, “Automatic news video segmentation and categorization based on closed-captioned text,” in IEEEInternational Conference on Multimedia and Expo (ICME 2001), 2001, pp. 829–832

[14] 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.

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

[16] A. D. Bimbo, Visual Information Retrieval. San Francisco, CA: MorganKaufman, 1999

[17] Josh Hanna, Fatma Patlar, Akhan Akbulut, Engin Mendi And Coskun Bayrak, “HMM Based Classification of Sports Videos Using Color Feature ”, in 6th IEEE International Conference, IEEE-2012

[18] Y. Abdeljaoued, T. Ebrahimi, C. Christopoulos, and I. M. Ivars, “Anew algorithm for shot boundary detection,” in Proceedings of the 10th European Signal Processing Conference, 2000, pp. 151–154.

[19] G. Wei, L. Agnihotri, and N. Dimitrova, "TV program classificationbased on face and text processing," in IEEE International Conference on Multimedia and Expo, vol. 3, 2000, pp. 1345–1348.

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

[21] P. Wang, R. Cai, and S.-Q. Yang, "A hybrid approach to news video classificationmultimodal 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.

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

[23] W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, "Integrating visual,audio and text analysis for news video," in Seventh IEEE InternationaConference on Image Processing (ICIP 2000), 2001

[24] K. Subashini, S. Palanivel, and V. Ramaligam. Audiovideo based segmentation and classification using SVM. In Third InternationalConference on Computing Communication Networking Technologies (ICCCNT), pages 1–6, July 2012

[25] C. M. Bishop, Pattern Recognition and Machine Learning. New York,NY: Springer, 2006.

[26] M. Kalaiselvi Geetha and S. Palanivel. "HMM Based Automatic Video Classification Using Static and Dynamic Features", in International Conference on Computational Intelligence and Multimedia Applications, IEEE-2007