ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

A Survey on: Content Based Video Classification

Hansa Shimpi[1], Sudeep Thepade[2]

P.G. Student, Department of Computer Engineering , Pimpri Chinchwad College of Engineering, Nigdi, Pune, India

Professor, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Nigdi, Pune, India

ABSTRACT: As we know that the today's world is the Internet world, the amount of the generation of multimedia data is very vast. The sheer size of video data makes it impossible for the human to classify it manually into different classes so the user gets appropriate result. With such large growth of video data, Content-based video classification comes into the picture. Content based video classification and retrieval have attracted more and more focused in last decade. Although successful results were acquired in recent year, automaticanalysing the semantic content of the video is still very challenging at the current state-of-the-art. In order to map the low-level feature to high-level semantic content, many efforts are lead to the semantic indexing and modeling of video content through semi-automatic approach. In this paper, some recent advances in content-based video classification and retrieval are reviewed.

KEYWORDS: KeyFrame, Feature Extraction, Video Classification.

I.INTRODUCTION

Due to significant improvement in processing technologies, network subsystems, and availability of large storage systems, a large quantity of video data is being generated all over the world. This videos may come from surveillance cameras, cell phones, movies or animation. With large databases containing video data becoming even larger. 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 a video of interest. One method that viewers use to narrow their choices is to look for video within specific categories or genre. Because of the huge amount of video to categorize, research has begun on automatically classifying video.

That automated methods of classifying video are an important and active area of research is demonstrated by the existence of the TRECVid video retrieval benchmarking evaluation campaign [1]. TRECVid provides data sets and common tasks that allow researchers to compare their methodologies under similar conditions. While much of TRECVid is devoted to video information retrieval, video classification tasks exist as well such as identifying clips containing faces or on-screen text, distinguishing between clips representing outdoor or indoor scenes, or identifying clips with speech or instrumental sound [2].

In this review, we focus on approaches to video classification and distinguish this from video indexing. The choices of features and approaches taken for video classification are similar to those in the video indexing field. The main difference is that in video indexing all features of videos placed into the database and when the feature of query video matched with feature database, it retrieves the video accurately [3]. In contrast, video classification algorithms place all videos into categories, typically with a meaningful label associated with each (e.g., 'news video' or 'cartoon video').

The various number of approaches have been attempted for the accomplishment of automatic classification of the video. After rigorous literature survey, we found that these approaches could be divided into four groups: text-based approaches, audio-based approaches, visual-based approaches, and those that used some combination of text, audio, and visual content.

II.BASIC CONCEPT OF VIDEO CLASSIFICATION

In the process of content based video classification, contents of a video can be analyzed by extracting features from the key frames of that video. Various feature extraction techniques can be used to represent an image or a key frame in the form of a feature vector. These feature extraction techniques work in spatial or transform domains. The extracted features are then pre-processed using various data pre-processing techniques popularly used with numeric data. Data pre-processing techniques are used to reduce redundancies and inconsistencies in data that is to be used for classification.

A. Key frame extraction

We know that videos are processed in the form of frame. Once the frames are separated it is obvious that there are multiple frames with redundant information. To process every frame. it need a very complex algorithm and also require large memory space. To use available memory efficiently, it is necessary to select frame which give lot of information about the video. Such a task of selecting a particular frame is done by the key frame extraction technique. Such a technique remove the frame which have redundant information. .By removing these frames the number of frames to process is reduced thus improves the processing time of the system. The most common method to calculate the frame difference is Histogram based method. The histogram difference of two consecutive frames Ii and Ij is calculated as [4]

$$\text{diff}\ (Ii, Ij) = (Hi - Hj)\ 2\ /\ (Hi + Hj) \qquad (1)$$

Where Hi and Hj are Histograms of Ii and Ij.

The mean and variance of all the pixels in the frame Ii is taken as the reference frame and is compared with that of frame Ij. When the variance is lesser than the threshold, the frame Ij is

rejected. Similarly, when the variance of the frame Ij is greater than the threshold, then Ij becomes the new reference frame. This step is repeated to all frames until all redundant frames are removed. The reference frames stored are known as the key frames.

B. Feature extraction

In image processing, feature extraction starts from an initial set of measured data and forms resulting features suggested to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction. When the input data to an algorithm is too large to be processed and it is supposed to be redundant (e.g. the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features. Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data. [5] After reviewing the literature of methods, we found that there are various features (e.g. Text, audio, visual, color, edge, motion) are extracted to reduce the data of video for classification purpose.

C. Classification

The Video Classification model is a training-testing model in which the classifier is initially trained on the basis of features extracted from training videos and then testing is performed. Various classifiers like Support Vector Machine (SVM), NAÏVE BAYESIAN, Neural Network (NN), HMM, CNN, J48, oneR, KNN are used for the classification purpose .According to literature survey it is find that the SVM and HMM classifier gives better accuracy.

III. APPROACHES TO VIDEO CLASSIFICATION

For performing automatic classification of video, various approaches have been attempted. After review process of methods we can categorize this approaches into four groups: text-based approaches, audio-based approaches, visual-based approaches and combination of text, audio & visual.The following Fig.1.shows the hierarchy of video classification.

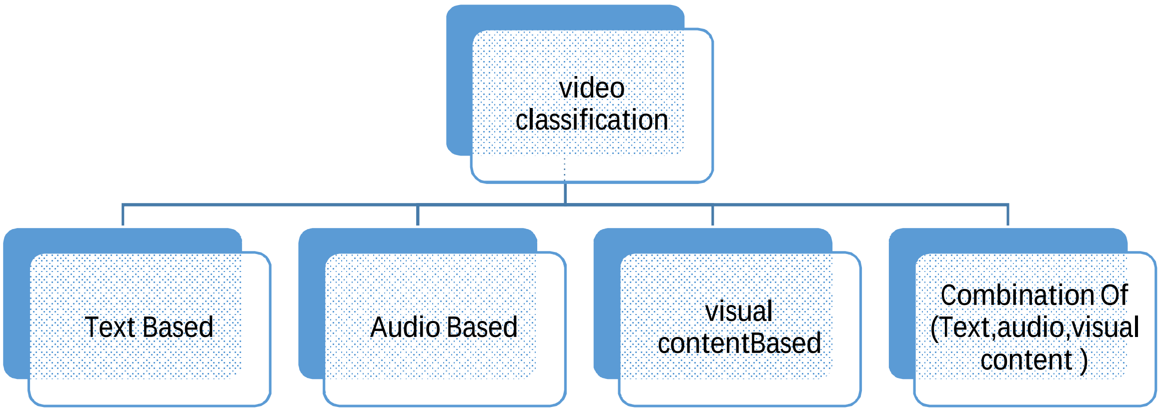

Fig.1. Approaches of Video Classification

A. *Text Based Approach*

The text based approach is not that much popular in the video classification. The text which is extracted from video divided into the two category.

The first category is the viewable text. Viewable text could be the text on the object that are filmed (scene text), like participant's name on jersey or the name on a building or it could be the text place at the bottom of the display (graphic text) like the score for sports event. With the help of optical character recognition (OCR), the text features are generated from this viewable text by identifying text objects.

The second category is the closed captioning (CC). CC is the processes of displaying text on a television, video screen, or other visual display to provide additional or interpretive information. The main purpose of CC is to provide transcription of the audio portion of a program as it occurs. The term "closed" indicates that the captions are not visible until activated by the viewer, usually via the remote control or menu option. Closed captioning was first demonstrated at the First National Conference on Television for the Hearing Impaired in Nashville.

In [6] the weighted voting method was used for automatic news video story categorization based on the closed captioned text. Initially News videos were segmented into stories using the separations in the closed captioned text, then a set of keywords were extracted to form a feature vector for further processing. The classification is accomplished by computing the likelihood score for each category and the data base is updated incrementally in linear time. The author had used the proposed method to classify 425 news stories from CNN and compared the classifying performance with SNoW and Bayes decision method.

The third category is Open Captioning (OC). It is also referred as Subtitles .The open captioning serve the same purpose of closed captioning but the main difference in CC and OC is that in OC text is a part of the video and it is extracted using text extraction methods and OCR.

In [7], a new method were propose to classify video content by analyzing the corresponding subtitles based on the WordNet lexical database and the WordNet domains .The subtitles were segmented into sentences by using Mark Hepple's POS tagger to picked the correct meaning of each word in WordNet.TextRank algorithm were used for the keyword extraction from the subtitles. The author also applied a Word Sense Disambiguation (WSD) method to assign a sense to each word in the text. After identifying the correct sense of each extracted keyword, the author made use of the WordNet domains to derive the domains which these sense correspond to. For picking appropriate class label for each video, they defines mappings between WordNet domains and category labels and finally to assign the class label to video they compared WordNet Domain with the extracted WordNet keywords.

The main advantage of text-based approaches is that they can utilize the whole body of document for text classification. Another advantage is that it is very easy to understand the relationship between the features (i.e., words) and specific genre for human.

On the other side, there are some disadvantages of transcript text which is used for video classification. One is that transcript is just the text of dialog. It does not describe the content of video which is being seen. Second is that not all video have closed caption or subtitle.

B. *Audio Based Approach*

According to literature survey, it comes to know that the audio based approach is widely used than the text based approach for the classification of video. One of the main advantage of this approach is that it requires very less computational resources than the visual method. And if we want to store the feature for the further process, audio feature require very less space for that. The audio feature can be extracted from the time domain as well as the frequency domain.

In [8], the author proposes a new method which uses audio content to classify the sports. The audio stream of sports contain of various parts like announcer's speech, advertisement, music, audience voice and environment noise. The author were use announcer's speech to extract the keywords which are used as features to distinguish different sports using a two-pass segmentation approach which uses a metric-based algorithm to do segmentation followed by a model-based classification to extract the speech segments. The author adopted 128-mixture GMM to perform the classification.

In [9], the author adopted a wavelet transform-based analysis of audio tracks accompanying videos for the problem of automatic program genre detection. And then compare the result of classification with the conventional features derived from Fourier and time analysis for the task of discriminating TV programs such as news, commercials, music shows, concerts, motor racing games, and animated cartoons. The author studied three different classifiers namely the Decision Trees, SVMs, and k-Nearest Neighbors to analyze the performance of wavelet features which are used for the purpose of classification.

C. *Visual Based Approach*

After reviewing the literature of methods, we found that most of the approaches of video classification are based on visual element either alone or with combination text and audio.Generally, videos are structured according to a descending hierarchy of video clips, scenes, shots, and frames as shown in Fig. 2

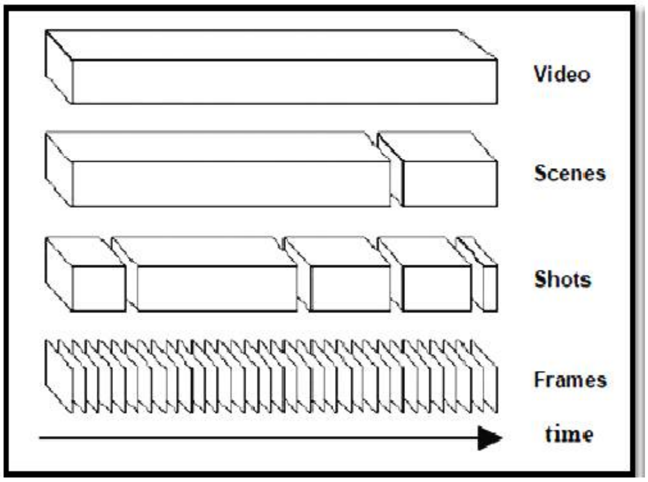

Fig. 2 Video Structure.

A video is a collection of images known as frames. All of the frames within a single camera action are called a shot. A scene is one or more shots that form a semantic unit. Video structure investigation aims at segmenting a video into a number of structural elements that have linguistic contents, including shot boundary detection, key frame extraction, and scene segmentation. Most of the approaches extract the feature from frame or the shot since shot is a natural way to segment a video and each of these segments may represent a higher-level concept to humans like “action scene”, “Car falling in a valley”. One major disadvantage of using shot-based methods is that the methods for automatically identifying shot boundaries don’t always perform well [10]. Also the feature which are extracted from frame consist of more raw data which affect the space. To avoid such a raw information, keyframe extraction technique is used which results into the reduction of data.

Various visual feature such as color, motion, and object are used for the purpose of classification of video into category.

*1) Color based feature:*A video is composed of frames and each frame again composed of number of pixel. Each pixel of frame is represented by the various color spaces. Among all, the RGB and the HSV color spaces are very popular. In RGB color space, the color of pixel is represented by the combination of red, green and blue color .In HSV color space, the color of pixel is represented by hue, saturation and value i.e. intensity of the color. The overall distribution of tonal in the frame of video is generally given by the histogram. It plots the number of pixels for each tonal value. Histogram difference is used for the comparison of two frame but if the frames are captured with different lightning condition, it may be affect the comparison result.

In [11], the author uses color feature to classify the sport video. Each pixel of frame of video is composed of color. Color data from each pixel in RGB color space is gathered and averaged for each frame. The author make the use of red .green, blue saturation for calculation speed of color change by subtracting each color’s saturation from the saturation of the previous frame. This speed of color change is then used as an observation sequence in HMM for the classification of sports namely golf, hockey and football.

In [12], the author uses Hidden Marlow Model for the classification of video into cartoon, news, sports, and unknown category. To reduce the number of frame per video author chooses keyframe of video and then apply color histogram and edge histogram to extract the features from keyframe. The label sequences from color feature space and edge feature space are concatenated to form a final symbol sequence which is used as input for the HMM classifier.

*2) Motion based feature:*Motion is one of the important attribute in video. Every video has its different pattern of motion. There are two types of source which causes motion in video namely camera and objects.

In [13] the author adopt the motion feature along with color and shape feature to classify Five popular TV broadcast genre namely cartoon, commercials, cricket, football and tennis. They use a simple and effective method where motion is extracted by pixel-wise differencing of consecutive Frames .Finally SVM classifier is applied to the set of features which are extracted using color, shape and motion.
In [14],The author proved that the motion feature in video sequencing give better accuracy to classify video as either cartoon or non-cartoon. The visual feature chosen for classification is the motion of foreground objects, which are detected using pixel based frame differencing. The dimensionality of this signal is reduced by applying a DCT. The authors also investigated how many DCT coefficients to use for constructing the feature vector and found the best results occurred by keeping 4–8 coefficients. Classification is performed using a GMM.

*3) Object based feature:*Object based features are rarely used for the purpose of video classification since it require lots of computation in detecting and identifying the object. Various approaches uses Face as an object for further processing.

In [15], the author uses skin tone extraction method. They manually label the skin ton pixel in large image dataset and generate the distribution graph of skin-tone pixels in YIQ color coordinate. By applying Morphological operations to skin-tone regions, thin protrusions and small isolated regions are removed and also break the narrow bridges. Next the shape analysis is perform on the given data to detect the face. For each detected face the mean and standard deviation in color, height, and width and center position are computed. All these features constitute a face model. The face model is used for tracking faces in the future frames until the next cut is detected. Finally feature from face and text are concatenate and give as an observation sequences to train HMM's and evaluate the probabilities of the given clip being one of the four categories of TV programs namely commercial, news, sitcom and soap

D. *Combination of Text, audio And Visual Based Approach*

Many author make a use of combination of text, audio and visual to avoid the cons of using each approach alone. The difficulty in this approach is how to form the feature vector from these three different area. From literature survey, it is found that some author train each e feature vector separately and make use of another classifier on them to predict the category of test video and some author combine all features from text, audio and visual content to form the final feature vector which is used as a input to classifier for the classification of video. [16]

In [17], the author classify a stream of news video into types of news stories. Audio and visual features are first used to detect video shots and then these shots are grouped into scenes if necessary. The closed captions and any scene text detected using optical character recognition (OCR) are the features used by a support vector machine for classifying the news stories.

In [18], the author combine the audio and video feature for segmentation and classification. The classification system classify the audio and video data into one of the predefined categories such as news, advertisement, sports, serial and movies. Mel frequency cepstral coefficients is used as acoustic features and color histogram is used as visual features for segmentation and classification. Support vector machine (SVM) is used for both segmentation and classification.

IV. COMPARATIVE STUDY OF FEATURES

The following table shows the comparative study of feature with their advantages and disadvantages

TABLE I
Feature Comparison

Feature Type		Advantages	Disadvantages
Text Feature		Full Document is used for classification . Easy to understand the relationship between the features (i.e., words) and specific genre for human.	Transcript only contain text of dialog. Not all video have closed caption or subtitle
Audio Feature		Require fewer computational resources than visual features, clips are typically shorter in length and smaller in file size than video clips,	difficult to distinguish between multiple sounds
Visual Feature	**Color Feature**	Easy to implement and process	Crude representation
	Motion Feature	Motion is important attribute of video	Difficult to distinguish between types of motion, computational requirements range from low (MPEG motion vectors, frame differencing) to high (optical flow)
	Object Feature	-	Difficult, limited on number of objects, computationally expensive

V. CONCLUSION

We have reviewed the video classification literature and found that a large variety of approaches have been explored. Features are drawn from three modalities: text, audio, and visual. Most of the literature describes approaches that uses features from a single modality. Also some the literature describes approaches that uses features from all three modality. Various classifier namely SVM, NAÏVE BAYESIAN, NN, HMM, CNN, J48, oneR, KNN are used for classification of video into particular category. Among these classifier HMM and SVM gives nearby 92% accuracy of classification.

REFERENCES

1. A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval. New York, NY, USA: ACM Press, 2006, pp. 321–330.
2. 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
3. S. W. Smoliar and H. Zhang, "Content based video indexing and retrieval," IEEE Multimedia, vol. 1, no. 2, pp. 62–72, 1994.
4. Shaoshuai Lei, Gang Xie, and Gaowei Yan, "A Novel Key- Frame Extraction Approach for Both Video Summary and Video Index", The Scientific World Journal Volume 2014 (2014),Article ID 695168, pp 9-18.
5. https://en.wikipedia.org/wiki/Feature_extraction
6. W. Zhu, C. Toklu, and S.-P. Liou, "Automatic news video segmentation and categorization based on closed-captioned text," in IEEE International Conference on Multimedia and Expo (ICME 2001), 2001, pp. 829–832
7. P. Katsiouli , V. Tsetsos and S. Hadjiefthymiades, Semantic video classification based on subtitles domain terminologies, Proceedings of SAMT Workshop on Knowledge Acquisition from Multimedia Content (KAMC)(Genoa, Italy, 2007).
8. Li Lu, Qingwei Zhao and Yonghong Yan, Kun Liu, "A Study on Sports Video Classification Based on Audio Analysis and Speech Recognition", 978-1-4244-5858-5/10/$26.00 ©2010 IEEE
9. P. Q. Dinh, C. Dorai, and S. Venkatesh, "Video genre categorization using audio wavelet coefficients," in Fifth Asian Conference on Computer Vision, 2002.
10. 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.
11. Hanna J," HMM Based Classification of Sports Videos Using Color Feature", Intelligent Systems (IS), 2012 6th IEEE International Conference, pp . 388-390, Date 6-8 September 2012
12. Narra Dhana Laxmi, Y.MADHAVEE LATHA, A.DAMODARAM, "Implementation of Content Based Video Classification using Hidden Markov Model", 2017 IEEE 7th International Advance Computing Conference
13. Suresh, V., Mohan, K.C., Swamy, K.R., Yegnanarayana, B.: Content-based Video Classification Using Support Vector Machines. In ICONIP-04, Calcutta, India (2004) 726- 731
14. M. Roach, J. S. Mason, and M. Pawlewski, "Motion-based classification of cartoons," in Proceedings of the International Symposium on Intelligent Multimedia, 2001, pp. 146–149.
15. N. Dimitrova, L. Agnihotri, and G. Wei, "Video classification based on HMM using text and faces," in European Signal Processing Conference (EUSIPCO2000), 2000
16. D. Brezeale and D. J. Cook, "Automatic video classification: a survey of the literature," IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 38, no. 3, 2008
17. 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 Internationa Conference on Image Processing (ICIP 2000), 2001
18. K. Subashini, S. Palanivel, and V. Ramaligam. Audiovideo based segmentation and classification using SVM. In Third International Conference on Computing Communication Networking Technologies (ICCCNT), pages 1–6, July 2012