Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 496 – 502

Procedia
Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Reducing Semantic Gap in Video Retrieval with Fusion: A survey

D.Sudha [a], J.Priyadarshini [b]*

[a]*School of Computing Science and Engineering, VIT University, Chennai, India.*
[b]*School of Computing Science and Engineering, VIT University, Chennai, India.*

Abstract

Abstract: Multimedia provides a rich content of information and huge amount of data's are available in the field of video retrieval. Now day's enormous videos are available on web and online to accessible from internet or retrieve videos from smart phones, digital cellular assistants. There is the drastic growth in the amount of multimedia field of improving data storage, acquisition and communication technologies, which are all supported by major improvements in processing of video and audio. Researches focused on more efforts in video retrieval that contain certain visual information rather than image of their interest. Such a search is facilitated by Content Based Video Retrieval (CBVR) methods. Specifically segmentation of video is the most prominent step as the retrieved results are based on the segmentation boundaries. The shot boundary detection can be performed using various different techniques like Motion/hybrid DCT, edge tracking, histogram, HSV Model, Motion vector and Block matching methods. This paper mainly presents a study of different methods/algorithm that has been proposed in literature for video retrieval to reduce the semantic gap between low and high level features. Semantic gap between these two feature level is improving by its efficiency with the help of advanced algorithms and techniques using machine learning with fusions.

Keywords: Shot Boundary Detection; CBVR; Segmentation; Semantic Gap; Key Frame Selection.

1. INTRODUCTION

The basic multimedia information is required for dynamic video indexing and retrieval of video from media warehouse. There have been a drastic challenges faced in the field of multimedia, researcher concentrate in

* Corresponding author. Tel.: 7373832002.
E-mail address: dsudha.89@gmail.com

doi:10.1016/j.procs.2015.04.020

multimedia video retrieval from media or database to end-users. Among several multimedia resources, video is a key component which comprises mainly upon four parts. The first one is that the vigorous video provides the rich contents than that of text as well as images. The second is a huge amount of raw data in video retrieval field. Next, splitting entire video retrieval process to scenes, shots and frames. Last process is to segment the video into small units which includes shot boundary detection, Key Frame Extraction, Scene segmentation and Audio Extraction. The segmented objects in each image frame which have been used in many applications, such as Object manipulation, Surveillance, Scene composition and Video retrieval.

In this paper, mainly focused on Information Based Video Retrieval (IBVR) system which includes various processes with several algorithms for different techniques to extract feature in order to reduce the semantic gap. Initially video is segmented into shots for shot boundary detection, while the next process it to perform key frame selection which is to selects the key frame for representing the shot using Euclidean Distance algorithm. After selecting the key frame, feature extraction had been processed and stored into the feature vector. Generally, features are of spatial as well as temporal in nature. [11] A spatial feature includes color, shape, and edge. Similarly, temporal features are mainly classified into two, motion and the other is audio. Finally [13][14] indexing process to be done with the help of clustering tree algorithm , B+ tree algorithm. Indexing is to index the key frames with the help of retrieving process easily following video retrieval process for similarity matching operations with the user query using Euclidean Distance algorithm and final stage is to displays the result to end-users.

2. RELATED WORK

The need for video retrieval information from cloud has captured more and more attention with two heterogeneous objects. Research efforts have led to the development of methods and tools that provides accessing to both images and video data. Nevertheless, several techniques roots are tuned from computer vision, Surveillance and pattern recognition fields. The most recent methods are used to find the similarities in visual information content are only extracted from low level features. Finally, these features are then clustered for generation of database indices. The following section describes in detail that literature survey on the use of these pattern recognition methods which enable to retrieve from both image and video content.

Video is manipulated / created by taking a set of shots and composing them together. Extracting structure is the purpose of segmenting video that involves the detecting of boundaries between scenes and shots. Specifically, video includes shots, scenes and frames. [12]The multiple clients are communicated through network to server in various processes as segmentation, key frame selection, feature extraction, classification and clustering, Indexing and matching similarity in data warehouse for retrieval in video for challenging the semantic gap between low level and high level features.

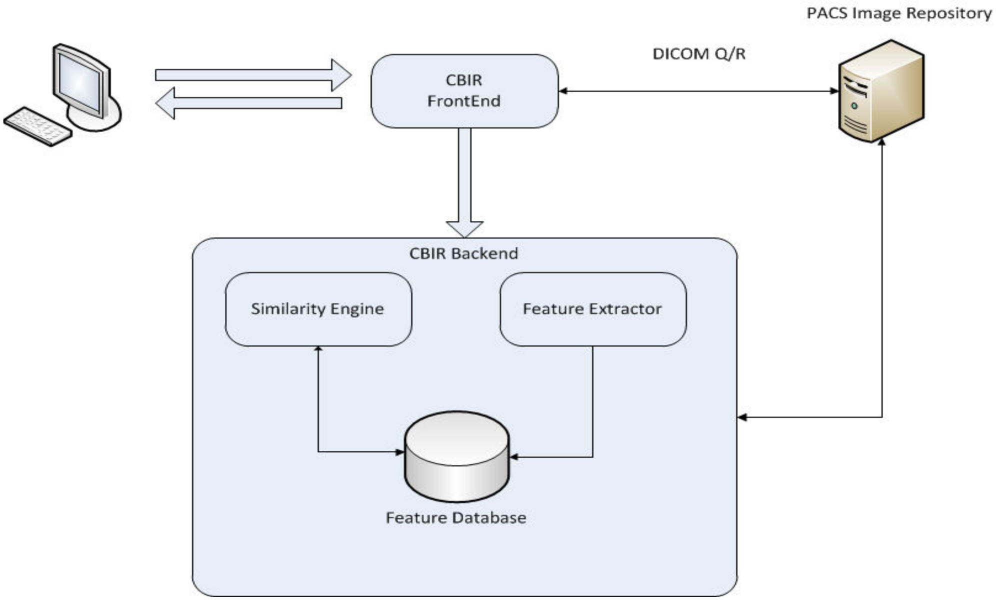


Fig 1. Basic Flow of system for retrieval

The above figure shows that the flow of system; in this system multiple clients are interacting with the server through the network . The server contains the segmentation step process into scenes, then converted into shots which in turn finally converted into frames. [9]Next key frame selection is to selects the most appropriate key frame among the extracted frames of entire video using Euclidean Distance algorithm. After feature extraction, [8]Support Vector Machine applies to learn a framework where high- level video index is visualized through the step of features. Finally matching similarity process to be done for the retrieval of video according to the user query.

Oscar D.Robles et.al. [2] are presented the two new methods for representing the content of a video in order to be used in content based video retrieval system. The proposed techniques are "Towards A content based video retrieval system using Wavelet Based Signature and compute first a multi-resolution representation using Haar Transform .In order to extract the two types of signatures mainly based on global and local multi-resolution color histogram.

"A semantic video retrieval approach using audio analysis" is presented by Lew[4][6] in which the audio can be automatically categorized into expressions , music, speech etc., Most of the researcher have been focussed on the semantic gap between the local and global features. In their research literature, significant attention has been given to the visual aspects of video, however relatively a little amount of work was directly focussed on audio content for video retrieval.

Visser.et.al.[5] are presented the "recognition objects in video sequences" for segmenting blobs from video by using Kalman Filter as well as classify the blobs using probability random test and apply several different temporal methods , has been led to the sequential mode of classification over the video sequences.

Zhang.et.al.[1] proposed "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback" for ranking to the results. It is to improve the retrieved results in the next future rounds. Relevance Feedback is an example of exploiting the fact that MMIR systems are smart search tools operated by human users. Video search tools engines can be the killer application in this field. Furthermore, real-time interactive mobile technologies are evolving introducing new ways for people to interact based on the ranking method.

J.C.Haartsen et.al[3][7] presented " To adaptive display for enhancing audio playback" in the field of audio retrieval by metadata coined as the various different terms as artist, song, title, album etc., Researchers have also used Annotation- based approaches for audio retrieval.

3. ARCHITECTURE DIAGRAM

The following figure shows that the architecture block diagram for video retrieval with scenes, shots as well as frames. Majorly, it divides into two separate processes together that is off-line processing and the other ne is on-line processing. First, one is to upload the video/clips data by the administrator and gives it to the media descriptors. Media descriptors works for feature extraction and pass it to the search engine .Finally indexing is for retrieving the video from multimedia data or data warehouse.

3.1. OFF-LINE PROCESSING

In this block the administrator uploads the various video clips/data and forward into the media descriptor. It performs the feature extraction based on the given video and then key frame is chosen from the available frame and pass over to the search engine .Next, Indexing is done on that key frame and these indexes are stored on data warehouse along with the indexes and various other features also gets stored on data warehouse .

3.2 ON-LINE PROCESSING

In this block user takes a one short video clip and forwarded to the user query media then it moves to media descriptor. Next, media descriptors performs the feature extraction based on the given video then these features are given to the search engine, it takes action of requesting to data warehouse taking the various features and similarity matching operations are done with the features of requested video with already stored video from data warehouse and final matched results is given to the end-users.

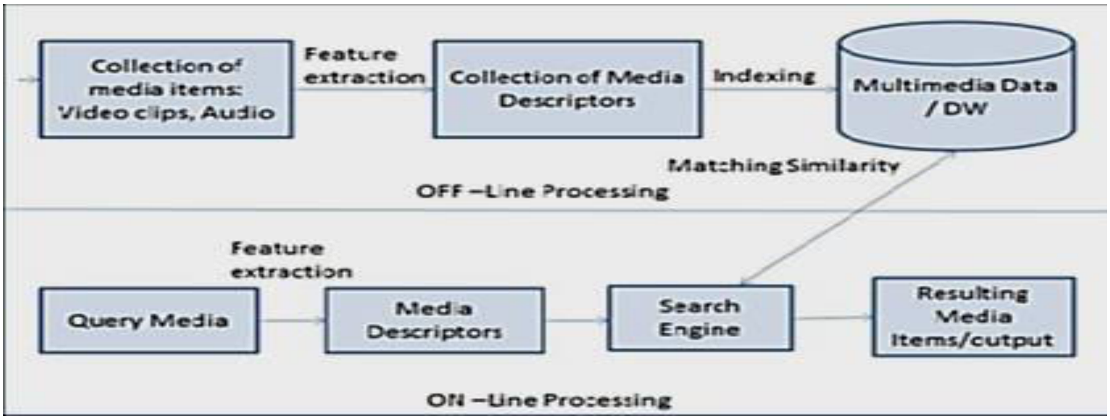


Fig 2.Architectural Diagram

4. PROPOSED WORK

The proposed work describes that analysis of video retrieval, which have semantic gap between low and high-level features, segmentation, boundary detection of shots, feature extraction, Key frame selection, Classification and clustering, Indexing and matching similarity as follows.

4.1. Video analysis

Video data are rich sources of information and in order to analyse the video, the information content of the entire data has to be analysed and thoroughly progressive up-to-date knowledge in video processing. Video analysis is further distributed into two stages; one is to divide the video sequence into a group of shots (shot boundary detection) while the second stage is the process of selecting key frame(s) to represent each slot.

4.1.2. Video shot boundary detection

A shot is a series of frames taken by using one camera. A frame at a boundary point of a shot differs in background as well as content from its successive frames that belongs to the next shot. There is a drastic change in boundary point of two frames because of switching from one camera to another, and this is the basic principle that most algorithms for detecting scene changes depend upon. Specifically, there are two trends to segment video data as shown in figure 3. The first one is uncompressed domain that is broadly classified into two categories, template matching and histogram based. While, the other one is compressed domain that classified into three kinds as discrete cosine transform, motion vectors, hybrid motion/DCT.

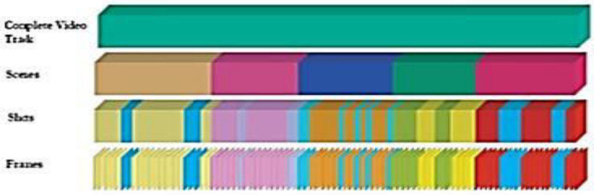


Fig 3. Video Segmentation

4.1.3 Video segmentation and key frame selection

Video segmentation or shot boundary detection involves identifying the frames changes occur from one shot to another. This abrupt changes occurs when the absolute difference of mean blocks between two consecutive frames exceed a threshold value. Changes occurred between two frames, which are called a cut or a break. A complete video is entirely divided into shots for which key frames need to be identified. After segmentation, key frame selection would be obviously compared to entire frame based on the similarity measure, with least difference frames has to be selected for the process .It needs extensive computations and it is not practical for most applications. While on the other hand, choosing the first frame seems to be the natural choice, as all the rest of the frames in the scenes can be considered, as being logical and continuous extensions of the first frame, but it may not be the best match for all the frames in the scenes

4.2 Feature Extraction

It is most difficult to step or process in video retrieval to describe the video frame with the minimum number of media descriptors. This includes the low-level features as color, shape and texture and high-level features as object annotation.

4.2.1. Low Level Features

In video, only low level features is still lagging to gain better frame content for retrieval the content from data warehouse or media repository. Most of the research efforts, to focus on the semantic gap between the high level and low level features have to be united as multimodality in upcoming future trends in video retrieval. The big advantage over here is multimodality can promotes a better video retrieval content than low level features

4.2.2 Color feature

The proposed method for color feature extraction is a histogram- based. The main idea from this method is if two frames having an unchanging background as well as objects will show only little differences in their respective histograms. Let Hi(j) denotes the histogram value for the ith frame, where j is one of the possible grey levels. The frame difference formula is computes as below:

$$HDi = \sum Gj=1 \ / Hi(j) - Hi +1(j) \qquad [Eq1]$$

Where
G is the number of grey levels,
j is the grey value,
i is the frame number,
and H(j) is the value of the histogram for the grey value level j.

If the overall difference HDi is larger than a given threshold T, a segment boundary is declared.

The above equation is to convert the color intensities into grey levels and only used for grey level frames with color frames.

4.2.3 Texture Feature

Texture feature extraction has been extensively used in many applications effectively which have information of texture as a salient feature in video. Many texture based algorithms are available for specific purposes. On the mode of using frequent texture models are Simultaneous auto regressive models, co-occurrences matrices, wavelet transformation based texture features, Tamura features, orientation features. Texture is a feature which provides significant information about the spatial arrangement of color or intensities in an image or identifying objects or regions of interest in an image .It can be defined technically as repetitive occurrence of the same pattern. The proposed method for extracting the texture feature is Entropy, which is a statistical measure of randomness can be used to characterize the texture of the given input image. The value of entropy can be determined as

$$ENT = \sum^{m}_{(k=1)} 〖pklogl /pk〗 \qquad [Eq2]$$

Where
ENT=Entropy of I/P,
M=Total number of samples,
P=Probability of I/P occurrences.

4.2.4 Shape Feature

It is the most prominent features among the others which determines to shape of the given objects in an image, can be extracted .More techniques of detecting the edges in the image. One is Edge Histogram Descriptor (EHD) has been used for capturing the spatial edges distribution, for TRECVid -2005, 2006 the video search task by Hauptmann et al. To determine the shape is very important step in a given image, for detecting edge techniques as Sobel, Prewitt, Roberts and Canny edge detectors respectively. Among all these four, canny detector's performance is more accurate than others to prove its efficiency, that's why Peak signal to noise ratio (PSNR) measure was used to provide a statistical method for its better performance.

4.3 High Level Features

Research efforts have been led to the development of challenging the semantic gap between low level and high level features that's why the high level feature term is introduced to improve video retrieval process more successively in future trend models as combined both high and low level features, indexing, multimodality, Data warehouse /media data and global features also be the killer application in this field of video retrieval. The proposed model is used here is object annotation for highlighting the regions or object of interest. Object classes were learned from a set of labelled training images in Label Me database. These dataset contains spatial annotations of thousands of object categories in hundreds of thousands of images.

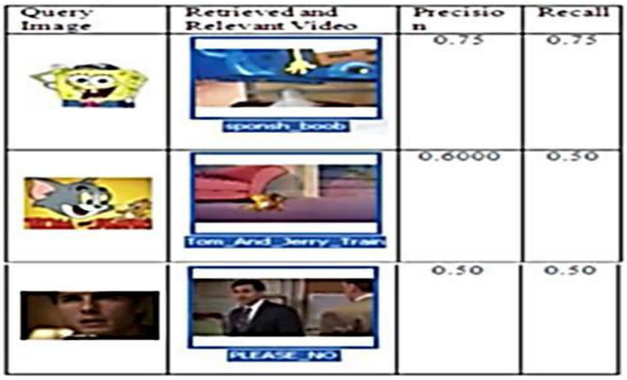
5. PERFORMANCE EVALUATION

The proposed system has been evaluated using several kinds of video sequences. We report here some results obtained on a part of a video sequence utilized for retrieval, its performance was compared to the performance of a video retrieval system based only on color feature. In order to evaluate the quality of the proposed system, recall and precision rates of the retrieved results against manual human opinions are used. Recall is a measure of how well the proposed system performs in finding relevant items, while precision indicates how well it performs in not returning irrelevant items and F-measure is an average of the formers. Recall, precision and F-measure are shown in formulas as follows:

Precision=|{relevant videos }-{retrieved videos}|/|{retrieved videos}|[Eq 3]

Recall= |{relevant videos }-{retrieved videos}|/|{retrieved videos}|[Eq 4]

The following figures shows that based on low and high level features as improved the results as 16% to 20% as compared to the previous results of existing works.



Query Image	Retrieved and Relevant Video	Precision	Recall
	sponsh_boob	0.75	0.75
	Tom_And_Jerry_Train	0.6000	0.50
	PLEASE_NO	0.50	0.50

Fig 4.Existing System results using color feature only

6. CONCLUSION

In this report, we briefly review the significance of video retrieval and indexing techniques. The field of video retrieval is a boon to adopt various video techniques involved to changes for the development of retrieval and to reduce the semantic gap between low and high level features. The first stage, shot boundary detection divides the video streams into their shots. Key frame selection is to select the key frame stored feature vectors as metadata. Finally the retrieval system accepts a user query, compares indexes derived from the user query with those stored into the metadata then returns search results sorted according to the similarity to the query. The proposed system is to reduce the semantic gap between multimodal features. Finally retrieved videos will be ranked according to their similarity and higher similarity than threshold value is returned to the user. The future work is to focus the multiple videos to retrieve the multiple shots using multimodality features with better algorithm to check with the similarity measures as Precision and Recall respectively.

7. REFERENCES

[1] X.Zhou , T.Zhang., et al. “ Relevance Feedback in content based image retrieval some recent Advances”, Proc. Of the 6th Joint Conf. on Information Sciences. 2002 15-18.

[2] Robles, Oscar D., et al. “Towards a content – based video retrieval system using wavelet-based signatures.” 7th IASTED International Conference on Computer Graphics and Imaging – CGIM. 2004.

[3] J.C.Haartsen et.al., “Adaptive Display For Enhancing audio Playback .” U.S. Patent Application 12/209,300, filed September 12, 2008.

[4]Bakker E.,Lew., et al.” Semantic video retrieval using audio analysis” In : International Conference on Image and video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer(2002) 260-267.

[5] visser, R., Sebe, N., Bakker, E.,et al. “Object Recognition for video retrieval” .In : International Conference on Image and video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer(2002) 250-259.

[6] Lew, Y.chang,D.J.Lee, Y.Hong, and J.Archibald, “Unsupervised video shot detection using clustering ensemble with a color global scae invariant feature transform descriptor,” EURASIP J. Image video process., vol.2008, pp.1-10, 2008.

[7] Yuan, H.Wang, L.Xiao, W.Zheng., et al. “ A formal study of shot boundary detection ,” IEEE Trans circuit Syst. Video Technol., vol.17, no.2 pp.168-186, Feb.2007.References

[8] Hatim G.Zaini and T.Frag,”Multi feature content based Video Retrieval Using High Level Semantic Concepts”,IPASJ International Journal of Computer Science,vol.2(2014) :15-25.

[9] Muhammad Nabeel Asghar, Fiaz Hussain, Rob Manton et.al,” Video Indexing : A Survey”, International Journal of Computer and Information Technology, vol.3(2014): 148-169.

[10] Shikui Wei, Yao Zhao, Zhenfeng Zhu and Nan Liu,”Multimodal Fusion for Video Search Reranking”, IEEE Transactions on Knowledge and Data Engineering,vol.22(2010): 1191-1199.

[11]Eng Wang, Zhanhu Sun, Yu-Gang Jiang and Chong-Wah Ngo, “ Video Event Detection using Motion Relativity and Feature Selection”, IEEE Transactions on multimedia,vol.16(2014) : 1303-1315.

[12] Luca Baroffio, Matteo Cesana, Alessandro Redondi, Marco Tagliasacchi and Stefano Tubaro,”Coding Visual Features Extracted From Video Sequences”, IEEE Transactions on Image Processing, vol.23(2014): 2262-2276.

[13] B V Patel and B B Meshram,” Content Based Video Retrieval Systems” , International Journal of UbiComp(IJU),vol.3(2012): 13-30.

[14] Weiwei Wang, Dong Zhai, Tao Li and Xiang chu Feng,” Salient edge and region aware image retargeting , Elseiver: Signal Processing : Image Communication:29(2014): 1223-1231.