Overview of Existing Content Based Video Retrieval Systems

Shripad A.Bhat, Omkar V.Sardessai, Preetesh P.Kunde and Sarvesh S.Shirodkar
Department of Electronics and Telecommunication Engineering
Goa College of Engineering
Farmagudi Ponda Goa

Abstract - With the advent of computer technology, a large amount of data is generated and stored. The data is stored in various ways such as in the form of text, image, videos, etc. As the stored data is increasing enormously over the years, efficient retrieval of the data has become difficult. Video sharing on the web is constantly on the rise due to the fact that mobile phones equipped with cameras has increased over the years and number of people accessing internet has increased manifold and video is one of the most preferred type of data owing to its content richness. Content-based retrieval of video data has been still an active research area. The efficient retrieval of video data is proven a difficult because of the unstructured way the videos are stored. In this paper we review the various content based retrieval systems and discuss the different parts of the content based retrieval systems.

Index Terms - Video annotation, Video database, Video indexing, Video retrieval, Video segmentation.

I. INTRODUCTION

There has been a tremendous growth in the usage of digital data. The most wide spread and common among the digital data is digital video which has become an essential part of our lives. Even if there are many tools to retrieve and process the digital data, it is a little cumbersome, less effective and less efficient. In the present day scenario the technologies that allow easy capture and sharing of digitized video have been rapidly developing, for example internet enabled mobile phones equipped with digital camera.

To date, handheld devices and the Internet have become a common method to create and transport video documents. As a result, there has been a huge increase in the utilization of video .So as to keep pace with the growth of video data; enhancing the current solutions for Content Based Video Retrieval (CBVR) is an important task.

Several content-based video retrieval (CBVR) systems have been proposed in the past, but they still suffer from the following challenging problems: (a) Semantic gap; (b) Semantic video concept modeling; (c) Semantic video classification; (d) Concept-oriented video database indexing and access.

So as to meet the demands of a user for video retrieval, numerous techniques have been proposed to bridge the semantic gap between the features that can be extracted fully without human intervention and the richness of user queries in Video retrieval.

One of them is domain-specific approach which exploits the typical characteristics of a particular video genre to design the most effective tools for the content extraction and indexing, and the associated browsing and retrieval.

In most Content Based Video Retrieval Systems (CBVR) the entire system could be broadly divided into four parts which are: (a) Video Segmentation (b) Video Annotation (c) Video Indexing (d) Video database. Depending on the effectiveness of each of the part mentioned above the efficiency and accuracy of the retrieval systems is dependant.

Figure 1: Block diagram of CBVR
II. VIDEO SEGMENTATION

In general, segmentation can be viewed as partitioning of data in classes having similar characteristics. Same is in case of videos but the difference is that, data being partitioned is a video. Video Segmentation is done so as to identify and group together similar content present in a video. This helps significantly in retrieving the relevant content from the video when there is a user query.

Frame: Frame is a smallest unit of a video which appears in the form of an image.

Shot: Frames recorded in a single camera operation.

Scene: Several shots combine together to form a scene.

Video: A Video is composed of different story units such as shots, scenes, and sequences arranged according to some logical structure (defined by the screen play).

A. TEMPORAL SEGMENTATION

Temporal segmentation of video as the name suggests deals with partitioning video in time into shots and scenes. This segmentation is done in a way so as to put together logically related frames together. Temporal segmentation for shots is performed by detecting the transitions between the shots. The transition is characterized by a cut between the shots. The transition is either sudden or gradual and detection of the transition in former is easier by comparing the consecutive frames of video. If the transition is gradual it is difficult to detect the transition as compared to sudden transition.

One of the simplest algorithms used for temporal segmentation is Pixel wise comparison. In this algorithm successive frames are compared. If the difference between the pixels is greater than the threshold, then transition is detected.

\[
D(i, i+1) = \frac{\sum_{x} \sum_{y} |P_i(x, y) - P_{i+1}(x, y)|}{XY}
\]

\(D(i, i+1)\) is difference between pixels of \(i^{th}\) and \((i+1)^{th}\) frame.

\(X\) and \(Y\) are the dimensions of the frame.

\(P_i(x, y)\) corresponds to intensity value of pixel of \(i^{th}\) frame at coordinates \((x, y)\).

A cut is detected whenever the difference between the pixels of frames is greater than the threshold. The main disadvantage associated with this algorithm is that it fails when there is a large change in a small part or small change in a large part of the frame. A false cut is detected in the above cases.

A small improvement over this method could be obtained by counting the number of pixels that undergo change between successive frames. If the change in the number of pixels is greater than threshold, then a cut is detected.

Shot Boundary Detection (SBD) is process of automatically detecting boundaries of shots of a video and research in this field has matured over the years. The methods used for the purpose of Shot Boundary detection have been described in [1], [2], [3]. In [3] a study of seven years of TRECVID activity has been presented. The most predominant methods used are: (1) Color histogram (2) Edge information (3) Luminance values. Combinations of these methods were also used and Color histogram along with edge information proved to be better in most cases. An average of 79% for precision and recall was obtained and no more significant improvement was achieved further. In [4], a local key point matching algorithm is presented to detect the shot changes using a so-called color context histogram (CCH) [5] feature computed around Harris corner points [6]. Hannes, in his paper [6], describes another method to achieve the same and describes it as, shot boundary detection using Self Organizing Maps (SOMs).

In [8] a two-pass, graph-based shot boundary detection algorithm, inspired from the graph based image segmentation algorithm proposed by Felzenszwalb and Huttenlocher is used for the purpose of Shot Boundary detection. In [7] segmentation is achieved by constructing an undirected graph \(G = (V, E)\) from the video frames; each frame \(F\) is a vertex \(v\) and edge weights are computed as the distances between the frames, \(w(v_i, v_j) = d(F_i, F_j)\). Edge weights are the distances between the pixels. In the first pass of this algorithm abrupt transitions are detected, while in the second pass gradual transitions are detected.

B. SPATIO-TEMPORAL SEGMENTATION

Spatio-temporal segmentation deals with extracting moving regions present in videos. Some spatio-temporal segmentation approaches rely on initially applying spatial segmentation
method independently to each frame. Spatio-temporal objects are subsequently formed by associating the spatial regions formed in successive frames using their low-level features. In this, general approach is apprehended by assuming a seeded region-growing scheme for performing spatial segmentation. The temporal correspondence of spatial regions between adjacent frames is established by examining the overlapping of these seeds, which in the case of the selected region growing technique may be quite large, ranging from 1/3 to 2/3 of the final region size.

In [9] two contributions have been made. Primarily, an efficient spatio-temporal video segmentation algorithm, which naturally incorporates long-range motion cues from past and future frames by exploiting groups of point tracks with coherent motion, is developed. Next, a new track grouping cost function that includes occlusion reasoning, in the form of depth ordering constraints, as well as motion similarity along the tracks is devised.

In [10] an efficient way to perform spatio-temporal segmentation is proposed. The proposed scheme deals with the temporarily stopping problem successfully. The temporal segmentation produces a temporal mask that indicates moving regions and static regions for each frame. For localization of moving objects, a block-based motion detection method considering a novel feature measure is proposed to detect changed regions. These changed regions are coarse and need accurate spatial compensation. An edge-based morphological dilation method is presented to achieve the anisotropic expansion of the changed regions. Furthermore, to solve the temporarily stopping problem of moving objects, the inertia information of moving objects is considered in the temporal segmentation.

C. KEYFRAME

In a video lot of redundancy is present in terms of the information present in each of the frames within a shot. Keyframe is the one which gives the best possible information by getting rid of the redundancy. The extracted key frames should contain as much relevant content of the shot as possible and avoid as much redundancy as possible. The features used for key frame extraction include colors (particularly the color histogram), edges, shapes, optical flow, MPEG-7 motion descriptors such as temporal motion intensity and spatial distribution of motion activity [10], MPEG discrete cosine coefficient and motion vectors [11], camera activity, and features derived from image variations caused by camera motion [12], [13].

In [14], six different categories for extraction of Key frames have been discussed.

1. Sequential comparison between frames [15], [16] Frames subsequent to a previously extracted key frame are sequentially compared with the key frame until a frame which is very different from the key frame is obtained.

2. Global comparison method [17], [18] Based on global differences between frames in a shot distribute key frames by minimizing predefined objective function that depends on the application.

3. Reference frame based [19], [20] These algorithms generate a reference frame and then extract key frames by comparing the frames in the shot with the reference frame.

4. Clustering [21], [22], [23] These algorithms cluster frames and then choose frames closest to the cluster centers as the key frames.

5. Curve Simplification [24] These algorithms represent each frame in a shot as a point in the feature space. The points are linked in the sequential order to form a trajectory curve and then searched to find a set of points which best represent the shape of the curve.

6. Object/Events [25], [26], [27], [28] These algorithms jointly consider key frame extraction and object/event detection in order to ensure that the extracted key frames contain information about objects or events.
III. VIDEO ANNOTATION

Video annotation is a marked-up comment made to information present in a video. Video annotation is also termed as metadata, tags, comment, etc added to a video. Video annotation is the method by which information present in a video is described so that retrieval of the information present in video could be made possible.

The availability of semantically annotated video assets constitutes a critical prerequisite for the realization of efficient content based retrieval systems pertaining to realistic user needs. In the early years of the research on Video Annotation, manual annotation techniques were developed. In case of manual annotation a person is required to annotate the videos or parts of the video. This is an extremely tedious and inaccurate task to achieve. Hence there is research being conducted so as to automate this task of annotation of videos.

Metadata is attached to the video for the purpose of video annotation. The metadata could be of various types as listed below.

1) **Content independent metadata:** Metadata associated with the video but not directly describing it. For example, Name of the Author, Date, Location, etc.

2) **Content dependent metadata:** It refers to the low level and intermediate level features in the video. The low level features that can be used for annotation are color, texture, edge, motion, etc.

3) **Content descriptive metadata:** It refers to content semantics. It is related to the content of the video, the entities or objects, events, emotions and meaning of scenes.

A. VIDEO ANNOTATION TECHNIQUES

In [29] various techniques for annotation of videos have been described.

1) **Free text descriptions**

Free textual descriptions could be added to video. There is no pre-defined structure for the annotation [30]. For example, user can add description about the video while uploading a video on YouTube. Any combination of words or sentences can be used. Such type of annotation helps in accessing the video. Since no structuring exists, annotation is an easy task but, efficient retrieval techniques must be used.

2) **Based on the text in video**

The textual information that exists in images and video sequences are called collateral text. For example, the text of news, documentary programs, movies and even newspaper film reviews [31]. Textual data is a source of highly semantic information and thus, if available, would allow the filtering and searching of video data by users in a more intuitive and natural way. Text embedded video; especially captions provide brief and important content information, such as the name of players or speakers, the title, location date of an event, etc. [32].

3) **Based on machine learning**

From the video low-level features can be extracted. Various machine learning techniques such as support vector machine (SVM), Bayesian networks, Clustering, similarity and metric learning can be used. In [33] [34] and [35] different approaches for video annotation based on machine learning have been discussed.

4) **Based on Rule Learning**

Visual features can be extracted from the video. These low-level features can be used for annotation but gap exists between the information that can be extracted automatically from visual data and the interpretation that the same data has for a user in a given situation: the semantic gap. Rules are made to infer a set of high-level concepts from low-level descriptors. Jardon et al. introduced a rule-based approach for the generation of inference rules using fuzzy logic [36]. In [37] the proposed system automatically annotates the video shots based on the pre-annotated data set.

5) **Based on Graph**

Graph-based learning is a semi-supervised method. Graph with labelled and unlabelled vertices are used. These vertices are samples; the edges reflect the similarities between sample pairs. A function is estimated on the graph based on a label smoothness assumption. These methods have already been successfully applied in image and video content analysis [38], [39].

6) **Based on ontology**

Explicit specification of a conceptualization can be defined as Ontology. Classification of different aspects of life into hierarchical categories is achieved in Ontology. Ontology consists of entities and their relationships, which may be organized as classes and subclasses, each class may also consist of one or more instances.

For example, it can be found that a “room” is a subclass of the class “house”. Jin-Woo Jeong, Hyun-Ki Hong, and Dong-Ho Lee [40], have presented an automatic video annotation technique which makes use of ontology to facilitate video retrieval and sharing process in smart TV environment. In [41], ontology enriched semantic annotation of CCTV video is proposed. Visual and text semantics are linked with appropriate keywords provided by domain experts. These visual semantics are annotated by keywords of CCTV ontology.
IV. VIDEO INDEXING

Video indexing can be defined as the process of extracting from the video data, temporal location a feature and its value. Segmentation of videos is achieved first in content based video retrieval systems and next phase is the annotation of the different parts of the video. Once these two things are accomplished, video indexing is used so as to find out which part of the video is relevant when a user provides a query.

Video indexing approaches can be categorized based on the two main levels of video content:

1) **Low-level (perceptual) features** [41], [42]
   The main benefit of this approach is that it can be fully automated. Users can do the search based on the features such as color, shape, etc. But the problem associated with this approach is that it ignores the semantic contents, while users search is usually based on the semantic contents.

2) **High-level (semantic) annotation** [43], [44]
   Advantage of semantic annotation is that it supports more powerful, natural and flexible ways of querying. This type of indexing is achieved through manual intervention because there is a need to map the lower level features to higher level features by bridging the semantic gap.

A. VIDEO INDEXING TECHNIQUES

Different Video indexing techniques have been devised based on various features of the video. Classification of these techniques has been presented below.

1. SEGMENT BASED INDEXING TECHNIQUES

A video can be decomposed into a hierarchy similar to storyboard in filmmaking [45], [46], [47]. In this video indexing technique video is divided based on the different stories present in it which is further divided into scenes and shots. Let us consider a example of a news video which consists of a sports, politics and entertainment sections. Segment based video indexing technique divides the entire video into the sections as mentioned above. Further each of these sections can be divided into scenes and shots depending on the various content present in them.

Keyframes are an essential part of the segment based video indexing techniques. Keyframes are used to represent the segments of the video. Advantage offered by them is less storage space required as compared to the video.

2. OBJECT BASED INDEXING TECHNIQUES

The aim of the Object based video indexing techniques is to distinguish different objects present in a video so as to capture content changes. Extraction of objects in an important and difficult part of the of object based video indexing technique. If the video is compressed using MPEG-4 object based coding standard, it is relatively easier to extract the objects. Kim and Hwang in their work [48] have proposed object tracking using edge detection. Yajima et al [49] proposed an indexing framework to allow users querying objects’ movements that include player and ball and querying by drawing a moving direction on a video screen directly. Their framework depends on tracing the spatio-temporal relationship between moving objects.

3. EVENT BASED INDEXING TECHNIQUES

Event based video indexing techniques intend to detect interesting events in a video by tracking the objects present in it. Event can be defined as a visual flow of relations between the objects in a time interval. One of the major applications of this technique is sports video indexing. This technique is more favourable as compared to segment and object based video indexing technique.

Detection of events in sport videos has been based on manual work [50] and automatic analysis of visual features using technologies such as camera or editing techniques analysis [51], global motion estimation [52], foreground-background extraction including visible object recognition [53], and the detection of CC (closed caption) streams [54].

4. ANNOTATION BASED INDEXING TECHNIQUE

Annotation based video indexing is based on the process of annotating the videos. Description of the content present in video is done with the help of annotation. In this technique user queries can be managed using standard query language. Text data management, such as Information Retrieval (IR) techniques have matured and successfully supported by traditional DBMSs.

However, the major limitation of this approach is the fact that it would be extremely tedious and ineffective to manually annotate every segments of video. Despite these limitations, this approach still needs to be explored as annotations can still be the closest representation of video semantic content. Moreover, a video application such as sport and news can share many keywords; therefore, a
‘glossary’ of keywords can be constructed to assist a more uniform indexing and making queries easier.

Annotation based video indexing have been used in [55] for the efficient retrieval of video. In this paper tags provided to the images on web used for the purpose of annotation of the video and further efficient retrieval.

V. VIDEO DATABASE

Database is a collection of information organized in such a way that a computer program can quickly select desired pieces of data. It is analogous to an electronic filing system.

Traditional databases are organized by fields, records, and files. A field is a single piece of information; a record is one complete set of fields; and a file is a collection of records. For example, a telephone book is analogous to a file. It contains a List of records, each of which consists of three fields: name, address, and telephone number. Video database is a collection of videos organized in a way that enables the user to retrieve a particular video. In [8] videos are processed offline and their mpeg-7 compatible xml representations are stored in an xml database. The users can enter their queries on clients visual query interface. The query interface then interacts with the query processing server using tcp/ip and retrieves the required video. Using a distributed architecture, the query processing speed can be increased and this would reduce the query response time. In [56], a prototype system is designed that enables fast and robust Near-Duplicate Video Clip detection based on visual content present in the video.

V. CONCLUSION

In this paper, we have presented a review on recent developments in content-based video indexing and retrieval. In this work, we have summarized various techniques used in content based video retrieval systems. Some of the most efficient and popular tasks used in CBVR include, video segmentation, video annotation, video indexing and creation of a video database. In this review paper, we conclude that, the recent work done on segmentation and database has matured, but the major challenge and research has to be performed in bridging the semantic gap through efficient annotation and indexing techniques.

REFERENCES


Shripad A.Bhat
Residing at Chinchinim Salcete Goa
Born on 30th August 1993
Currently pursuing B.E degree in Electronics and Telecommunication Engineering from Goa College of Engineering
Farmagudi Ponda Goa
A member of IEEE GEC
Student branch

OmKar V.Sardessai
Residing at Porvorim Goa
Born on 22nd September 1990
Currently pursuing B.E degree in Electronics and Telecommunication Engineering from Goa college of Engineering
Farmagudi Ponda Goa
A member of IEEE GEC
Student branch

Preetesh P.Kunde
Residing at Navelim Goa
Born on 22nd April 1991.
Currently pursuing B.E degree in Electronics and Telecommunication Engineering from Goa college of Engineering
Farmagudi Ponda Goa
A member of IEEE GEC
Student branch
Sarvesh S. Shirodkar

Residing at Panjim Goa
Born on 29th April 1993
Currently pursuing B.E degree in Electronics and Telecommunication Engineering from Goa College of Engineering Farmagudi Ponda Goa
A member of IEEE GEC Student branch