Front. Comput. Sci., 2013, 7(5): 782–799
DOI 10.1007/s11704-013-1276-6

REVIEW ARTICLE

A comprehensive review of significant researches on content based indexing and retrieval of visual information

R. PRIYA (✉), T. N. SHANMUGAM

Department of Mathematics, Anna University-Chennai, Chennai 600 025, India

Abstract Developments in multimedia technologies have paved way for the storage of huge collections of video documents on computer systems. It is essential to design tools for content-based access to the documents, so as to allow an efficient exploitation of these collections. Content based analysis provides a flexible and powerful way to access video data when compared with the other traditional video analysis techniques. The area of content based video indexing and retrieval (CBVIR), focusing on automating the indexing, retrieval and management of video, has attracted extensive research in the last decade. CBVIR is a lively area of research with enduring acknowledgments from several domains. Herein a vital assessment of contemporary researches associated with the content-based indexing and retrieval of visual information. In this paper, we present an extensive review of significant researches on CBVIR. Concise description of content based video analysis along with the techniques associated with the content based video indexing and retrieval is presented.

Keywords nultimedia information, content based video retrieval (CBVR), content based video indexing and retrieval (CBVIR), shot segmentation, object segmentation, feature extraction, indexing, motion estimation, querying, key frame, retrieval, and indexing.

1 Introduction

With the arrival of broadband networks, high-powered workstations and compression standards, the importance of multimedia information systems has increased. There is a heavy requirement to efficiently index, store and retrieve the visual information from multimedia database, as visual media needs huge quantities of storage and processing [1]. The modern-day growth of various multimedia compression standards along with a considerable increase in desktop computer performance and a decrease in the cost of storage media, has led to the extensive exchange of multimedia information. The availability of cost effective means for obtaining digital video gained the easy storage of digital video data, which can be widely distributed over various types of networks or storage media [2]. The last two decades have resulted in a considerable development in the multimedia and storage technology that has led to the creation of a large repository of digital image, video and audio data. Video is swiftly becoming the most popular media, owing to its high information and entertainment power [3]. Video sequences reveal more information about how objects and scenarios change over time, as compared to still images, which they are subjected to appropriate processing step as like in image [4,5]. They can present more information than text, graphics and static images. This can relate to the position, distance, temporal and spatial relationships that are built-in in the video data implicitly. But, video needs more space for storage and wider bandwidth for transmission [6].

A video or video sequence is hierarchical in nature and it can be perceived as a set of story units. Each story unit includes a set of scenes and a scene is a set of interrelated shots which are unified by the same point of interest. A shot is defined as one or more image recorded contiguously that

Received November 19, 2011; accepted January 17, 2013

E-mail: priyarajendranphd@gmail.coma; priyacrs@yahoo.com

represents a continuous action in time or space. There is little change in the visual content of a shot. A shot consists of multiple real-world objects and captures their semantic, their dynamic and their syntax the way objects are merged to obtain an image sequence, such as spatial/temporal inter-object relationships. A shot is commonly supposed to be composed of rigid objects or objects consisting of rigid parts connected together. Two shots are separated by a cut which is a transition at the image boundary between two successive shots. A cut can be thought of as an "edge" in time [7]. The benefit of using video as an alternative of a single image for recognition is that, the different poses of the object can be visible and also the smoothness in pose transitions can be used to the maximum advantage for recognition [8]. The structural content of a typical video is depicted in Fig. 1 [9]. Interpretative video-analysis is a quite new, though quickly expanding innovation in social science methodology [10]. Immense researches are available in a digital video analysis, a few of them are shape based [11], object based [12], motion based [13], event based [14] and more. In addition to that, content based analysis plays a key role in video analysis techniques.

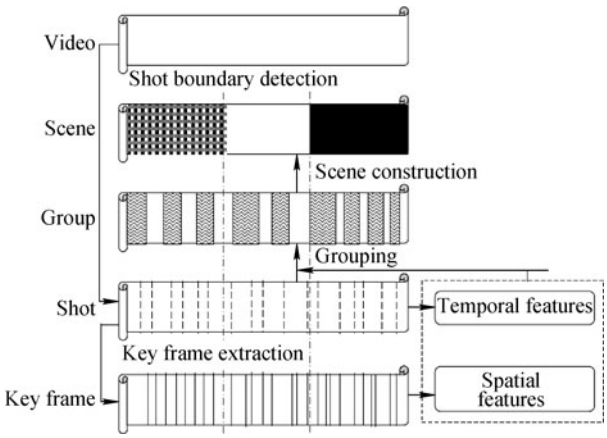

Fig. 1 Structural content of a typical video

In recent times, content-based video analysis has emerged as a common and interesting research matter that is accompanied by numerous multimedia applications where effective access to video data based on its content is essential [15]. Content-based indexing and retrieval of visual information has gained attention in the research community over the past decade [16]. Content-based searching, browsing and retrieval is more natural, friendly and semantically significant to users. The need of content-based multimedia retrieval encourages the research of feature extractions of the information embedded in text, image, audio and video. A video data model should include elements that signify inbuilt structural properties of video as well as elements that represent the video content. Thus in this multimedia environment, queries can be based not only on exact matching of textual data, but also on the degree of similarity of visual features. So, a suitable query mechanism is necessary to formulate queries on these properties. The query procedure fundamentally consists of three interactive steps: query formulation, query processing and query results presentation. This involves locating expressive methods for conveying what is desired, the capability to match what is expressed with what is there and ways to evaluate the outcome of the search [17].

In past decades, a huge range of researches and techniques were available in content based indexing and retrieval of visual information for effective analysis. However, most of the the reviews focus only on a particular stage of content based video indexing and retrieval system. As a result, they lag in delivering impact of combination of different techniques in different stages of the system. Hence, it is essential to review the system in a wider way so that different stages of the system have to be reviewed. Despite the work does not provide research contribution, it presents a broad review on different techniques in the literature on different stages of the content-based indexing and retrieval along with its processing and analysis techniques. Unlike other reviews, we review all the stages of the system that have been proposed in the literature and we demonstrate the performance variation. An overview of different video content modeling, retrieval and classification techniques employed in existing content-based video indexing and retrieval (CBVIR) systems are shown. Based on the modeling requirements of a CBVIR system, we analyze, categorize and review the existing modeling approaches. Additionally, we propose a concise description about content-based video analysis along with the various procedures involved in content based video indexing and retrieval. We hope, that this would provide a good knowledge about the impact of various techniques on the retrieval performance. The other details are as follows: in Section 2, a brief description of the processes involved in content-based indexing and retrieval system is presented. A broad review on the study of significant research methods in content-based indexing and retrieval of visual information is provided in Section 3. Performance analysis of content based video indexing and retrieval paper review. Section 5 describes directions for the future research and Section 6 wind up the paper.

2 Content based video indexing and retrieval

Video retrieval, searching and retrieving the videos relevant

to a user defined query, is one of the most accepted topics in multimedia research [18]. With the rapid enhancements in both centralized video archives and distributed WWW video resources, content-based video retrieval is gaining its importance. To support such applications efficiently, content-based video indexing must be addressed. Content-based video retrieval is developing technologies to automatically parse video, audio and text to identify meaningful composition structure and to extract and represent content attributes of any video sources [19]. Content-based video retrieval requires many changes in a multimedia database management system, mostly in the modeling and querying techniques. Valuable content-based retrieval of imagery and video can be performed at three abstraction levels [20–22].

- Raw data: At the lowest abstraction level, objects are merely aggregations of raw pixels. Comparison between objects or regions is done on a pixel-by-pixel basis using similarity measures such as the correlation coefficient and the Euclidean distance.
- Feature: A feature is a distinguishing primitive characteristic or attribute (e.g., luminance, shape descriptor, gray scale texture, color histogram and spatial frequency).
- Semantic: At the maximum abstraction level, retrieval assumes that features have been grouped into meaningful objects and semantic descriptions have been attached to scenes. Search is carried out on entities with well defined spatiotemporal properties [23].

On using low-level features to segment video into shots, each of which is contained of a sequence of consecutive frames recording a video scene or event contiguous in time and space [24,25]. After the video is segmented into shots, a number of consecutive frames can be extracted from each shot to signify the salient content of the shot for video analysis, indexing and retrieval purposes [26,27]. Based on the extracted feature vectors from video data, video retrieval can be done by evaluating the similarity between the query vector(s) and the feature vectors in the database [19,28,29]. Figure 2 illustrates the flow chart of video segmentation, frame extraction, feature extraction and video indexing and retrieval; and the subsequent sections explain the processes.

2.1 Video segmentation

In content-based video retrieval, video segmentation creates video shots distinguished by a certain degree of visual cohesiveness [30]. Video segmentation is the first preprocessing step to further examine the video content for indexing and browsing. Video segmentation or shot boundary detection involves temporal segmentation of video sequences into shots. A shot in a video is a contiguous sequence of video frames recorded from a single camera operation, representing a continuous action in time and space [31]. Shot transitions can be abrupt (cuts) or gradual (fades, dissolves and wipes). Many methods have been developed to detect the video shot boundaries [16]. Video object segmentation is a hot topic in the communities of computer vision and pattern recognition, due to its potential applications in background substitution, video tracking, general object recognition and content based video retrieval [32]. The motive of object segmentation is to locate the semantically meaningful objects in the observed scene [33].

2.2 Feature extraction

In feature extraction, features are extracted from images based on different image information [34]. Features are

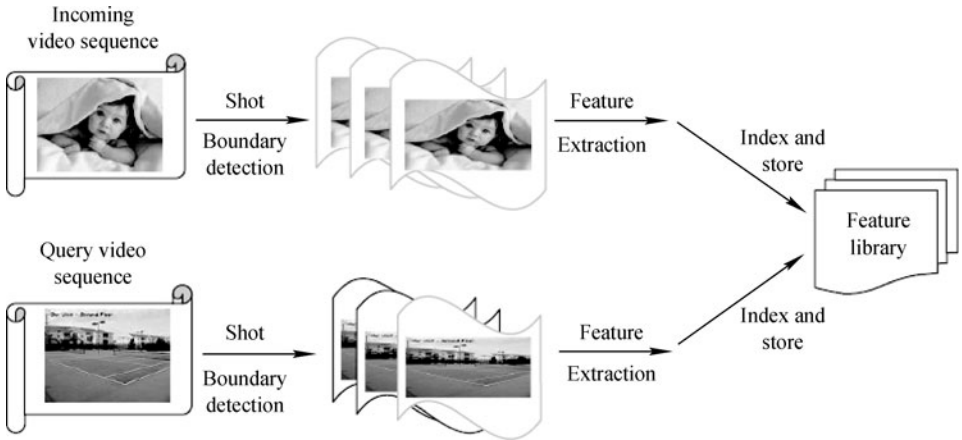

Fig. 2 Flowchart of content based video indexing and retrieval system

objects in images carrying similar properties, e.g., regions of a certain color, corners, lines, parabola and more. Such features can then be used in many computer vision techniques, including object or scene recognition, solving for 3D structure from multiple images, stereo correspondence and motion tracking [35]. Using color and motion segment information, a few numbers of key frames or scenes that offer sufficient information about the content of a video sequence is extracted. Key-frames are those frames significant to understand video content and key-frame definition is quite subjective. It could be related to motion, or objects, or events [36]. This can be used for reduction of the amount of stored information that is necessary in order to provide search capabilities in a multimedia database. Furthermore, instead of performing a query on all available video frames, one can only consider the selected ones, because they include most information about the content of the database [37].

2.3 Video indexing

Once the video segmentation is operated at a desired level, the indexing of the document is performed by creating some meta-data [38]. The variation of the content of meta-data depends on the application towards which the database is oriented. For generic video documents, this data normally includes video object (shot, scene) boundaries along with some characteristic and visual representation. One common representation is the choice of one or more key frames within the shot or the sequence. Depending on the assumptions under which the segmentation has been performed, all frames within a basic shot should normally be consistent with each other. Heuristics for choosing one or more key-frames can therefore be derived. The simplest relies on the global position of the frames within the shot (first, middle and end). Some other characteristics such as the corresponding audio stream may also be used for efficient key frame detection [39]. The process of re-segmenting the shots with respect to some heuristics reflecting a comprehension of the video content is referred to as video micro-segmentation [40].

2.4 Video retrieval

In the context of imagery and video, the goal of content-based retrieval (CBR) systems is to recover a set of images or video clips from a large collection on the basis of the internal content of the desired items, in addition to associated alphanumeric keywords and attributes. This has two phases-online and the offline phase. During the offline phase, broadcast videos in multiple languages are stored in a video database. Cuts and theme changes are identified in each video for the segmentation. These videos are further processed frame by frame for identifying possible text regions. A selected set of features is extracted from these text regions and stored in a feature database during the online phase; a search query is entered through a graphical user interface. This query string is rendered as an image and the corresponding set of features is extracted. These features are identical to those employed in the offline process. A matching algorithm then computes the degree of similarity between the features of the search query and those present in the feature database. The results are then ranked based on the similarity measure [41].

2.5 Video querying

The problem of searching a video document calls for that of formulating a meaningful and clear query. For content-based image retrieval, the system of query-by-example is comparatively intuitive as it caters for cases which would be hard to solve using simple text queries. For video documents however, things are no so straightforward since a query-by-example would want the user to have a video at hand already. One of the major problems is the extreme dimensionality of the search space induced by the temporal information. In order to decrease this dimensionality, different approaches are used [40].

2.5.1 Visual query

In this most basic category the user is informed to select one image which is supposed to be similar to the one he is searching for. As a result, the system returns in decreasing order of the images until it finds the most similar to the given example. An enrichment of such a system allows the user to choose more than one example so that the query combines all familiar features in these images. The generalization of that system uses feedback so that the search follows an online learning of the user's expectations. Negative and positive feedback rates are used as equivalent AND & NOT operators. Another refinement of this principle consists in using only parts of the images for the query. This requires the definition of a perceptually-meaningful segmentation technique through which images in the database are pre-processed.

2.5.2 Motion query

The motion is the only means of representing the temporal information contained in a video document. Motion-based query is therefore an attractive feature of a video search engine. The main problem is the formulation of such a query.

Motion-based queries can be seen as counter-intuitive in the sense that the user is asked to signify a motion in some still fashion. It is obvious that solving the problem of the formulation of a motion-query is to be made in parallel with that of representing the motion information in the indexing task.

2.5.3 Textual query

Textual query is a promising approach for querying in video databases. It offers a more natural interface, as there are many powerful clues hidden in the video clips. Audio and the textual content in videos can be of immense use in indexing. Textual information is available as captions on the video or printed/handwritten text in the scene. If we can detect and recognize the textual content, it can be effectively used for characterizing the video clips, as a complementary measure to the visual appearance-based features [41].

2.6 Applications

Potential users may come up from different horizons. In the case of CBVIR system, major users can be listed as follows [40].

- News broadcasting. The need to maintain complete archives induces a huge quantity of typically short documents. The specificity of news report video documents makes an automated CBVR very striking to retrieve documents with respect to a specific topic, place or appearance of a given character. One should also note that the audio stream attached to all such documents as well as captions usually present in news reports form important cues for an automated retrieval process.
- Advertising. Here the documents are typically short. An example of retrieval may concern the fact of retrieving all document w.r.t. a style of shooting. Such a characteristic is usually complex to express using text so that a visual-based approach is highly desirable.
- Music video clips. Here again, the characteristics of such documents may be complicated to express based on given textual annotations stored on a database. For this type of specific applications, one may think of retrieving documents on the basis of some dance step described by e.g., a sketch.
- Distant learning. Two types of educational documents can be distinguished. Firstly, lectures where the main content is given by the audio stream and practical courses when both the visual contents and the audio stream are of importance.
- Video archiving
- Medical applications

3 Extensive review of significant researches on content based video indexing and retrieval system

A range of research methodologies employed for developing the successful framework for content based indexing and retrieval of visual information is presented in this section. The reviewed CBVIR models are classified and described in the following subsections.

3.1 Feature extraction

Zhang and Chen [42] have developed a technique to extract objects from video sequences based on spatiotemporal independent component analysis (stICA) and multiscale analysis. The preliminary source images with moving objects in video sequences are extracted using statistical formulation which is based on stICA. A mathematical framework has presented in the context of the video frame analysis. One of the most important advantages of using this statistical analysis is that it has captured both the spatial and temporal characteristics of moving video objects in frames without getting into the detailed pixel-based processing. Now the source image data after stICA analysis is further processed using wavelet-based multiscale image segmentation and region detection techniques. Based on these techniques, an automated video object extraction system has been developed.

Chang et al. [43] have presented the development of a color feature extraction algorithm and they have also proposed a clustering strategy using clustering ensembles for video shot detection. Based on Fibonacci lattice-quantization, they have developed a color global scale-invariant feature transform (CGSIFT) for enhanced description of color contents in video frames for video shot detection. To start with, CGSIFT first quantizes a color image representing it with a small number of color indices and then used SIFT to extract features from the quantized color index image. The second step is to represent the global color features. To achieve this, they have developed a space description method using small image regions. Clustering ensembles focusing on knowledge reuse are then applied to obtain better clustering results than using single clustering methods for video shot detection.

A hierarchical framework for analyzing high-level events in soccer video by combining low level feature analysis with high level semantic knowledge has been presented by

Kolekar et al. [44]. The sports domain semantic knowledge encoded in the hierarchical classification reduced the cost of processing data drastically and also significantly improved the classifier accuracy. The hierarchical framework enabled the utilization of simple features and organized the set of features in a semantically meaningful way. The next task addressed in their work was semantic concept mining based on proposed sequential association distance. They have observed hopeful results for mining of the Goal scored by team-A, Goal scored by team-B, Goal saved by team-A and Goal saved by team-B concepts. The results have demonstrated that their proposed mining technique was effective and they could achieve average recall 90.83% and precision 86.57%. For the sake of soccer video they considered only the goal and save concept.

Jiang and Crookes [45] proposed fast automation motion segmentation method. Their proposed method uses labeled regions to segment different video objects from the background. Their proposed approach widely differs from the conventional pixel or edge based motion segmentation. Moreover, Bayesian logic is used to cluster the facets into objects. Since the facets are lower in number than edges and points, facets are utilized to reduce the complexity of motion segmentation. Hence their proposed method can handle the complexity of video object motion tracking more effectively and efficiently and it also provides the potential for real-time content based video annotation.

Basharat et al. [46] proposed a framework for matching video sequences using the spatiotemporal segmentation of videos. They have used the interest point trajectories to generate video volumes rather than using appearance features for region correspondence across frames. The temporal correspondence between the estimated motion segment is then created based on most common SIFT correspondences. A two pass correspondence algorithm is utilized to process the splitting and merging regions. Spatiotemporal volumes are extracted by the consistently tracked motion segments. After that, a set of features including color, texture, motion and SIFT descriptors are extracted to signify a volume. They have made use of an Earth mover's distance (EMD) based approach for the comparison of volume features. Experiments for video retrieval were conducted on a large number of videos obtained from different sources like BBC motion Gallery and hopeful results were obtained.

3.2 Shot detection

A framework to automatically group similar shots into one scene, where a scene is generally referred to as a group of shots taken place in the same site have been proposed by Ngo et al. [13]. Their motion-based video representation scheme has been proposed for scene change detection by integrating the motion characterization and background reconstruction techniques. With the help of this scheme, an adaptive keyframe selection and formation method has been derived. Their approach is helpful in not only selecting the keyframes from shots but also in reconstructing the new images such as background panoramas as new keyframes based on the annotated motion of shots. This proposed method can be used for scene change detection as it effectively performs the histogram intersection, which measures similarity between features based on the intersection of feature points and more over the proposed video representation scheme is also quite compact. Experiment results are very hopeful after the combination of the histogram intersection for similarity measure along with time constraint grouping algorithm.

Kuo and Chen [47] suggested an approach for segmenting videos into shots on moving picture experts group (MPEG) coded video data. By finding the probability for each frame, their approach has detected the shot changes. In order to avoid the time for decoding and processing video data frame by frame and pixel by pixel, the MPEG coded video are decoded only partially. A set of masks for different types of MPEG coded frames (I, P and B frames) is defined for the computation of the shot change probability. A number of experiments based on various parameters are done in order to show an average detection rate of 95%. On further detection of the dissolve effect, the result was improved to reach an average 98% recall and 96% precision. A video indexing tool based on their approach was implemented. Their approach was more efficient that it also provides fast video browsing and query-by-image capabilities. Thus the results of detected shot changes are maintained so that the video retrieval and browsing can be provided.

Chen et al. [48] have proposed a mechanism to automatically parse sports videos in compressed domain and then to construct a concise table of video content employing the super imposed closed captions and the semantic classes of video shots. Initially GOP (Group of Pictures)-basedvideo segmentation is used to examine the shot boundaries. Color-based shot identification is then exploited to automatically identify meaningful shots. The proficient approach of closed caption localization is proposed to detect caption frames in meaningful shots. Then instead of frames, caption frames are selected as targets for detecting closed captions based on long-term consistency without size constraint. Moreover, in order to support discriminate captions of interest automatically, a

tool-font size detector has proposed to recognize the font size of closed captions using compressed data in MPEG videos. The experimental outcomes have showed that the efficiency and the viability of their proposed mechanism.

Duan et al. [49] have proposed a unified framework for semantic shot classification, with a stress on knowledge representation and acquisition. Their framework is relied on mid-level representations instead of exhaustive low-level characteristics. Experiments have proved that the accurateness and flexibility of shot classification can be enhanced through appropriate construction of mid-level representations. Their proposed mid-level representations can be stretched out to additional sports video analysis process such as event detection, highlight extraction and much more. The wedding between machine learning algorithms and human constructed knowledge proved effectual for obtaining mid-level representations. They justified the proposed mid-level representations through the task of video shot classification.

A semantic analysis and annotation approach by using multimodal analysis of video and audio information tested in basketball videos are presented by Liu et al. [50]. Motion prediction information is used in detecting shot and scene boundaries. In addition, motion features, describing the total motion, camera motion and object motion, are utilized for the purpose of scene classification. Identically, their proposed HMM-based method for audio key sound detection outperformed the earlier SVM-based technique, especially for the excited commentator speech and excited audience sounds. This conformed to the fact that the HMM based technique efficiently captures rich contextual information so as to advance various key sounds separately. Experimental results have also illustrated the effectiveness of event detection by using the mixture of audio and visual information.

Han et al. [51] have been proposed a rough-fuzzy operator based on rough-fuzzy sets for feature reduction and the dissimilarity function. Shot transition can be divided into three types based on the characteristics of new scenes and they are, cut transition, gradual transition and no transition. The effectiveness of the presented method is tested on more than two hrs of news programs and 98.0% recall with 96.6% precision has been achieved.

A method for video segmentation and a model for anchorperson detection scheme for news story parsing have been proposed by Gao and Tang [52]. In their proposed system by utilizing the fuzzy C-means clustering algorithm they first segmented the news program into video shots and to improve the performance of shot segmentation they introduced several techniques such as two-pass strategy, post processing based on fuzzy membership values and gradual transition differentiation through binary pattern matching. Once the news video is parsed into camera shots, by using the graph-theoretical cluster (GTC) algorithm they have found the anchorperson shots. However the individual news stories can be reconstructed based on the structural model of the news program. Experimental outcomes on a data set are much greater than most news video experiments in the literature. It has been demonstrated that both the shot boundary detection method and the anchorperson detection method are well organized.

Zhai and Shah [53] have proposed a general framework for temporal scene segmentation. Their proposed method is formulated in a statistical fashion and utilizes the Markov chain Monte Carlo (MCMC) technique to determine the boundaries between video scenes. In their approach, a set of arbitrary scene boundaries are initialized at random locations and are automatically updated using two types of updates: diffusion and jumps. Diffusion is the process of updating the boundaries between adjacent scenes. Jumps consist of two reversible operations: the merging of two scenes and the splitting of an existing scene. The posterior probability of the target distribution of the number of scenes and their corresponding boundary locations is computed based on the model priors and the data likelihood. They have tested their proposed method on two video domains: home videos and feature films. This has provided them with accurate results.

To support content-based video analysis, modeling, representation, summarization, indexing and access, several practical algorithms were proposed by Fan et al. [54]. Initially, a multilevel video database model is given and it affords a reasonable approach to bridge the gap between low-level representative features and high-level semantic concepts from a human point of view. Subsequently, some model-based video analysis techniques are proposed to detect the video shots. They have presented a technique, which can adapt the threshold for scene cut detection to the activities of variant videos or even different video shots. A seeded region aggregation and temporal tracking technique was proposed for generating the semantic video objects. The semantic video scenes can then be generated from these extracted video access units (e.g., shots and objects) according to some domain knowledge. Finally, in order to categorize video contents into a set of semantic clusters, an integrated video classification technique was developed to support more efficient multilevel video representation, summarization, indexing and access techniques.

3.3 Query processing

Dönderler et al. [55] have focused on the task of spatiotem-

poral query processing. In their approach, video clips are segmented into shots whenever the current set of relations between video objects changes, thereby helping us to determine parts of the video, where as the spatial relationships do not change at all. They have also proposed an SQL-like video query language that has the capability to handle a broad range of spatiotemporal queries. The language is rule-based in that it allows users to express spatial conditions in terms of Prolog-type predicates. Their rule-based spatiotemporal query processing strategy and query language take advantage of this segmentation technique to provide precise (fine-grained) answers to spatiotemporal video queries. Spatiotemporal query processing is carried out in three main stages: query recognition, query decomposition and query execution.

Erozel et al. [56] have described a user interface based on natural language processing (NLP) to a video database system. The video database is based on a content-based spatiotemporal video data model. The data model is focused on the semantic content which includes objects, activities and spatial properties of objects. Spatiotemporal relationships between video objects and also trajectories of moving objects can be queried with this data model. In the video database system, a natural language interface allows to make possible querying. The queries, which are given as English sentences, are parsed using Link Parser. By using the proper information extraction techniques, the semantic representations of the queries are extracted from their syntactic structures. Those extracted semantic representations are used to call the related parts of the underlying video database system to return the results of the queries. Apart from the exact matches the similar objects and activities are also returned from the database with the help of the conceptual ontology module. This module is implemented using a distance-based method of semantic similarity search on the semantic domain-independent ontology, WordNet.

Smeaton et al. [57] have concentrated on the application of direct content access to video where user queries are matched directly against video content. They have presented three example systems which demonstrate differing approaches to content-based video search, each developed in the context of a large scale worldwide benchmarking activity where dozens of video indexing and retrieval systems are benchmarked on the same video dataset.

Schonfeld and Lelescu [58] have proposed modified method to multiple objects tracking from compressed multimedia databases. Their method have been proposed aimed at operating in distributed environment, where users initiate video searches and retrieve relevant video information concurrently from multiple compressed video archives. The proposed approach tracks the region of interest and finds the positions in the image. This leads to more complexity of query formulations in terms of the relative positions of the target objects in the image. In order to track the objects in the video bit stream, motion vectors such as analysis and motion information is used. Based on the request, the query related video sequence is displayed as soon as the system completes the search.

Affendey et al. [17] have presented a video data model that would permit users to create hybrid queries on different attributes of video. The video data model receives the hierarchical structure of video (sequence, scene, shot and key frame), as well as high-level concepts (object, activity and event) and low-level visual features (color, texture, shape and location). With this illustration, queries for the content and/or the specific hierarchical structure by similarity-based matching of low-level visual features as well as exact matching of textual attributes are supported. Experiments to evaluate query formulation using single types with hybrid query proved that hybrid query provides exact results.

3.4 Retrieval

A well-organized and valuable retrieval model for handling visual and multimedia digital libraries has been presented by Vrochidis et al. [59]. Their proposed retrieval model has adopted three methods for retrieval: two autonomous and one combinational. The ontology-based method utilize the formal, logic-based representation of semantic mark-up metadata accompanying each collection, while an illustrative user interface is used for graphical query information. Their method was suitable only when the user is interested in semantically similar outcomes. The low level visual characteristic of the multimedia material is utilized in the content based method so as to retrieve items with similar appearance. Though the search engine has deal with 2D images of cultural heritage content, there is the potential of extension based on their proposed model to include video content. A noteworthy feature of their work is its modular and extensible ontology infrastructure, which has provided mappings to CIDOC-CRM with an intention to achieve operability with other ontologies from the cultural domain. The hybrid method has pooled the use of preceding two methods. Thus is presenting a complete outcome set to the user, which comprised both visually and semantically similar items, where as the input query is either ontology based or content based.

Wen et al. [60] have applied the technology of moving-

object tracking to content base video retrieval. In order to overcome the shadow problem and to detect the moving pixels they have used the background subtraction technique. They have also used the connected components labeling and morphological operations to eliminate noise and mend the moving pixels. Then by using color histograms, color similarity and "motion vector", the target's image and information for content base video retrieval in the database have been extracted. Their method can be applied to image frame retrieval in single-CCD (Charge Coupled Device) or multi-CCD surveillance systems. For multi-surveillance retrieval, abrupt environment changes (such as light), CCD shift, viewing angle and position (it will cause same object with different size) will cause detection and retrieval error.

Lili et al. [61] have presented a framework architecture that utilized the multimodality features as the knowledge representation scheme in order to model the behaviors of a number of human actions in the video scenes. The main focus of their work placed on the design of two main components (model classifier and inference engine) for a tool abbreviated as video action scene detector (VSAD) for retrieving and detecting human actions from video scenes. They initiated their work by presenting the workflow of the retrieving and detection process and the automated model classifier construction logic. Then they proceeded on to demonstrate how the constructed classifiers can be used with multimodality features for detecting human actions. As a final point, behavioral explanation manifestation was discussed.

Hoi and Lyu [62] have proposed a multimodal and multilevel (MMML) ranking framework to attack the challenging ranking problem of content-based video retrieval. They have represented the video retrieval task by graphs and suggests a graph based semi-supervised ranking (SSR) scheme, which can be learnt with small samples in effect and integrate multimodal resources for ranking smoothly. They have proposed a multilevel ranking framework that unifies several different ranking approaches in a cascade fashion so as to build the semi-supervised ranking solution practical for large-scale retrieval tasks. They also conducted empirical evaluations of their proposed solution for the purpose of automatic search tasks on the benchmark test bed of TRECVID2005.

Tjondronegoro and Chen [63] have found that successful content-based video data management systems depend on three most important components: key-segments extraction, content descriptions and video retrieval. The computer can understand more accurately the content of the video type by identifying the typical events that happens just before or after the key-segments (specific-domain approach), but it cannot perceive the content of the video in order to identify the key segments. Thus, they have proposed a concept of customizable video segmentation module, which integrates the suitable segmentation techniques for the current type of video. The identified key-segments are then described using standard video content descriptions to enable content-based retrievals. For retrieval, they have implemented XQuery, which is the most recent XML query language and the most powerful compared to older languages, such as XQL and XML-QL.

Ma and Zhang [64] have presented a generic motion representation, named motion texture. In this presentation most of the important motion characters are preserved. Based on such representation, they have improved the performance of motion-based video retrieval as well as they have devised a semantic classification scheme by which the motion patterns can be mapped to semantic conceptions. Experimental results have indicated that motion texture is a compact, generic and effective representation of a motion pattern.

A technique for implementing video search functions (retrieval of near-duplicate videos and identifying actions) in surveillance video is illustrated by DeMenthon and Doermann [65]. Videos are alienated into half-second clips whose stacked frames generate 3D space-time volumes of pixels. Pixel areas with reliable properties of color and motion are extracted from these 3D volumes by a threshold-free hierarchical space-time segmentation technique. Each area is then illustrated by a high-dimensional point whose components stand for the position, orientation and color of the region, when possible. These points are assigned labels that stipulate their video clip of origin in the indexing phase for a video database. All the labeled point for each clip is accumulated into a single binary tree for effective k-nearest neighbor retrieval. The retrieval phase utilizes video segments as queries. Half-second clips of these queries are fragmented again by space-time segmentation in order to generate sets of points and for each point the labels of its nearby neighbors are retrieved. The labels that obtain the highest numbers of votes match up to the database clips that are nearly similar to query video segment. Initially, they have explained the retrieval test for dynamic logos and for video queries that vary from the indexed broadcasts by the adding up huge overlays. Also they illustrated experimentation in recognizing office actions (such as pulling and closing drawers, taking and storing items, picking up and putting down a phone).

Yang et al. [66] have proposed an integrated frame work for the retrieval of flash movies. Two major components are involved in their approach: (1) a content-based retrieval component, which explored the characteristics of Flash movie

content at compositional and semantic levels; and (2) a context-based retrieval component, which explored the contextual information including the texts and hyperlinks surrounding the movies. In order to explain the possibility of the proposed framework an experimental Flash search engine system has been implemented.

Dagtas et al. [67] have offered the models that utilized the object motion information so as to characterize the events to allow subsequent retrieval. By using various signal and image processing techniques, algorithms for different spatiotemporal search cases in terms of spatial and temporal translation and scale invariance have been developed. They have developed a prototype video search engine, pictorial information and content transformation unified retrieval engine for spatiotemporal queries (PICTURESQUE) to verify their proposed methods.

3.5 Indexing and retrieval

Lu et al. [68] have addressed the problem of content-based video indexing and proposed an efficient solution, called the ordered VA-file (OVA-File) based on the VA-file. OVA-File is a hierarchical structure and has two novel features: (1) partitioning the whole file into slices such that only a small number of slices are accessed and checked during k nearest neighbor (kNN) search and (2) efficient handling of insertions of new vectors into the OVA-File, such that the average distance between the new vectors and those approximations near that position is minimized. In order to assist the search, they have presented an efficient approximate kNN algorithm named ordered VA-LOW (OVA-LOW) based on the proposed OVA-File. Extensive experimental studies using real video data sets were conducted and the results showed that their methods yielded a significant speed-up over an existing VA-file-based method and iDistance with high query result quality.

Zhang and Nunamaker [69] have presented an interactive multimedia-based e-Learning environment that enables users to interact with it to obtain knowledge in the form of logically segmented video clips. They have developed a two-phase approach to conduct content-based video indexing and retrieval to identify video clips appropriate to addressing users interests. Their approach integrates natural language processing, named entity extraction, text and frame based video indexing and information retrieval techniques. The relevance of video clips to questions is measured based on the similarity between generated templates of questions and clip content. Their research explores a way to access instructional videos in interactive e-Learning. Some results have shown that their approach has achieved advanced precision than the traditional keyword-based approach.

Video retrieval and browsing is a complicated task. Most prior work in the field is roughly categorized into two classes. One is based on image processing technique, often called content-based image and video retrieval, in which video frames are indexed and searched for visual content. The other is based on spoken document retrieval, which relies on automatic speech recognition and text queries. Both approaches have major limitations. In the first approach, semantic queries pose a great challenge, while the second, speech-based approach and do not support efficient video browsing. Amir et al. [70] have described a system that groups the advantages of both the approaches where speech is used for efficient searching and visual data for efficient browsing. A fully automatic indexing and retrieval system has been developed and tested. Automated speech recognition and phonetic speech indexing support text-to-speech queries. From the original video new browsable views are generated. A special synchronized browser allows instantaneous, context-preserving switching from one view to another. The system was successfully used to produce searchable-browsable video proceedings.

Chiu et al. [71] proposed a framework for constructing a content-based human motion retrieval system. The major components such as, indexing and matching are discussed along with their corresponding algorithm. According to the distribution of the raw data, in indexing they have proposed affine invariant posture representation and proposed a SOM-based (Self-Organizing Map) index map. In order to find the clips from the given motion collection, the start and end frames are used. Then the similarity between the query example and each candidate clip is computed by using the dynamic time warping algorithm. The high-dimension feature space of the entire skeleton are decomposed into the direct sum of low-dimension feature spaces of skeletal segments so as to avoid the curse of dimensionality. Several experimental results have shown that their proposed indexing method have performed well than the conventional fixed-grid and k-d tree methods in the aspects of retrieval accuracy and computation time.

An adaptive video indexing (AVI) technique based on a template-frequency model, together with self-training retrieval architecture, was proposed by Munesawang and Guan [72], so as to allow full use of temporal information. For relevance feedback analysis of the dynamic content of video data, AVI considers spatiotemporal information also. The AVI indexing method is effectively adapted for video shot, scene and story queries to facilitate multiple-level access to a video

database. Through its signal propagation process, the indexing structure is built-in by the system to a self-training neural network which has implemented automatic adaptive retrieval. As the relevance feedback is implemented in automatic and semi-automatic fashions, the search time for video transmissions is greatly reduced over the Web. In order to achieve a user-friendly environment, the AVI structure works well both in fully automatic mode and the user-interaction interface system. Experimentally for the retrieval of CNN news videos, they have demonstrated their proposed indexing method and automatic relevance feedback.

An original approach for content-based video indexing and retrieval has been described. Fablet et al. [73] have aimed at providing a global interpretation of the dynamic content of video shots without any prior motion segmentation and without any use of dense optic flow fields. In order to complete they depend on statistical modeling of the distribution of local motion-related measurements using nonparametric causal, Gibbs distribution fitted at the maximum likelihood (ML) sense. To be free from camera movement, they have considered an efficient model complexity reduction scheme based on likelihood ratios. Thus with the help of this property they are able to develop a general statistical framework for video indexing and retrieval with query-by-example. They have built a hierarchical structure of the processed video database according to motion content similarity. This results in a binary tree where each node is associated to an estimated causal Gibbs model.

Yi et al. [74] introduced a technique for motion indexing in which the formation of a pixel change ratio map (PCRM) is based on the motion content in a video sequence which also mines the motion histogram from the same. Bins that are present in motion histogram are adaptively quantized in order to encompass a higher discriminating power amidst of various classes. They have illustrated the effectiveness of motion histogram technique by means of several applications in video retrieval, video clustering and video classification. It not only retrieves sequences containing similar motion as the query sequence, but it is also provides an indication of the size of the moving objects. The clusters formed by using motion histogram as characteristics are extremely similar in motion content to what is perceived by the human visual system. Through their proposed motion histogram technique, categorization of video sequences ends up in a higher classification rate.

Babu and Ramakrishnan [75] have presented an object-based video indexing and retrieval system using the motion information acquired from the compressed MPEG video. The major contribution of their proposed system was the usefulness of the motion information of MPEG video for global and object-based retrieval of generic video that are readily available. The retrieval results are nearer to the user expectation. The computational burden is reduced to a very large extent as the sparse motion vectors are used for the extracting features. The number of objects present in the video shot is determined by establishing the proposed K-means algorithm on the refined motion data and the object features are acquired by segmenting the objects by EM algorithm. In order to retrieve the video files both global and object characteristics are combined with the user given weights.

Snoek et al. [76] have proposed a video retrieval. Initially, a large lexicon of semantic concepts has been detected. They have combined query-by-concept, query-by-example, query-by-keyword, and user interaction into the MediaMill semantic video search engine. The measurement of the impact of increasing lexicon size on interactive video retrieval performance, they have performed two experiments against the 2004 and 2005 NIST TRECVID benchmarks, using lexicons containing 32 and 101 concepts, respectively. The results have been indicated that as of all factors that play a role in interactive retrieval, a large lexicon of semantic concepts matters most. In fact by using large lexicons, lot of video search questions was solvable without using query-by-keyword and query-by-example.

Abdelali et al. [77] have focused on hardware implementation of content based video indexing techniques by using the field programmable gate arrays (FPGA) technology. Their goal is to offer hardware modules that can satisfy requirements of applications under tough constraints, such as real time applications and applications of high complexity. They offered two instances of micro-architectures related to the colors descriptor and they are, dominant and compact. The synthesis and simulation results for the provided solutions are also given in their work.

Doulamis et al. [78] presented a fuzzy representation of visual content. Their proposed method is very much beneficial for multimedia applications, such as content-based image indexing and retrieval, video browsing and summarization. Video sequence analysis techniques are used to collect the features and based on the collection of these features, multi-dimensional fuzzy histogram is constructed for each video frame. Their approach has been implemented on both for video summarization, in the context of a content-based sampling algorithm and for content-based indexing and retrieval. In the first case only a small but a meaningful amount of information is retained. This is achieved by removing shots or

frames of similar visual content. In the second case, similar images of the query are extracted and this is obtained by employing content-based retrieval scheme. The experimental results revealed that the proposed scheme on real life video recordings performed well when compared to that of the other methods.

An approach that is content-based multi-resolution indexing technique was presented. For the content storage and retrieval, histogram is considered to be a more proficient feature. In order to define index for the scene Lee and Dickinson [79] have utilized two representative frames of a scene. Since they used a temporally vague query image their proposed method overcomes the difficulty of retrieving a scene and the multi-resolution matching leads to the benefit of allowing some degree of vagueness in user queries. Their experiment outcomes revealed that difference of histograms (DOH) measure is effective frame distance measure for scene-change detection among the four measures compared. Due to the hierarchical search procedure the proposed multiresolution video indexing technique for subband-coded video databases is more efficient and effective.

Fan et al. [80] presented a content based video classification method so as to support semantic categorization, high dimensional indexing and multilevel access. Their contributions are in four points: (1) to begin with they have presented a hierarchical video database model that captures the structures and semantics of video contents in databases; (2) secondly they have proposed a set of useful techniques for exploiting the basic units (e.g., shots or objects) in order to access the videos in database; (3) the third one is that they have suggested a learning-based semantic classification technique in order to exploit the structures and semantics of video contents in the database; (4) and the fourth one is the cluster-based indexing structure used to speed-up query-by-example and organizes a database that helps for effective and efficient browsing. The applications of their proposed multilevel video database representation and indexing structures for MPEG-7 were also discussed.

Albanese et al. [81] have proposed video segmentation algorithm and also a formal model for the video segmentation process. Their proposed approach was able to detect the mathematical characterization of the most common transition effects; more over it is also based on the computation of an arbitrary similarity measure between consecutive frames of a video. The algorithm has been experimented adopting a similarity metric based on the Animate Vision theory and results have been reported.

Cheung [82] has proposed the implementation of an intelligent video database using evolutionary control. Hence the performance of video retrieval can be done by utilizing the automatic video indexing techniques. He has implemented an automatic video indexing system (AVIS) using information retrieval and machine learning techniques. His system was experimented using the original movie scripts from "Star Wars – return of the JEDI" (139 movie segments) and "Star Wars – A new hope" (476 movie segments). The experimental outcomes revealed that the information retrieval and machine learning techniques can be implemented to video information systems successfully.

3.6 Other researches

Erol and Kossentini [83] have proposed two local motion descriptors for the retrieval of video objects. The level of rank obtained after utilizing their descriptors intimately match with the human ranking. According to the average normalized modified retrieval rank (ANMRR) scores obtained, the angular circular local motion (ACLM) descriptor offered a better retrieval rate than the angular radial transform (ART)-based descriptor. Given that each descriptor value is quantized to [0–255] range, ACLM descriptor requires 16 bytes and the ART-based descriptor requires 8 bytes to represent. ACLM descriptor is computationally less complex to extract. Nevertheless, if the ART coefficients of the video object is already computed and attached to the video objects as metadata for shape retrieval, then the extra computations required to extract the local motion descriptors based on the ART coefficients are minima. Based on their application, anyone of their proposed descriptors could be used for the video retrieval by local motion.

Xu et al. [84] have discussed that their research achievement on semantics extraction and automatic editorial content creation and adaptation in sports video analysis. They have proposed a generic multi-layer and multi-modal framework for sports video analysis and they introduced several mid-level audio/visual features which are able to bridge the semantic gap between low-level features and high-level understanding. They have also discussed emerging applications on editorial content creation and content enhancement/adaptation in sports video analysis, including event detection, sports MTV generation, automatic broadcast video generation, tactic analysis, player action recognition, virtual content insertion and mobile sports video adaptation.

Lee and Yoo [85] proposed video fingerprinting method based on the centroid of gradient orientations. The centroid of gradient orientations is chosen due to its pairwise indepen-

dence and robustness against common video processing steps that include lossy compression, resizing, frame rate change, and more. A threshold used to reliably determine a fingerprint match is theoretically derived by modeling their proposed fingerprint as a stationary ergodic process and the validity of the model is experimentally verified. Based on the performance the proposed finger print was experimentally evaluated and compared with that of other widely-used features. Their experimental results have showed that the proposed fingerprint outperformed the considered features in the context of video fingerprinting.

In order to build an advanced superior video database indexing and access, Fan et al. [86] have proposed a framework, called Class View. To reduce the semantic gap, a hierarchical semantics-sensitive video classifier has been proposed. This hierarchical structure of the semantics-sensitive video classifier was derived from the domain-dependent concept hierarchy of video contents present in a database. Relevance analysis is used for choosing the discerning visual characteristics with appropriate significance. In addition, expectation-maximization (EM) algorithm is used to identify the classification rule for every visual concept node present in the classifier. A hierarchical video database indexing and summary presentation technique was proposed to support extra effectual video access over a large-scale database. The innate hierarchical video database indexing tree structure is combined with presentation of visual summary. Combining video access with a well-organized database indexing tree structure provides huge opportunities for supporting more powerful video search engines.

Fan et al. [87] have proposed a framework to make some advances towards the final goal to solve (a) semantic gap; (b) semantic video concept modeling; (c) semantic video classification; (d) concept-oriented video database indexing and access problems. The framework especially includes: (1) a semantic-sensitive video content representation framework by using major video shots to improve the excellence of features; (2) semantic video concept interpretation by using flexible mixture model to viaduct the semantic gap; (3) a semantic video classifier training framework by combining feature selection, parameter estimation and model selection flawlessly in a single algorithm; (4) a concept-oriented video database organization technique through a certain domain-dependent concept hierarchy to facilitate semantic sensitive video retrieval and browsing.

The main goal of Khan et al. [88] was the prediction of video quality integrating the application and network level parameters for all content types. Initially, video sequences are classified into groups representing various content types using cluster analysis. The categorization of contents is based on the characteristics of temporal (movement) and spatial (edges, brightness) extraction. Subsequently, the manners of video quality for wide range variations of a set of elected parameters are learned and examined. At last, to increase two learning models based on (1) ANFIS to estimate the visual perceptual quality in terms of the mean opinion score (MOS) and decodable frame rate (Q value) and (2) regression modeling to calculate the visual perceptual quality in terms of the MOS. Primary results have proved that excellent prediction accuracy was acquired from both models. However, the regression based model performed better in terms of the correlation coefficient and the root mean squared error.

Dumont and Merialdo [89] have presented a technique to summarize rushes video based on the detection of repetitive sequences, with the aid of the smith-waterman algorithm to determine the matching sequences. They have relied on the evaluation technology that has been introduced in the TRECVID BBC rushes summarization task. They have proposed an automation of the manual TRECVID evaluation using the machine learning techniques to train an automatic assessor.

Thyagharajan and Ramachandran [90] have discussed the segmentation techniques for creating video summary and they have proposed a hierarchical scheme, which has decomposed a video sequence into different content-resolution levels to improve the transmission and user interaction. Mathematical models have been derived in order to represent the video structure. Therefore, streaming parameters such as bandwidth, buffer requirements and initial delay are estimated for each segment at different stages to present a jitter-free playback. The algorithm developed in their work can be integrated to a video encoder to calculate the streaming parameters and the result can be provided to a streaming server to produce an effectual transmission schedule.

4 Performance review

CBVIR have a wide range of applications such as quick browsing of video folders, analysis of video frames. Here, focusing on automating the indexing, retrieval and management of video and each changes in a multimedia database management system, mostly in the modeling and querying techniques. The performance of content-based video indexing and retrieval based performance results is shown in below Table 1.

Table 1 An overview of different feature extraction methods performance results

Feature extraction of CBVIR	Recalls /%	Precision /%
Chang et al. [43]	96.6	96.6
Kolekar et al. [44]	90.83	86.57
Basharat et al. [46]	68	70
Kuo and Chen [47]	98	96
Chen et al. [48]	91	94
Duan et al. [49]	85	95
Liu et al. [50]	97.8	92
Han et al. [51]	97	95.4
Gao and Tang [52]	97.25	97.64
Zhai and Shah [53]	84	91.3
Erozel et al. [56]	88	95
Lili et al. [61]	88.93	98.01
Munesawang and Guan [72]	62.34	92.03

As can be seen from Table 1, Zhang and Chen [42] have presented a technique based on the stICA and a multiscale analysis system and they get improve the accuracy of the extracted object. Chang et al. [43] have presented clustering strategy using clustering ensembles for video shot detection and they achieve Clustering ensemble recalls 96.6% and Precision 90.6%. Kolekar et al. [44] have used automated indexing and semantic labeling for broadcast soccer video sequences and they achieved average recall 90.83% and precision 86.57%. Jiang and Crookes [45] proposed fast automation motion segmentation method and they reduce complexity and improves performance.

Basharat et al. [46] have proposed a framework for matching video sequences using the spatiotemporal segmentation and they achieved average recall 68% and precision 70%. Kuo and Chen [47] have proposed a technique an approach for segmenting videos into shots on MPEG coded video data and them given an average 98% recall and 96% precision of the detection. Chen et al. [48] have proposed a mechanism to automatically parse sports videos in compressed domain and they get an average result of 91% recall and 94% precision of this technique. Duan et al. [49] have proposed a unified framework for semantic shot classification and they have achieved good accuracy of recall 85% and precision 95%. Liu et al. [50] have designed multiple-modality method for extracting semantic information from basketball video and they get achieved of recall 97.8% and precision 92.0%. Han et al. [51] have been proposed a rough-fuzzy operator based on the rough-fuzzy sets for feature reduction and the dissimilarity function and result of 97.0% recall and 95.4% precision were achieved. Gao and Tang [52] have presented an unsupervised video-shot segmentation method and a model-free anchorperson detection scheme and they achieve a precision of 97.64% and recall of 97.25% for anchorperson shot detection. Zhai and Shah [53] have proposed a general statistical framework for the temporal scene segmentation of videos and achieved a precision of 91.3 % and recall of 84 % for equally important. Fan [54] have integrated content-based video retrieving and browsing approach, called MultiView and they can be supported by our multilevel video representation and indexing structures.

Donderler et al. [55] have proposed an architecture for a Web-based video database management system (VDBMS) providing an integrated support and semantic queries and they get better results in query recognition, query decomposition, and query execution. Erozel et al. [56] have described a user interface based on natural language processing (NLP) to a video database system they get a recall of 88% and precision of 95%. Smeaton et al. [57] have concentrated on the application of direct content access to video and they give TRECVid benchmarking activity. Schonfeld and Lelescu [58] have proposed modified method to multiple objects tracking from compressed multimedia databases and this system may decompress and display the query-relevant video sequences upon request. Vrochidis et al. [59] have developed a model for handling visual and multimedia digital libraries is presented in an efficient and effective manner and they get insights into its performance.

Wen et al. [60] have applied the technology of moving-object tracking to content base video retrieval and they get parameters according to the environmental condition. Lili et al. [61] have presented an approach to characterize and abstract human activity to support high level video indexing and they get recall of 88.93% and precision of 98.01%. Hoi and Lyu [62] have proposed a MMML ranking framework to attack the challenging ranking problem of content-based video retrieval and they get ranking solutions are effective and very competitive with the state-of-the-art solutions in the TRECVID evaluations. Tjondronegoro and Chen [63] multimedia technology enable ease of capturing and encoding digital video and they get new XML query language and compared to older languages, such as XQL and XML-QL. Ma and Zhang [64] have presented a generic motion representation, named motion texture and they get motion texture compact, generic, and effective representation. Yang et al. [66] have investigated the problem of web-based Flash retrieval and they get better Flash search engine system. Dagtas et al. [67] have proposed two complementary models for motion-based video characterization that lead to an effective content-based retrieval mechanism and they get the signal-to-noise ratio (PSNR) technique in traditional signal processing. Lu

et al. [68] have proposed index structure, OVA-File, for fast similarity query in typical multimedia applications and they achieved much more efficient performance. Zhang and Nunamaker [69] have presented an interactive multimedia-based e-Learning environment and they get precision and recall of this approach are better than those of the traditional keyword based approach. Chiu et al. [71] have proposed a novel framework for constructing a content-based human motion retrieval system. Munesawang and Guan [72] have presented an AVI model and integrated with an adaptive signal propagation network and they were achieved at 92.03% precision and recall 62.34%. Fablet et al. [73] have a global characterization of motion content in video sequences and they obtained promising results on a set of various real image sequences. Yi et al. [74] have introduced motion content of the video at pixel level, is represented as a pixel change ratio map (PCRM) and they demonstrated the usefulness of the motion histogram with three applications, viz., video retrieval, video clustering and video classification. Babu and Ramakrishnan [75] have presented an object-based video indexing and retrieval system using the motion information acquired from the compressed MPEG video. Doulamis et al. [78] presented fuzzy representation of visual content and they presented to indicate better performance of the proposed scheme on real-life video recordings. Lee and Dickinson [79] have proposed an indexing technique appears to be promising for its speed and its inherent hierarchical search procedure. Fan et al. [80] develop a content-based video classification approach to support the semantic categorization, high-dimensional indexing and multilevel access and achieved multi-level video database representation and indexing structures for MPEG-7. Albanese et al. [81] have proposed video segmentation algorithm and also a formal model for the video segmentation process.

Erol and Kossentini [83] have proposed two local motion descriptors for the retrieval of video objects and they show that ranking obtained by querying with descriptors closely match with the human ranking. Xu et al. [84] have discussed on semantics extraction and automatic editorial content creation and adaptation in sports video analysis. Lee and Yoo [85] proposed video fingerprinting method based on the centroid of gradient orientations and showed that the proposed fingerprint outperformed the considered features in the context of video fingerprinting. Fan et al. [86] have proposed a framework, called Class View and they get provided a great opportunity for supporting more powerful video search engines. Fan et al. [87] have used framework to make some advances toward the final goal to solve these problems and they improved the classification accuracy significantly. Khan et al. [88] have used video quality combining the application and network level parameters for all content types and they get development of a reference-free video prediction model and quality of service (QoS) control methods. Snoek et al. [76] have proposed a video retrieval for semantic concepts. and they the lexicon-driven search engine outperforms all state-of-the-art video retrieval systems in both TRECVID 2004 and 2005.

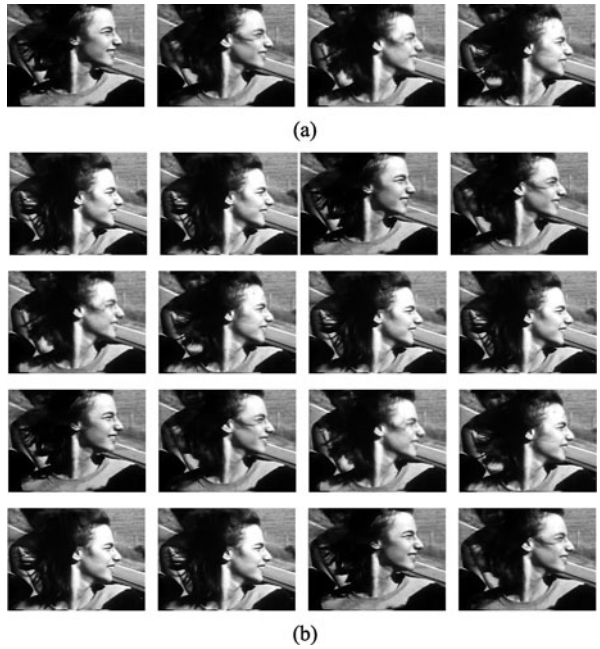

Fig. 3 (a) Query image; (b) retrieved frames of video input

5 Directions for the future research

In this review paper, various techniques utilized for content based video retrieval and indexing has been analyzed thoroughly. Different methods of feature extraction, shot detection, query processing and retrieval and indexing researches are reviewed along with their different features. News Broadcasting, Advertising, Music video clips, distant learning video archiving and medical applications are some of the different areas in which the CBVIR has been utilized and hence it have turned out to be popular. Because the utilization of different video databases, it has attained the better results in the research community. At the end of this review, we conclude that very few works are available in audio based video retrieval. In the future, we expect a copious amount of inventive brainwaves will rise by means of our review work. This paper will be a healthier foundation for the budding researchers in the field of content based video indexing and retrieval.

6 Conclusion

Over the past decade, content-based indexing and retrieval of visual information has been an increasing research area which has received immense response in the research community. Efficient use of multimedia materials needs an efficient way to support it in order to browse and retrieve it. The techniques of user annotation involved extra challenges by human, so automatic video shot annotation, indexing and retrieval of visual information is in great need. In this paper, we have presented an extensive review of the significant researches in existence for content based video indexing and retrieval system of visual information. An introduction to content based video analysis has been presented. A brief description about CBVIR and its application is also presented. The motive of this research survey is to help the budding researchers in the field of content based video indexing and retrieval to understand the available methods and to assist their further research.

References

1. Hanis A, Sziranyi T. Measuring the motion similarity in video indexing. In: Proceedings of the 4th EURASIP Conference Focused on Video/Image Processing and Multimedia Communications. 2003, 507–512
2. Calic J, Izuierdo E. Efficient key-frame extraction and video analysis. In: Proceedings of the 2002 International Conference on Information Technology: Coding and Computing. 2002, 28–33
3. Carbonaro A. Ontology-based video retrieval in a semantic-based learning environment. Journal of e-Learning and Knowledge Society, 2009, 4(3): 203–212
4. George A, Rajakumar B, Suresh B. Markov random field based image restoration with aid of local and global features. International Journal of Computer Applications, 2012, 48(8): 23–28
5. Kundra E H, Verma E M, Aashima E. Filter for removal of impulse noise by using fuzzy logic. International Journal of Image Processing (IJIP), 2011, 3(5): 195–202
6. Umamakeswari A, Rajaraman A. Object based video analysis, interpretation and tracking. Journal of Computer Science, 2007, 3(10): 818–822
7. Amer A. Object-based video retrieval based on motion analysis and description. Technical Report, University du Québec, 1999
8. Javed O, Shah M, Comaniciu D. A probabilistic framework for object recognition in video. In: Proceedings of the 2004 International Conference on Image Processing. 2004, 2713–2716
9. Radhakrishnan R, Divakaran A, Xiong Z, Otsuka I. A content-adaptive analysis and representation framework for audio event discovery from unscripted multimedia. EURASIP Journal on Applied Signal Processing, 2006: 1–24
10. Schnettler B, Raab J. Interpretative visual analysis developments: state of the art and pending problems. Historical Social Research/Historische Sozialforschung, 2009, 265–295
11. Ramoser H, Schlogl T, Beleznai C, Winter M, Bischof H. Shape-based detection of humans for video surveillance applications. In: Proceedings of the 2003 IEEE International Conference on Image Processing. 2003, 3: 1013–1016
12. Ahmad A M, Lee S Y. Fast and robust object-extraction framework for object-based streaming system. International Journal of Virtual Technology and Multimedia, 2008, 1(1): 39–60
13. Ngo C W, Pong T C, Zhang H J. Motion-based video representation for scene change detection. International Journal of Computer Vision, 2002, 50(2): 127–142
14. Zelnik-Manor L, Irani M. Event-based analysis of video. In: Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition. 2001, II-123–II-130
15. Ding Y, Fan G. Camera view-based american football video analysis. In: Proceedings of the 8th IEEE International Symposium on Multimedia. 2006, 317–322
16. Mohan C K, Dhananjaya N, Yegnanarayana B. Video shot segmentation using late fusion technique. In: Proceedings of the 7th International Conference on Machine Learning and Applications. 2008, 267–270
17. Affendey L S, Mamat A, Ibrahim H, Ahmad F. Video data modelling to support hybrid query. International Journal of Computer Science and Network Security, 2007, 7(9): 53–61
18. Aytar Y, Shah M, Luo J. Utilizing semantic word similarity measures for video retrieval. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. 2008, 1–8
19. Dimitrova N, Zhang H J, Shahraray B, Sezan I, Huang T, Zakhor A. Applications of video-content analysis and retrieval. IEEE MultiMedia, 2002, 9(3): 42–55
20. Bergman L D, Castelli V, Li C. Progressive content-based retrieval from satellite image archives. Technical Report, D-Lib Magazine, 1997
21. Chang S F, Smith J R, Meng H J, Wang H, Zhong D. Finding images/video in large archives. Technical Report, D-Lib Magazine, 1997
22. Gupta A, Jain R. Visual information retrieval. Communications of the ACM, 1997, 40(5): 70–79
23. Papadias D, Mantzourogiannis M, Kalnis P, Mamoulis N, Ahmad I. Content-based retrieval using heuristic search. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1999, 168–175
24. Koprinska I, Carrato S. Temporal video segmentation: a survey. Signal Processing: Image Communication, 2001, 16(5): 477–500
25. Hanjalic A. Shot-boundary detection: unraveled and resolved? IEEE Transactions on Circuits and Systems for Video Technology, 2002, 12(2): 90–105
26. Girgensohn A, Boreczky J. Time-constrained keyframe selection technique. In: Proceedings of the 1999 IEEE International Conference on Multimedia Computing and Systems. 1999, 756–761
27. Liu T, Kender J R. Optimization algorithms for the selection of key frame sequences of variable length. In: Proceedings of the 2002 European Conference on Computer Vision. 2002, 403–417
28. Aslandogan Y A, Yu C T. Techniques and systems for image and video retrieval. IEEE Transactions on Knowledge and Data Engineering, 1999, 11(1): 56–63

29. Lu G. Techniques and data structures for efficient multimedia retrieval based on similarity. IEEE Transactions on Multimedia, 2002, 4(3): 372–384
30. Lelescu D, Schonfeld D. Video skimming and summarization based on principal component analysis. In: Proceedings of the 4th IFIP/IEEE International Conference on Management of Multimedia on the Internet. 2001, 128–141
31. Gargi U, Kasturi R, Strayer S H. Performance characterization of video-shot-change detection methods. IEEE Transactions on Circuits and Systems for Video Technology, 2000, 10(1): 1–13
32. Huang Y, Liu Q, Metaxas D. Video object segmentation by hypergraph cut. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009, 1738–1745
33. Aydm Alatan A, Tuncel E, Onural L. A rule-based method for object segmentation in video sequences. In: Proceedings of the 1997 International Conference on Image Processing. 1997, 522–525
34. Otoom A F, Gunes H, Piccardi M. Feature extraction techniques for abandoned object classification in video surveillance. In: Proceedings of the 15th IEEE International Conference on Image Processing. 2008, 1368–1371
35. Van Cauwelaert D. Generic models for adaptive robust feature extraction in video. In: Proceedings of the 9th FirW PhD Symposium. 2008, 148–149
36. Zhong D, Chang S F. An integrated approach for content-based video object segmentation and retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 1999, 9(8): 1259–1268
37. Avrithis Y S, Doulamis A D, Doulamis N D, Kollias S D. An adaptive approach to video indexing and retrieval using fuzzy classification. In: Proceedings of VLBV. 1998
38. Idris F, Panchanathan S. Review of image and video indexing techniques. Journal of Visual Communication and Image Representation, 1997, 8(2): 146–166
39. Christel M G, Smith M A, Taylor C R, Winkler D B. Evolving video skims into useful multimedia abstractions. In: Proceedings of the 1998 SIGCHI Conference on Human factors in Computing Systems. 1998, 171–178
40. Marchand-Maillet S. Content-based video retrieval: an overview. 2000
41. Jawahar C, Chennupati B, Paluri B, Jammalamadaka N. Video Retrieval Based on Textual Queries. Technical Report, 2005
42. Zhang X P, Chen Z. An automated video object extraction system based on spatiotemporal independent component analysis and multiscale segmentation. EURASIP Journal on Applied Signal Processing, 2006, 2006: 184
43. Chang Y, Lee D J, Hong Y, Archibald J. Unsupervised video shot detection using clustering ensemble with a color global scale-invariant feature transform descriptor. Journal on Image and Video Processing, 2008, 9
44. Kolekar M H, Palaniappan K, Sengupta S, Seetharaman G. Semantic concept mining based on hierarchical event detection for soccer video indexing. Journal of Multimedia, 2009, 4(5): 298–312
45. Jiang R, Crookes D. Approach to automatic video motion segmentation. Electronics Letters, 2007, 43(18): 968–970
46. Basharat A, Zhai Y, Shah M. Content based video matching using spatiotemporal volumes. Computer Vision and Image Understanding, 2008, 110(3): 360–377
47. Kuo T C, Chen A L. A mask matching approach for video segmentation on compressed data. Information Sciences, 2002, 141(1): 169–191
48. Chen D Y, Hsiao M H, Lee S Y. Automatic closed caption detection and filtering in mpeg videos for video structuring. Journal of Information Science and Engineering, 2006, 22(5): 1145–1162
49. Duan L Y, Xu M, Tian Q, Xu C S, Jin J S. A unified framework for semantic shot classification in sports video. IEEE Transactions on Multimedia, 2005, 7(6): 1066–1083
50. Liu S, Xu M, Yi H, Chia L T, Rajan D. Multimodal semantic analysis and annotation for basketball video. EURASIP Journal on Applied Signal Processing, 2006: 182
51. Han B, Gao X, Ji H. A shot boundary detection method for news video based on rough-fuzzy sets. International Journal of Information Technology, 2005, 11(7): 101–111
52. Gao X, Tang X. Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing. IEEE Transactions on Circuits and Systems for Video Technology, 2002, 12(9): 765–776
53. Zhai Y, Shah M. Video scene segmentation using markov chain monte carlo. IEEE Transactions on Multimedia, 2006, 8(4): 686–697
54. Fan J, Aref W G, Elmagarmid A K, Hacid M S, Marzouk M S, Zhu X. Multiview: multilevel video content representation and retrieval. Journal of Electronic Imaging, 2001, 10(4): 895–908
55. Dönderler M E, Ulusoy Ö, Güdükbay U. Rule-based spatiotemporal query processing for video databases. The VLDB Journal, 2004, 13(1): 86–103
56. Erozel G, Cicekli N K, Cicekli I. Natural language querying for video databases. Information Sciences, 2008, 178(12): 2534–2552
57. Smeaton A F, Wilkins P, Worring M, De Rooij O, Chua T S, Luan H. Content-based video retrieval: three example systems from trecvid. International Journal of Imaging Systems and Technology, 2008, 18(2-3): 195–201
58. Schonfeld D, Lelescu D. Vortex: video retrieval and tracking from compressed multimedia databases—multiple object tracking from mpeg-2 bit stream. Journal of Visual Communication and Image Representation, 2000, 11(2): 154–182
59. Vrochidis S, Doulaverakis C, Gounaris A, Nidelkou E, Makris L, Kompatsiaris I. A hybrid ontology and visual-based retrieval model for cultural heritage multimedia collections. International Journal of Metadata, Semantics and Ontologies (IJMSO), 2008, 3(3): 167–182
60. Wen C Y, Chang L F, Li H H. Content based video retrieval with motion vectors and the rgb color model. Forensic Science Journal, 2007, 6(2): 1–36
61. Lili N, Noah S, Khalid F. Extracting and integrating multimodality features via multidimensional approach for video retrieval. International Journal of Computer Science and Network Security, 2009, 9(2): 252
62. Hoi S C, Lyu M R. A multimodal and multilevel ranking scheme for large-scale video retrieval. IEEE Transactions on Multimedia, 2008, 10(4): 607–619
63. Tjondronegoro D, Chen Y P P. Content-based indexing and retrieval using mpeg-7 and *x*-query in video data management systems. World Wide Web, 2002, 5(3): 207–227
64. Ma Y F, Zhang H J. Motion pattern-based video classification and retrieval. EURASIP Journal on Applied Signal Processing, 2003: 199–208

65. DeMenthon D, Doermann D. Video retrieval of near-duplicates using κ-nearest neighbor retrieval of spatio-temporal descriptors. Multimedia Tools and Applications, 2006, 30(3): 229–253
66. Yang J, Li Q, Wenyin L, Zhuang Y. Searching for flash movies on the web: a content and context based framework. World Wide Web, 2005, 8(4): 495–517
67. Dagtas S, Al-Khatib W, Ghafoor A, Kashyap R L. Models for motion-based video indexing and retrieval. IEEE Transactions on Image Processing, 2000, 9(1): 88–101
68. Lu H, Ooi B C, Shen H T, Xue X. Hierarchical indexing structure for efficient similarity search in video retrieval. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(11): 1544–1559
69. Zhang D, Nunamaker J F. A natural language approach to content-based video indexing and retrieval for interactive e-learning. IEEE Transactions on Multimedia, 2004, 6(3): 450–458
70. Amir A, Srinivasan S, Efrat A. Search the audio, browse the video—a generic paradigm for video collections. EURASIP Journal on Advances in Signal Processing, 1900, 2003(2): 209–222
71. Chiu C Y, Chao S P, Wu M Y, Yang S N, Lin H C. Content-based retrieval for human motion data. Journal of Visual Communication and Image Representation, 2004, 15(3): 446–466
72. Munesawang P, Guan L. Adaptive video indexing and automatic/semi-automatic relevance feedback. IEEE Transactions on Circuits and Systems for Video Technology, 2005, 15(8): 1032–1046
73. Fablet R, Bouthemy P, Pérez P. Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval. IEEE Transactions on Image Processing, 2002, 11(4): 393–407
74. Yi H, Rajan D, Chia L T. A new motion histogram to index motion content in video segments. Pattern Recognition Letters, 2005, 26(9): 1221–1231
75. Babu R V, Ramakrishnan K. Compressed domain video retrieval using object and global motion descriptors. Multimedia Tools and Applications, 2007, 32(1): 93–113
76. Snoek C G, Worring M, Koelma D C, Smeulders A W. A learned lexicon-driven paradigm for interactive video retrieval. IEEE Transactions on Multimedia, 2007, 9(2): 280–292
77. Abdelali A B, Mtibaa A, Bourennane E, Abid M. Design of hardware accelerators for content based video indexing. Asian Journal of Information Technology, 2006, 5(9): 976–984
78. Doulamis A D, Doulamis N D, Kollias S D. A fuzzy video content representation for video summarization and content-based retrieval. Signal Processing, 2000, 80(6): 1049–1067
79. Lee J, Dickinson B W. Hierarchical video indexing and retrieval for subband-coded video. IEEE Transactions on Circuits and Systems for Video Technology, 2000, 10(5): 824–829
80. Fan J, Zhu X, Hacid M S, Elmagarmid A K. Model-based video classification toward hierarchical representation, indexing and access. Multimedia Tools and Applications, 2002, 17(1): 97–120
81. Albanese M, Chianese A, Moscato V, Sansone L. A formal model for video shot segmentation and its application via animate vision. Multimedia Tools and Applications, 2004, 24(3): 253–272
82. Cheung R. Indexing an intelligent video database using evolutionary control. Journal of Digital Information Management, 2003, 1: 8–19
83. Erol B, Kossentini F. Retrieval by local motion. EURASIP Journal on Advances in Signal Processing, 1900, 2003(1): 41–47
84. Xu C, Cheng J, Zhang Y, Zhang Y, Lu H. Sports video analysis: semantics extraction, editorial content creation and adaptation. Journal of Multimedia, 2009, 4(2): 69–79
85. Lee S, Yoo C D. Robust video fingerprinting for content-based video identification. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(7): 983–988
86. Fan J, Elmagarmid A K, Zhu X, Aref W G, Wu L. Classview: hierarchical video shot classification, indexing, and accessing. IEEE Transactions on Multimedia, 2004, 6(1): 70–86
87. Fan J, Luo H, Elmagarmid A K. Concept-oriented indexing of video databases: toward semantic sensitive retrieval and browsing. IEEE Transactions on Image Processing, 2004, 13(7): 974–992
88. Khan A, Sun L, Ifeachor E. Content-based video quality prediction for mpeg4 video streaming over wireless networks. Journal of Multimedia, 2009, 4(4): 228–239
89. Dumont E, Merialdo B. Rushes video summarization and evaluation. Multimedia Tools and Applications, 2010, 48(1): 51–68
90. Thyagharajan K, Ramachandran V. An effective transmission and browsing methodology for streaming video. Journal of Computer Science, 2006, 2(4): 326–332

R. PRIYA received BS in Computer Science from Madras University and MS in Software Engineering from Annamalai University, in 1998 and 2001, respectively. She is currently a research scholar in Department of Mathematics, Anna University–Chennai, India. Her research interests are data mining, content based image and video retrieval.

Dr. T. N. Shanmugam is currently a professor in Department of Mathematics, Anna University, and Chennai, India. He received his BS in Mathematics from The New College, Chennai and MS from Ramanujan Institute for Advanced study Mathematics, University of Madras. He received his PhD from Anna University–Chennai in the year 1990. He has been a co-ordinator of the Video Technology Lab of Anna University–Chennai. His research interests are in complex function theory and multimedia streaming, with about eighty research papers in international journals and conferences.