Feature Extraction from Video Data for Indexing and Retrieval

[1]Ms. Amanpreet Kaur, M.Tech(CSE) , CGCTC, Jhanjeri

[2] Ms. Rimanpal Kaur, AP (CSE), CGCTC,Jhanjeri

Abstract- *In recent years, the multimedia storage grows and the cost for storing multimedia data is cheaper. So there is huge number of videos available in the video repositories. With the development of multimedia data types and available bandwidth there is huge demand of video retrieval systems, as users shift from text based retrieval systems to content based retrieval systems. Selection of extracted features play an important role in content based video retrieval regardless of video attributes being under consideration. These features are intended for selecting, indexing and ranking according to their potential interest to the user. Good features selection also allows the time and space costs of the retrieval process to be reduced. This survey reviews the interesting features that can be extracted from video data for indexing and retrieval along with similarity measurement methods.*

Keywords- Shot Boundary Detection, Key Frame Extraction, Scene Segmentation, Video Data Mining, Video Classification and Annotation, Similarity Measure, Video Retrieval, Relevance Feedback.

I. Introduction

Multimedia information systems are increasingly important with the advent of broadband networks, high-powered workstations, and compression standards. Since visual media requires large amounts of storage and processing, there is a need to efficiently index, store, and retrieve the visual information from multimedia database. Similar as image retrieval, a straightforward approach is to represent the visual contents in textual form (e.g. Keywords and attributes). These keywords serve as indices to access the associated visual data. This approach has the advantage that visual database can be accessed using standard query language like SQL; however, this entails extra storage and need a lot of manual processing. As a result, there has been a new focus on developing content-based indexing and retrieval technologies [1]. Video has both spatial and temporal dimensions and video index should capture the spatio-temporal contents of the scene. In order to achieve this, a video is first segmentation into shots, and then key frames are identified and used for indexing, retrieval.

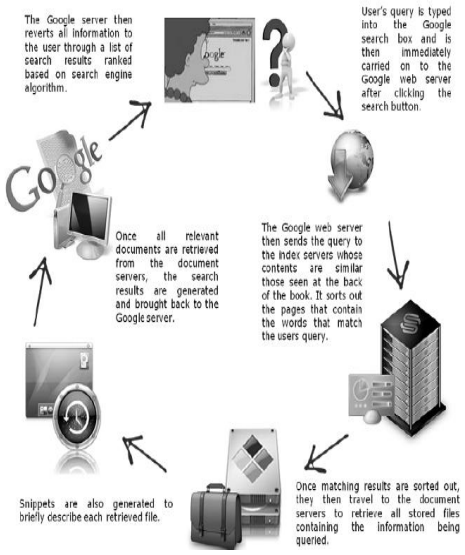

Fig 1. Working of Google search Engine

Figure 1 shows the working of Google search engine. When query is entered in the search box for searching the image, it is forwarded to the server that is connected to the internet. The server gets the URL's of the images based on the tagging of the textual word from the internet and sends them back to the client.

During recent years, methods have been developed for retrieval of videos based on their visual features [2]. Color,

texture, shape, motion and spatial-temporal composition are the most common visual features used in visual similarity match. Realizing that inexpensive storage, ubiquitous broadband Internet access, low cost digital cameras, and nimble video editing tools would result in a flood of unorganized video content; researchers have been developing video search technologies for a number of years. Video retrieval continues to be one of the most exciting and fastest growing research areas in the field of multimedia technology [1]. Content-based image retrieval (CBIR), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) is the application of computer vision to the video retrieval problem, that is, the problem of searching for video in large databases.

II. Video indexing and retrieval framework

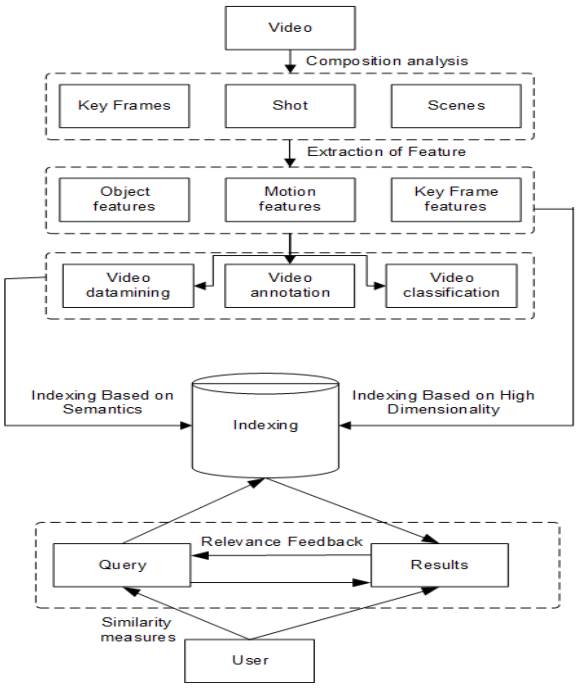

Fig 2. Video indexing and retrieval framework

Figure 2 shows the flowchart of how the individual components of the system interact.

1) structure analysis: for the detection of shot boundaries, key frame extracts, and scene fragments; 2) parts from segmented video units (scenes or stilled): it consists of the static feature in key frames, motion features and object features; 3) taking out the video data by means of extracted features; 4) video annotation: the extracted features and mined knowledge are being used for the production of a semantic index of the video. The video sequences stored within the database consists of the semantic and total index along with the high-quality video future index vector; 5) question: by the usage of index and the video parallel measures the database of the video is searched for the required videos; 6) visual browsing and response: the searched videos in response to the question are given back to the client to surf it in the form of video review, as well as the surfed material will be optimized with the related feedback.

III. Analysis Of Video Composition

Mostly, the hierarchy of video clips, scenes, shots and frames are arranged in a descending manner as shown in figure 3.

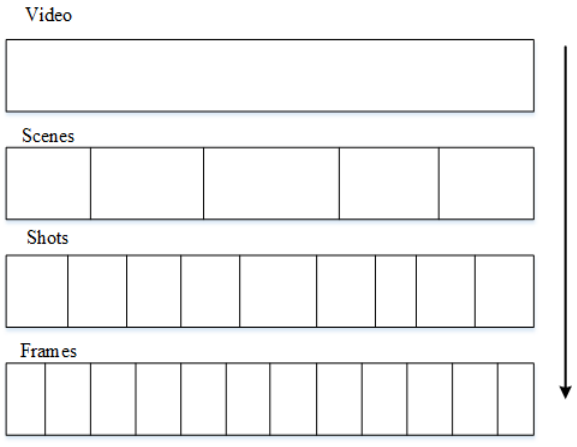

Fig 3. General Hierarchy of Video Parsing

The endeavour of the video structure analysis is segmenting a video in structural parts, which have semantic contents, segmentation of scene, and boundary detection of the shot and extraction of the key frame[3].

A. Shot Boundary Detection

Dividing the entire video into various fleeting sections is called shots. A shot may be characterized as a continuous sequence of frames created by a single non-stop camera operation. However, from the semantic point of view, its lowest level is a frame followed by shot followed by scene and, finally, the whole video. Shot boundaries are classified as cut in which the transition between successive shots is abrupt and gradual transitions which include dissolve, fade in, fade out, wipe, etc., stretching over a number of frames.

Methods for shot boundary detection usually first ex-tract visual features from each frame, then measure likeness between frames utilizing the extracted features, and, finally, detect shot boundaries between frames that are dissimilar. Frame transition parameters and frame estimation errors based on global and local features are used for boundary detection and classification. Frames are classified as no change (within shot frame), abrupt change, or gradual change frames using a multilayer perceptron network.

Shot boundary detection applications classified into two types. 1) Threshold based approach detects shot boundaries by comparing the measured pair-wise similarities between frames with a predefined threshold 2) statistical learning-based approach detects shot boundary as a classification task in which frames are classified as shot change or no shot change depending on the features that they contain.

B. Key Frame Extraction

The features used for key frame extraction include colors (particularly the color histogram), edges, shapes, optical flow. Current approaches to extract key frames are classified into six categories: sequential comparison-based, global comparison-based, reference frame-based, clustering based, curve simplification-based, and object/event-based. Sequential comparison-based approach previously extracted key frame are sequentially compared with the key frame until a frame which is very different from the key frame is obtained. Color histogram is used t find difference between the current frame & the previous key frame. Global comparison-based approaches based on global differences between frames in a shot distribute key frames by minimizing a predefined objective function. Reference frame- based Algorithms generate a reference frame and then extract key frames by comparing the frames in the shot with the reference frame.

Negative uniform evaluation method has been available for key frame extraction as a cause of the key frame subjectivity definition. In order to evaluate the rate of an error, the video compression is used for its measurements. Those key frames are favoured, which give the low error rate and high rate compression. Commonly, the low rate compression is associated with the low rate error rates. Error rates are dependable on the structures of the algorithms used for key frame extraction. Thresholds in global base comparison, frame based reference, sequential comparison based, algorithms clustering based along with that the parameters to robust the curves in the simplification based algorithms in the curve these are the examples of the parameters. The parameters are chosen by the users with that kind of error rate, which are acceptable

C. Scene Segmentation

Scene segmentation is also known as story unit segmentation. A scene is a group of contiguous shots that are coherent with a certain subject or theme. Scenes have higher level semantics than shots. Scene segmentation approaches can be classified into three categories: ocular and aural information based, key frame based, approach based on the background.

1) Ocular and aural information based: A shot boundary is selected by the following approach in which the contents of visual and acoustic change happen at the same time in the form of the boundary scene.

2) Key frame based: The following approach signifies every shot of the video in a set of key frames from which features have been taken out. In a scene, close shots along with the features are grouped temporally. The key based approach limitations are that key frames are not able to efficiently show all the dimensions of contents of the shots because in shots the scenes were usually related with the dimensions of the contents in the scene rather than in shots by frame based key similarities.

3) Background Based: The main theme about this approach is background similarity of same shots. The background base approach limitations are the hypothesis that the backgrounds are similar in the shots of the similar scene but sometimes backgrounds were different in single scene of the shots.

Current scene segmentation approaches are divisible according to the processing method, there are four categories: splitting based, merging based, shot boundary shot based classification, and model based statistics.

IV. Feature Extraction

Extracting features from the output of video segmentation. Feature extraction is the time consuming task in CBVR. This can be overcome by using the multi core architecture [4]. These mainly include features of key frames, objects, motions and audio/text features.

A. Features of Key Frames

Classified as color based, texture based and shape based features. Color-based features include color histograms, color moments, color correlograms, a mixture of Gaussian models, etc. split the image into 5×5 blocks to capture local color information [5]. Texture-based features are object surface-owned intrinsic visual features that are independent of color or intensity and reflect homogenous phenomena in images. Gabor wavelet filters is used to capture texture information for a video search engine [6]. Shape-based features that describe object shapes in the image can be extracted from object contours or regions. Edge histogram de-scriptor (EHD) is used to capture the spatial distribution of edges for the video search task in TRECVid-2005 [7]. Features based on colour: Colour histograms, a mixture of Gaussian models, colour moments, colour coral grams etc. are in the features of colour based. Colour based feature extraction are dependent on the spaces of colour for example, the HSV, RGB, YCBCR, HVC and normalized r-g and YUV.

B. Object Features

Object features include the dominant color, texture, size, etc., of the image regions corresponding to the objects. Construct a person retrieval system that is able to retrieve a ranked list of shots containing a particular person, given a query face in a shot [8]. Text-based video indexing and retrieval by, expanding the semantics of a query and using the Glimpse matching method to perform approximate matching instead of exact matching [9].

C. Motion Features

Motion features are closer to semantic concepts than static key frame features and object features. Motion-based features for video retrieval can be divided into two categories: camera-based and object-based. For camera-based features, different camera motions, such as “zoom-ing in or out,” “panning left or right,” and “tilting up or down,” are estimated and used for video indexing. Ob-ject-based motion features have attracted much more interest in recent work.

The distinguishing factor from the still images is the motion; it is the most important feature of the dynamic videos. By temporary variations, the visual content is represented by the motion information. As comparing to static key features and object features, the motion features

were near to the concepts of semantics. In the motion of the video, the motion background is added, which is formed by camera motion as well as the foreground motion this is formed by the objects which are moving. Hence, video retrieval could be divided in two categories for motion feature based they are as following: object based as well as camera based. For video indexing, the camera based features and camera motions like: the in and out zooming, left and right panning, and up or down tilting are used. The limitation of video retrieval is by using the camera based features, that the key objects motions are not describable. In modern work, a lot of interest had been grabbed by motion features of object based. Statics based, trajectory based, and spatial relationship based objects are the further categories of object based motion features

Statics Based: To model the distribution of local and global video motions, the motion's statistical features of frames points were extracted in the video. Such as, the casual Gibbs models have been used for the representation of the distribution of the spatio-temporal for the local measurements related motions, which is computed after balancing, in the original sequence, the leading motions image,

Trajectory Based: In videos, with modelling the motion trajectories of objects, the trajectory features based were extracted [99].

Relationship Based Objects: among the objects such features explains the spatial relationship.

V. Video Representation

In multilayered, iconic annotations of video content called Media Streams is developed as a visual language and a stream based representation of video data, with special attention to the issue of creating a global, reusable video archive. Top-down retrieval systems utilize high-level knowledge of the particular domain to generate appropriate representations.

Data driven representation is the standard way of extracting low-level features and deriving the corresponding representations without any prior knowledge of the related domain. A rough categorization of data-driven approaches in the literature yields two main classes [11]. The first class focuses mainly on signal-domain features, such as color histograms, shapes, textures, which characterize the low-level audiovisual content. The second class concerns annotation-based approaches which use free-text, attribute or keyword annotations to represent the content. [10] propose a strategy to generate stratification-based key frame cliques (SKCs) for video description, which are more compact and informative than frames or key frames.

VI. Mining, Classification, And Annotation

A. Video Mining

A process of finding correlations and patterns previously unknown from large video databases. The task of video data mining is, using the extracted features, to find structural patterns of video contents, behaviour patterns of moving objects, content characteristics of a scene, event patterns and their associations, and other video semantic knowledge, in order to achieve video intelligent applications, such as video retrieval. Object mining is the grouping of different instances of the same object that appears in different parts in a video. A spatial neighbourhood technique to cluster the features in the spatial domain of the frames [12]. Extract stable tracks which are combined into meaningful object clusters, used to mine similar objects [13].Special Pattern Detection applies to actions or events for which there are a priori models, such as human actions, sporting events, traffic events, or crime patterns [14].Pattern discovery is the automatic discovery of unknown patterns in videos using unsupervised or semi-supervised learning. The discovery of unknown patterns is useful to explore new data in a video set or to initialize models for further applications. Unknown patterns are typically found by clustering various feature vectors in the videos.

B. Video Classification

The task of video classification is to find rules or knowledge from videos using extracted features or mined results and then assign the videos into predefined categories. Video classification is an important way of increas-ing the efficiency of video retrieval. The semantic gap between extracted formative information, such as shape, color, and texture, and an observer's interpretation of this information, makes content-based video classification very difficult. Semantic content classification can be performed on three levels [11]: video genres, video events, and objects in the video. Video genre classification is the classification of videos into different genres such as "movie," "news," "sports," and "cartoon" .genre classification divides the video into genre relevant subset and genre irrelevant subset [15]. Video object classification which is connected with object detection in video data mining is conceptually the lowest grade of video classification. An object-based algorithm to classify video shots. The objects in shots are represented using features of color, texture, and trajectory. A neural network is used to cluster correlative shots, and each cluster is mapped to one of 12 categories [16].

C. Video Annotation

Video annotation is the allocation of video shots or video segments to different redefined semantic concepts, such as person, car, sky, and people walking. Video annotation is similar to video classification, except for two differences. Video classification has a different category/concept ontology compared with video annotation, although some of the concepts could be applied to both. Video classification applies to complete videos, while video annotation applies to video shots or video segments [18].

Learning-based video annotation is essential for video analysis and understanding, and many various approaches have been proposed to avoid the intensive labour costs of purely manual annotation. A Fast Graph-based Semi-Supervised Multiple Instance Learning (FGSSMIL) algorithm, which aims to simultaneously tackle these difficulties in a generic framework for various video domains (e.g., sports, news, and movies), is proposed to jointly explore small-scale expert labelled videos and large-scale unlabeled videos to train the models [17]. Skills-based learning environments are used to promote the acquisition of practical skills as well as decision making, communication, and problem solving.

VII. Query And Retrieval

Once video indices are obtained, content-based video retrieval can be performed. The retrieval results are optimized by relevance feedback.

A) Types of Query:

Classified into two types namely, semantic based and non semantic based query types. Non semantic-based video query types include query by example, query by sketch, and query by objects. Semantic-based video query types include query by keywords and query by natural language.

Query by Example: This query extracts low-level features from given example videos or images and similar videos are found by measuring feature similarity. Query by Sketch: This query allows users to draw sketches to represent the videos they are looking for. Features extracted from the sketches are matched to the features of the stored videos.

Query by Objects: This query allows users to provide an image of object. Then, the system finds and returns all occurrences of the object in the video database.

Query by Keywords: This query represents the user's query by a set of keywords. It is the simplest and most direct query type, and it captures the semantics of videos to some extent.

Query by Natural Language: This is the most natural and convenient way of making a query. Use semantic word similarity to retrieve the most relevant videos and rank

them, given a search query specified in the natural language.

B) Measuring Similarities of Videos

Video similarity measures play an important role in content based video retrieval. To measure video similarities can be classified into feature matching, text matching, ontology based matching, and combination-based matching. The choice of method depends on the query type. Feature Matching approach measures the similarity between two videos is the average distance between the features of the corresponding frames.

Text Matching matches the name of each concept with query terms is the simplest way of finding the videos that satisfy the query.

Ontology-Based Matching approach achieves similarity matching using the ontology between semantic concepts or semantic relations between keywords. Semantic word similarity measures to measure the similarity between texts annotated videos and users' queries. Combination-Based Matching approach leverages semantic concepts by learning the combination strategies from a training collection.

C) Relevance Feedback

Relevance feedback bridges the gap between semantic notions of search relevance and the low level representation of video content. Explicit feedback asks the user to actively select relevant videos from the previously retrieved videos. Implicit feedback refines retrieval results by utilizing click-through data obtained by the search engine as the user clicks on the videos in the presented ranking. Psuedo feedback selects positive and negative samples from the previous retrieval results without the participation of the user.

VIII. Conclusion

Many issues are in further research, especially in the following areas most current video indexing approaches depend heavily on prior domain knowledge. This limits their extensibility to new domains. The elimination of the dependence on domain knowledge is a future research problem. Fast video search using hierarchical indices are all interesting research questions. Effective use of multimedia materials requires efficient way to support it in order to browse and retrieve it. Content-based video indexing and retrieval is an active area of research with continuing attributions from several domain including image processing, computer vision, database system and artificial intelligence.

REFERENCES

[1] Xu Chen, Alfred O. Hero, III, Fellow, IEEE, and Silvio Savarese ,2012,"Multimodal Video Indexing and Retrieval Using Directed Information", IEEE Transactions On Multimedia, VOL. 14, NO. 1, ,pp.3-16.

[2] Zheng-Jun Zha, Member, IEEE, Meng Wang, Member, IEEE, Yan-Tao Zheng, Yi Yang, Richang Hong, 2012,"Interactive Vid-eo Indexing With Statistical Active Learning ", IEEE TransActions On Multimedia, VOL. 14, NO. 1,,p.17-29.

[3] Meng Wang, Member, IEEE, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, Senior Member, IEEE,and Tat-Seng Chua,2012," Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification 2012", IEEE TransActions On Multimedia, VOL. 14, NO. 4, pp.975-985.

[4] Q Miao, 2007,"Accelerating Video Feature Extractions in CBVIR on Multi-core Systems".

[5] R. Yan and A. G. Hauptmann, 2007, "A review of text and image retrieval approaches for broadcast news video," Inform. Retrieval, vol. 10, pp. 445– 484.

[6] J. Adcock, A. Girgensohn, M. Cooper, T. Liu, L. Wilcox, and E. Rieffel,2004 "FXPAL experiments for TRECVID 2004," in Proc. TREC Video Retrieval Eval., Gaithersburg.

[7] A. G. Hauptmann, R. Baron, M. Y. Chen, M. Christel, P. Duygu-lu, C. Huang, R. Jin, W. H. Lin, T. Ng, N. Moraveji, N. Paper-nick, C. Snoek, G. Tzanetakis, J. Yang, R. Yan, and H. Wact-lar,2003, "Informedia at TRECVID 2003: Analyzing and search-ing broadcast news video," in Proc.

[8] J. Sivic, M. Everingham, and A. Zisserman, 2005, "Person spot-ting: Video shot retrieval for face sets," in Proc. Int. Conf. Image Video Retrieval, pp. 226–236.

[9] H. P. Li and D. Doermann, 2002,"Video indexing and retrieval based on recognized text," in Proc. IEEE Workshop Multimedia Signal Process,2002, pp. 245–248.

[10] Xiangang Cheng and Liang-Tien Chia, Member, IEEE,2011," Stratification-Based Keyframe Cliques for Effective and Efficient Video Representation", IEEE Transactions On Multi-Media, VOL. 13, NO. 6, pp.1333-1342

[11] Y. Yuan, 2003,"Research on video classification and retrieval," Ph.D. dissertation, School Electron. Inf. Eng., Xi'an Jiaotong Univ., Xi'an, China, pp. 5–27.

[12] A. Anjulan and N. Canagarajah, 2009,"Aunified framework for object retrieval and mining," IEEE Trans. Circuits Syst. Video Technol., vol. 19, no. 1, pp. 63–76.

[13] Y. F. Zhang, C. S. Xu, Y. Rui, J. Q.Wang, and H. Q. Lu, 2007,"Semantic event extraction from basketball games using multi-modal analysis," in Proc. IEEE Int. Conf. Multimedia Expo, pp. 2190–2193.

[14] T. Quack, V. Ferrari, and L. V. Gool, 2006,"Video mining with frequent item set configurations," in Proc. Int. Conf. Image Video Retrieval, pp. 360–369.

[15] Jun Wu and Marcel Worring,*2012,"* Efficient Genre-Specific Semantic Video Indexing ", IEEE Transactions On Mul-Timedia, VOL. 14, NO. 2 , pp.291-302.

[16] G. Y. Hong, B. Fong, and A. Fong,2005, "An intelligent video categorization engine," Kybernetes, vol. 34, no. 6, pp. 784–802.

[17] Tianzhu Zhang,"A Generic Framework for Video Annotation via Semi-Supervised Learning", IEEE Transactions On Mul-Timedia , Volume: 14,issue 4, 1206 - 1219

[18] L. Yang, J. Liu, X. Yang, and X. Hua,2007, "Multi-modality web video categorization," in Proc. ACM SIGMM Int. Workshop Multimedia Inform. Retrieval, Augsburg, Germany, pp. 265–274.

