Volume 5, Issue 1, January 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Multimodal Techniques in Visual Content-Based Video Retrieval

Abinaya Sambath Kumar
M Phil Research scholar,
Department of Computer Science,
Dr N.G.P Arts and Science College,
Tamil Nadu, India

A. Nirmala
Assistant Professor (SG),
Department of Computer Applications,
Dr N.G.P Arts and Science College,
Tamil Nadu, India

Abstract - The multimedia storage grows and the cost for storing multimedia data is cheaper. There is huge number of videos available in the video repositories. It is difficult to retrieve the relevant videos from large video repository. It is very easy and flexible for searching and accessing the unstructured Multimedia data. This paper proposed an overview of the different existing techniques in multimodal content based video retrieval and different approaches to search in the long videos.

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

I. INTRODUCTION

Many users access the videos from large video repositories like YouTube. It is difficult to manually index and retrieve from large video repositories. It is also difficult to search with in long video clips in order to find portions of segments. Semantic gap between low-level information extracted from the video and the user's need to meaningfully interact with it on a higher level. To increase in videos with very similar con-tents (near duplicate videos) .The near-duplicate videos may be uploaded many times from many different users. So the problem of efficient identification of near duplicate videos on the web is an important issue for video management. Watching a large number of videos to grasp important information quickly is a big challenge. The evolution of the entire event is not directly observable by simply watching these videos. Some videos are weak or not relevant to the query. Content based video retrieval (CBVR) has wide range of applications such as consumer domain applications, quick browsing of video folders, remote instruction, digital museums, news event analysis, video surveillance, and educational applications.

These applications motivate the research in content based video retrieval. Videos have the following information.

1) Video metadata, which are embedded with the video like title, author and description about the video.
2) Sound track from audio channel.
3) Texts obtained by using optical character recognition (OCR) technology.
4) Visual information contained in the images.

Multimodality is the capacity of the system to communicate with a user along different types of communication channels and to extract and convey meaning automatically.

The framework consists of following steps.

1) **Video Segmentation** which includes shot boundary detection
2) **Feature Extraction** includes extracting feature from segmented video clips.
3) **Video mining** to the output of extracted feature.
4) **Video annotation** to build a semantic index.
5) **User query.**
6) **Feedback and Reranking** returns the video to user and feature retrieval are optimized using feedback.

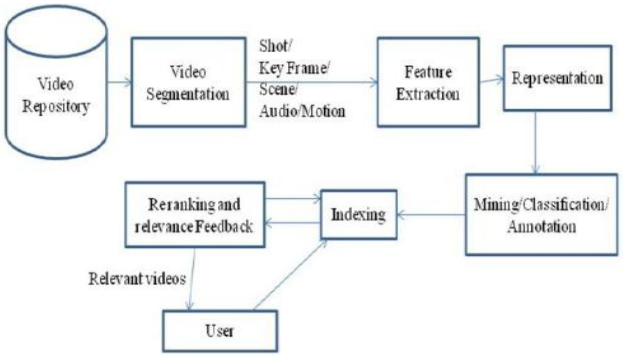

Fig 1: A framework for content based video retrieval

II. VIDEO SEGMENTATION

Video segmentation separates the video into small segments that includes shot boundary detection, key frame extraction, scene segmentation and audio extraction.

Shot Boundary Detection:

Dividing the whole video into a number of temporal segments is called shots. A shot may be defined as a continuous sequence of frames generated by a single non-stop camera operation. 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 similarities between frames using the extracted features, and, finally, detect shot boundaries between frames that are dissimilar.

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.

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.

Scene 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 is also known as story unit segmentation. Scene segmentation approaches can be classified into three categories: key frame based visual information integration-based, and background-based.

Key Frame-Based Approach: It represents each video shot by a set of key frames from which features are extracted. Limitation of the key frame-based approach is Key frames cannot effectively represent the dynamic contents of shots, as shots within a scene are generally correlated by dynamic contents within the scene rather than by key frame-based similarities between shots.

Visual Integration-Based Approach:

It selects a shot boundary where the visual and audio con-tents change simultaneously as a scene boundary. Limitation in this approach is it is Difficult to determine the relation between audio segments and visual shots.

Background-Based Approach:

This approach segments scenes under the assumption that shots belonging to the same scene often have similar back-grounds.

Audio segmentation:

Audio track is often a rich source of content information for all kinds of video genres. A large linguistic literature has shown that topic boundaries are indicated prosodically. In other words, major shifts in topic typical-ly show long pauses a higher maximum accent peak, and greater range intensity. Research has utilized these prosodic features (e.g. pausing, pitch change or rhyme duration) for topic segmentation.

A probabilistic model is used to integrate prosodic and lexical cues for the automatic segmentation of speech into topics. At first a large collection of prosodic features were extracted capturing two major types of speech prosody: duration features and pitch features. A decision tree learning algorithm was used to select salient prosodic features. Then lexical information was captured by statistical language models embedded in a Hidden Markov Model (HMM).

Audio is a promising source in lecture videos. Usually lecture videos contain duration of 60 – 90 minutes. Searching within the entire video to find portion of interest is a time consuming process. Uses the speech-recognition engine (SRE) to extract the text from audio layer and indexing techniques to index the transcript.

III. 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. These mainly include features of key frames, objects, motions and audio/text features.

Features of Key Frames:
It 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.
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.
Shape-based features that describe object shapes in the image can be extracted from object contours or regions.

Features of Objects:
Object features include the dominant color, texture, size, etc., Text-based video indexing and retrieval by, expanding the semantics of a query.

Features of Motion:
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.

Features of Audio:
One advantage of audio approaches is that they typically require fewer computational resources than visual methods another advantage of audio approaches is that the audio clips can be very short; many of the audio-based features are chosen to approximate the human perception of sound. Audio features can lead to three layers of audio understanding.

IV. VIDEO REPRESENTATION

The foundational work that has formulated the problem of computational video representation was presented. In multilayered, iconic annotations of video content called Media Streams. It 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.
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 [42]. 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.

V. MINING, CLASSIFICATION, AND ANNOTATION

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.
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.
Pattern discovery is the automatic discovery of unknown patterns in videos using unsupervised or semi-supervised learning.
Preference Mining For news videos, movies, etc., the user's preferences can be mined. A personalized multimedia news portal to provide, personalized news service by, mining the user's preferences.

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 increasing the efficiency of video retrieval. Semantic content classification can be performed on three levels. 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.
Statistic-based approach classifies videos by statistically modelling various video genres. First, video syntactic properties such as color statistics, cuts, camera motion, and object motion are analyzed. Second, these properties are used to derive more abstract film style attributes such as camera panning and zooming, speech, and music. Finally, these detected style attributes are mapped into film genres. An event can be defined as any human-visible occurrence that has significance to represent video contents. Each video can consist of a number of events, and each event can consist of a number of sub events.
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.

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.

VI. QUERY AND RETRIEVAL

Once video indices are obtained, content-based video retrieval can be performed.

The retrieval results are optimized by relevance feedback.

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.

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.

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.

VII. CONCLUSION AND FUTURE WORK

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. Video indexing and retrieval in the cloud computing environment, where the individual videos to be searched and the dataset of videos are both changing dynamically, will form a new and flourishing research direction in video retrieval in the very near future. Video affective semantics describe human psycho-logical feelings such as romance, pleasure, violence, sadness, and anger. Hierarchically organizing and visualizing retrieval results are all interesting research issues. This paper covers the following tasks: Video segmentation including shot boundary detection, key frame ex-traction, scene segmentation and audio segmentation , extraction of features of static key frames, objects ,audio features and motions, video data mining, video classification and annotation, video search including interface, similarity measure, video retrieval and relevance feed-back.

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 TRANS-ACTIONS ON MULTIMEDIA, VOL. 14, NO. 1,,p.17-29.

[3] Jun Wu and Marcel Worring, Member, IEEE," Efficient Genre-Specific Semantic Video Indexing, 2012," IEEE TRANSAC-TIONS ON MULTIMEDIA, VOL. 14, NO. 2pp.291-302.

[4] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou, Senior Member, IEEE,2012," Multi-View Video Summarization ", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 12, NO. 7, pp.717-729.

[5] L.-H. Chen, Y.-C. Lai, and H.-Y. M. Liao,2008, "Movie scene segmentation using background information," Pattern

[6] Q Miao, 2007,"Accelerating Video Feature Extractions in CBVIR on Multi-core Systems". Meng Wang, Member, IEEE, Richang Hong, Guangda Li

[7] Meng Wang, Member, IEEE, Richang Hong, Guangda Li, ZhengJun 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.

[8] Alexandre Karpenko, Student Member, IEEE, and Parham Aarabi, Senior Member, IEEE,2011," Tiny Videos: A Large Data Set for Nonparametric Video Retrieval and Frame Classification", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 33, NO. 3, p.618-630.

[9] Barbara André*, Tom Vercauteren, Anna M. Buchner, Michael B. Wallace, and Nicholas Ayache, 2012," Learning Semantic and Visual Similarity for Endomicroscopy Video Retrieval ", IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 31, NO. 6, pp.1276-1288.

[10] Xinmie Tian, Linjun Yang, Member, IEEE, Jingdong Wang, Member, IEEE, Xiuqing Wu, and Xian-Sheng Hua, Member, IEEE,2011."Bayesian Visual Reranking", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 13, NO. 4, pp.639-652.

[11] Tianzhu Zhang, Member, IEEE, Changsheng Xu, Senior Member, IEEE, Guangyu Zhu, Si Liu, and Hanqing Lu, Senior Member, IEEE,2012." A Multimedia Retrieval Framework Based on SemiSupervised Ranking and Relevance Feedback", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 14, NO. 4, pp/1206-1219.

[12] Huurnink, B.; Snoek, C.G.M.; de Rijke, M.; Smeulders, A.W.M., 2012."Content-Based Analysis Improves Audiovisual Archive Retrieval", Multimedia, IEEE Transactions on Volume: 14, Page(s): 1166 – 1178.

[13] Yu-Gang Jiang; Qi Dai; Jun Wang; Chong-Wah Ngo; Xiangyang Xue; Shih-Fu Chang ,2012."Fast Semantic Diffusion for LargeScale Context-Based Image and Video Annotation", Image Processing, IEEE Transactions on Volume: 21 , Issue: 6 2012, Page(s): 3080 – 3091.

