International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)

International Conference on Information Systems and Computing (ICISC-2013), INDIA.

A SURVEY ON MULTIMODAL CONTENT BASED VIDEO RETRIEVAL

Tamizharasan.C[1], Dr.S.Chandrakala[1]

[1]*Department of computer science and engineering, Velammal engineering college, Chennai, India*

tamizharasancseslm@gmail.com

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. It is difficult to retrieve the relevant videos from large video repository as per user interest. It is urgently required to make the unstructured multimedia data accessible and searchable with great ease and flexibility. This paper offers an overview of the different existing techniques in multimodal content based video retrieval and different approaches to search with in 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

Due to the increase of available network bandwidth, many users access the videos from large video repositories like youtube. For example, in YouTube, over 48 h of new videos are uploaded to the site every minute, and more than 14 billion videos were viewed in May 2010 [1]. 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 that the user might interested. Semantic gap between low-level information extracted from the video and the user's need to meaningfully interact with it on a higher level. However, the majority of ideas follow a paradigm of finding a direct mapping from low-level features to high level semantic concepts. Not only does this approach requires extremely complex and unstable computation and processing, but it appears to be unfeasible unless it targets a specific and contextually narrow domain.

Substantial increase in videos with very similar contents (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 [63]. 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. Even worse, some videos are indeed weakly 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 [13].

These applications motivate the research in content based video retrieval. Videos have the following information [14], [15]. 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 [17]. Multimodality of video media is the capacity of an author to express a predefined semantic idea, by combining a layout with a specific content, using at least two information channels, where the channels can be visual, auditory or textual. The number of low-level features extracted from various modalities is limited. Some of the research papers in CBVR are listed in **Table 1**. For example [47] give a good overview of the video annotation. [8] Describes different recent researches in reranking method.

The framework consists of following steps as shown in **Fig1**. 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** and 6) **Feedback and Reranking** returns the video to user and feature retrieval are optimized using feedback.

Table 1.
Papers related to video indexing and retrieval

TASK	PAPER	YEAR
Lecture video segmentation	[58] [57]	2011 2008
Video summarization	[4] [5]	2010 2012
Video indexing	[1],[2],[3]	2012
Multimodal CBVR	[10]	2012
Video Representation	[63]	2011
Semantic CBVR	[11],[12]	2012
Video annotation	[47],[12]	2012
Motion based video retrieval	[64]	2012
Video retrieval	[6] [7]	2011 2012
Multimedia information retrieval	[10]	2012
Reranking	[8] [9]	2011 2012
Video classification	[39] [40] [41]	2012 2010 2011

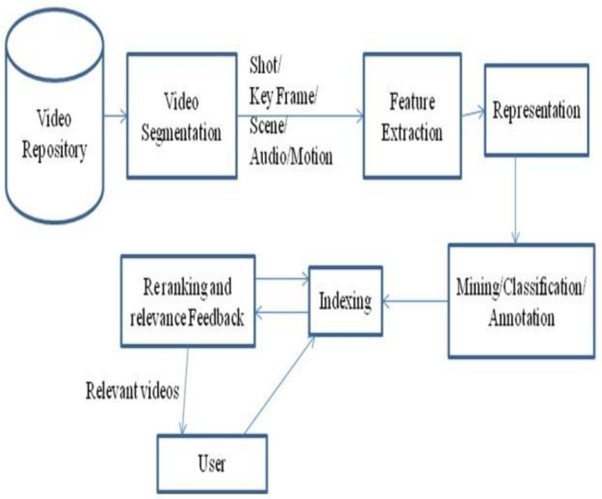
Video Repository
Video Segmentation
Shot/ Key Frame/ Scene/ Audio/Motion
Feature Extraction
Representation
Mining/Classification/ Annotation
Indexing
Reranking and relevance Feedback
Relevant videos
User

Fig 1. General framework for content based video retrieval

II. Video Segmentation

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

2.1 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. 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 extract visual features from each frame, then measure similarities between frames using 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 [17]. 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.

2.2 Key Frame Extraction

There are great redundancies among the frames in the same shot; therefore, certain frames that best reflect the shot contents are selected as key frames [18] to succinctly represent the shot. 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 [19].

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 [20].

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. Construct an alpha-trimmed average histogram describing the color distribution of the frames in a shot [21].

2.3 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: key frame based, visual information integration-based, and background-based. *Key Frame-Based Approach:* represents each video shot by a set of key frames from which features are extracted [22]. Temporally close shots with similar features are grouped into a scene. Compute similarities between shots using block matching of the key frames [24]. 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.

Audio and Vision Integration-Based Approach: Selects a shot boundary where the visual and audio contents change simultaneously as a scene boundary. A time-constrained nearest neighbor algorithm is used to determine the correspondences between these two sets of scenes [25]. 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 backgrounds. A mosaic technique is used to reconstruct the background of each video frame. Color and texture distributions of all the background images in a shot are estimated to determine the shot similarity and the rules of filmmaking are used to guide the shot grouping process [27].

2.4 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 typically 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. [57] Uses the speech-recognition engine (SRE) to extract the text from audio layer and indexing techniques to index the transcript. [58] Uses Sphinx-4 SRE and achieves a recall of 0.72 and an average precision of 0.84 as video retrieval results.

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 [28]. These mainly include features of key frames, objects, motions and audio/text features.

3.1 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 [29]. **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 [30]. **Shape-based features** that describe object shapes in the image can be extracted from object contours or regions. Edge histogram descriptor (EHD) is used to capture the spatial distribution of edges for the video search task in TRECVid-2005 [31].

3.2 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 [32]. 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 [62].

3.3 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 "zooming in or out," "panning left or right," and "tilting up or down," are estimated and used for video indexing. Object-based motion features have attracted much more interest in recent work.

3.4 Audio features

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 [27]: Low-level acoustics, such as the average frequency for a frame, mid-level sound objects, such as the audio signature of the sound a ball makes while bouncing, and high-level scene classes, such as background music playing in certain types of video scenes. Speech typically has a lower bandwidth than music.

IV. VIDEO REPRESENTATION

The foundational work that has formulated the problem of computational video representation was presented in [59]. In [60] 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 [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. **[63]** 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.

V. MINING, CLASSIFICATION, AND ANNOTATION

5.1 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, behavior 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 neighborhood technique to cluster the features in the spatial domain of the frames [33]. Extract stable tracks which are combined into meaningful object clusters, used to mine similar objects [34].**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 [35].**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. Using n-grams and suffix trees to mine motion patterns by analyzing event subsequences over multiple temporal scales. The mined motion patterns are used to detect unusual. **Preference Mining** For news videos, movies, etc., the user's preferences can be mined [37]. A personalized multimedia news portal to provide, personalized news service by, mining the user's preferences [38].

5.2 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. 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 [42]: 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 [43].

By using title-based information only, Song [44] proposes an incremental support vector machine (SVM), with the help of online Wikipedia propagation, to categorize large-scale web videos.

Statistic-based approach classifies videos by statistically modeling 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. Semantics in different granularities are mapped to a hierarchical model in which a complex analysis problem is decomposed into sub problem [45].

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 [46].

5.3 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 [48].

Learning-based video annotation is essential for video analysis and understanding, and many various approaches have been proposed to avoid the intensive labor 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 labeled videos and large-scale unlabeled videos to train the models [47]. Skills-based learning environments are used to promote the acquisition of practical skills as well as decision making, communication, and problem solving [49].

VI. QUERY AND RETRIEVAL

Once video indices are obtained, content-based video retrieval can be performed. On receiving a query, a similarity measure method is used, based on the indices, to search for the candidate videos in accordance with the query. The retrieval results are optimized by relevance feedback.

6.1 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 [51].

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 [52].

6.2 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 [53]. **Text Matching** matches the name of each concept with query terms is the simplest way of finding the videos that satisfy the query. Normalize both the descriptions of concepts and the query text and then compute the similarity between the query text and the text descriptions of concepts by using a vector space model.

Normalize both the descriptions of concepts and the query text and then compute the similarity between the query text and the text descriptions of concepts by using a vector space model [54].

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 [55].**Combination-Based Matching** approach leverages semantic concepts by learning the combination strategies from a training collection.

6.3 Relevance Feedback and Reranking

The videos obtained in reply to a search query are ranked either by the user or automatically. This ranking is used to refine further searches. 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. Adjust the weights embedded in the similarity measure to reflect the user's feedback [56]. 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. Model the textual and visual information from the probabilistic perspective and formulate visual reranking as an optimization problem in the Bayesian framework, termed Bayesian visual reranking [8].

VII. FUTURE DIRECTIONS AND CONCLUSION

Many issues are still open and deserve 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 psychological feelings such as romance, pleasure, violence, sadness, and anger. Affective computing-based video retrieval is the retrieval of videos that produce these feelings in the viewer. Affective semantics in videos are very interesting research issues. It is interesting to simulate human perception to exploit new video retrieval approaches.
- The layout of multimodel information in the human computer interface, the effectiveness of the interface to quickly capture the results in which users are interested, the suitability of the interface for users' evaluation and feedback, and interface's efficiency in adapting to the users' query habits and expressions of their personality are all topics for further investigation.
- Fusions of multiple model information in multiple levels are all difficult issues in the fusion analysis of integrated models.
- Current approaches for semantic-based video indexing and retrieval usually utilize a set of texts to describe the visual contents of videos. There are many unanswered questions.
- The effective use of motion information is essential for content-based video retrieval to distinguish between background motion and foreground motion, detect moving objects and events, combine static features and motion features, and construct motion-based indices are all important research areas.
- Hierarchically organizing and visualizing retrieval results are all interesting research issues.
- Dynamic, online, and adaptive updating of the hierarchical index model, handling of temporal sequence features of videos during index construction and updating, dynamic measure of video similarity based on statistic feature selection, and fast video search using hierarchical indices are all interesting research questions.

This paper covers the following tasks: Video segmentation including shot boundary detection, key frame extraction, 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 feedback.

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 Video Indexing With Statistical Active Learning ", IEEE TRANSACTIONS 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 TRANSACTIONS 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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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 Semi-Supervised Ranking and Relevance Feedback", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 14, NO. 4, pp/1206-1219.

[10] 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.

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

[12] Hong Qing Yu; Pedrinaci, C.; Dietze, S.; Domingue, J.2012," Using linked data to annotate and search educational video resources for supporting distance learning", Learning Technologies, IEEE Transactions on Volume: 5 , Issue: 2 Page(s): 130 – 142.

[13] Nevenka Dimitrova Philips Research," Applications of Video-Content Analysis and Retrieval".

[14] A. F. Smeaton, "Techniques used and open challenges to the analysis, indexing and retrieval of digital video," Inform. Syst., vol. 32, no. 4, pp. 545–559.

[15] Y. Y. Chung, W. K. J. Chin, X. Chen, D. Y. Shi, E. Choi, and F. Chen,2007, "Content-based video retrieval system using wavelet transform," World Sci. Eng. Acad. Soc. Trans. Circuits Syst., vol. 6, no. 2, pp. 259–265.

[16] W.-N. Lie and K.-C. Hsu,2012," A Survey on Visual Content-Based Video Indexing and Retrieval".

[17] Laurence Nigay and Jo¨elle Coutaz.1993," A design space for multimodal systems: concurrent processing and data fusion. In CHI '93: Proceedings of the SIGCHI conference on Human factors in computing systems, pages 172–178, New York, NY, USA,. ACM Press.

[18] K. W. Sze, K. M. Lam, and G. P. Qiu,2005, "A new key frame representation for video segment retrieval," IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 9, pp. 1148–1155.

[19] B. T. Truong and S. Venkatesh, 2007,"Video abstraction: A systematic review and classification," ACM Trans. Multimedia Comput., Commun. Appl., vol. 3, no. 1, art. 3, pp. 1–37.

[20] X.-D. Zhang, T.-Y. Liu, K.-T. Lo, and J. Feng,2003, "Dynamic selection and effective compression of key frames for video abstraction," Pattern Recognit. Lett., vol. 24, no. 9–10, pp. 1523–1532.

[21] K. Matsumoto, M. Naito, K. Hoashi, and F. Sugaya,2006, "SVM-based shot boundary detection with a novel feature," in Proc. IEEE Int. Conf. Multimedia Expo, pp. 1837–1840.

[22] B. T. Truong, S. Venkatesh, and C. Dorai,2003, "Scene extraction in motion pictures," IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 1, pp. 5–15.

[23] N. Goela, K. Wilson, F. Niu, A. Divakaran, and I. Otsuka,2007,"An SVM framework for genre-independent scene change detection," in Proc. IEEE Int.

[24] A. Hanjalic, R. L. Lagendijk, and J. Biemond,1999, "Automated high-level movie segmentation for advanced video-retrieval systems," IEEE Trans. Circuits Syst. Video Technol., vol. 9, no. 4, pp. 580–588.

[25] H. Sundaram and S.-F. Chang,2000, "Video scene segmentation using video and audio features," in Proc. IEEE Int. Conf. Multimedia Expo., New York, pp. 1145–1148.

[26] W.-N. Lie and K.-C. Hsu, 2008,"Video summarization based on semantic feature analysis and user preference," in Proc. IEEE Int. Conf. Sens. Netw., Ubiquitous Trustworthy Comput, pp. 486–491.

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

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

[29] 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.

[30] 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.

[31] A. G. Hauptmann, R. Baron, M. Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W. H. Lin, T. Ng, N. Moraveji, N. Papernick, C. Snoek, G. Tzanetakis, J. Yang, R. Yan, and H. Wactlar,2003, "Informedia at TRECVID 2003: Analyzing and searching broadcast news video," in Proc.

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

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

[34] 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.

[35] 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.

[36] J. Tang, X. S. Hua, M. Wang, Z. Gu, G. J. Qi, and X.Wu,2009,"Correlative linear neighborhood propagation for video annotation," IEEE Trans. Syst., Man, Cybern., B, Cybern., vol. 39, no. 2, pp. 409–416.

[37] K. X. Dai, D. F. Wu, C. J. Fu, G. H. Li, and H. J. Li, 2006,"Video mining: A survey," J. Image Graph., vol. 11, no. 4, pp. 451–457.

[38] V. Kules, V. A. Petrushin, and I. K. Sethi, 2001,"The perseus project: Creating personalized multimedia news portal," in Proc. Int. Workshop Multimedia Data Mining, pp. 1–37.

[39] C. G. M. Snoek ,2012"Adult movie classification system based on multimodal approach with visual and auditory features.

[40] J. Biemond, 2010,"Optimizing support vector machine based classification and retrieval of semantic video events with genetic algorithms".

[41] B. Huurnink ,2011"Boosting video classification using cross-video signals".

[42] 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.

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

[44] Y.-G. Jiang, C.-W. Ngo, and J.Yang,2007, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in Proc. CIVR, pp. 494–501.

[45] P. Chang, M. Han, and Y. Gong,2002, "Extract highlights from baseball game video with hidden Markov models," in Proc. IEEE Int. Conf. Image Process, vol. 1, pp. 609–612.

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

[47] Tianzhu Zhang ,"A Generic Framework for Video Annotation via Semi-Supervised Learning", IEEE TRANSACTIONS ON MULTIMEDIA , Volume: 14,issue 4, 1206 - 1219

[48] 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.

[49] Sargin, M.E.; Aradhye, H,2012,"Boosting video classi_cation using cross-video signals Semantic Annotation of Ubiquitous Learning Environments".

[50] Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki / youtube.

[51] J. Sivic and A. Zisserman, 2006,"Video Google: Efficient visual search of videos," in Toward Category-Level Object Recognition.. Berlin, Germany: Springer, pp. 127–144.

[52] Y. Aytar, M. Shah, and J. B. Luo,2008, "Utilizing semantic word similarity measures for video retrieval," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 1–8.

[53] P. Browne and A. F. Smeaton,2005, "Video retrieval using dialogue, keyframe similarity and video objects," in Proc. IEEE Int. Conf. Image Process., vol. 3, pp. 1208–1211.

[54] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, 2007,"Adding semantics to detectors for video retrieval," IEEE Trans. Multimedia, vol. 9, no. 5, pp. 975–985.

[55] Y. Aytar, M. Shah, and J. B. Luo,2008, "Utilizing semantic word similarity measures for video retrieval," in Proc. IEEE Conf. Comput. Vis. Pattern Recog, pp. 1–8.

[56] L.-H. Chen, K.-H. Chin, and H.-Y. Liao,2007, "An integrated approach to video retrieval," in Proc. ACM Conf. Australasian Database, vol. 75, Gold Coast, Australia, pp. 49–55.

[57] Stephen Repp,2008"Browsing within Lecture Videos Based on the Chain Index of Speech Transcription".

[58] Vijaya kumar,2011,"Automated tagging to enable fine-grained browsing of lecture videos".

[59] Marc Davis,1995",Media streams: representing video for retrieval and repurposing".

[60] Marc Davis.1994," Media streams: representing video for retrieval and repurposing".

[61] Chih-Wen Su, Hong-Yuan Mark Liao, Senior Member, IEEE, Hsiao-Rong Tyan, Chia-Wen Lin, Senior Member, IEEE, Duan-Yu Chen, and Kuo-Chin Fan, Member, IEEE,2007," Motion Flow-Based Video Retrieval", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 9, NO. 6, pp.1193-1201.

[62] 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.

[63] Xiangang Cheng and Liang-Tien Chia, Member, IEEE,2011," Stratification-Based Keyframe Cliques for Effective and Efficient Video Representation", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 13, NO. 6,pp.1333-1342

[64] Wei-Ta Chu; Shang-Yin Tsai,2012," Rhythm of Motion Extraction and Rhythm-Based Cross-Media Alignment for Dance Videos", Multimedia, IEEE Transactions on Volume: 14