A Survey on Visual Content-Based Video Indexing and Retrieval

Weiming Hu, *Senior Member, IEEE*, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank

***Abstract*—Video indexing and retrieval have a wide spectrum of promising applications, motivating the interest of researchers worldwide. This paper offers a tutorial and an overview of the landscape of general strategies in visual content-based video indexing and retrieval, focusing on methods for video structure analysis, including shot boundary detection, key frame extraction and scene segmentation, extraction of features including static key frame features, object features and motion features, video data mining, video annotation, video retrieval including query interfaces, similarity measure and relevance feedback, and video browsing. Finally, we analyze future research directions.**

***Index Terms*—Feature extraction, video annotation, video browsing, video retrieval, video structure analysis.**

I. Introduction

MULTIMEDIA information indexing and retrieval [44] are required to describe, store, and organize multimedia information and to assist people in finding multimedia resources conveniently and quickly. Dynamic video is an important form of multimedia information. Videos have the following characteristics: 1) much richer content than individual images; 2) huge amount of raw data; and 3) very little prior structure. These characteristics make the indexing and retrieval of videos quite difficult. In the past, video databases have been relatively small, and indexing and retrieval have been based on keywords annotated manually. More recently, these databases have become much larger and content-based indexing and retrieval are required, based on the automatic analysis of videos with the minimum of human participation.

Content-based video indexing and retrieval have a wide range of applications such as quick browsing of video folders, analysis of visual electronic commerce (such as analysis of interest trends of users' selections and orderings, analysis of correlations between advertisements and their effects), remote instruction, digital museums, news event analysis [96], intelligent management of web videos (useful video search and harmful video tracing), and video surveillance.

It is the broad range of applications that motivates the interests of researchers worldwide. The following two examples of research activity are particularly noteworthy. 1) Since 2001, the National Institute of Standards and Technology has been sponsoring the annual Text Retrieval Conference (TREC) Video Retrieval Evaluation (TRECVid) to promote progress in video analysis and retrieval. Since 2003, TRECVid has been independent of TREC. TRECVid provides a large-scale test collection of videos, and dozens of participants apply their content-based video retrieval algorithms to the collection [260], [263], [266]. 2) The goal of video standards is to ensure compatibility between description interfaces for video contents in order to facilitate the development of fast and accurate video retrieval algorithms. The main standards for videos are the moving picture experts group (MPEG) and the TV-Anytime Standard [254]. There exist many investigations that adopt the MPEG-7 to extract features to classify video contents or to describe video objects in the compressed domain [78].

A video may have an auditory channel as well as a visual channel. The available information from videos includes the following [66], [67]: 1) video metadata, which are tagged texts embedded in videos, usually including title, summary, date, actors, producer, broadcast duration, file size, video format, copyright, etc.; 2) audio information from the auditory channel; 3) transcripts: Speech transcripts can be obtained by speech recognition and caption texts can be read using optical character recognition techniques; 4) visual information contained in the images themselves from the visual channel. If the video is included in a web page, there are usually web page texts associated with the video. In this paper, we focus on the visual contents of videos and give a survey on visual content-based video indexing and retrieval.

The importance and popularity of video indexing and retrieval have led to several survey papers, which are listed in Table I, together with the publication years and topics. In general, each paper covers only a subset of the topics in video indexing and retrieval. For example, Smeaton *et al*. [263] give a good review of video shot boundary detection during seven years of the TRECVid activity. Snoek and Worring [262] present a detailed review of concept-based video retrieval. They emphasize semantic concept detection, video search using semantic concepts, and the evaluation of algorithms using the TRECVid databases. Ren *et al*. [278] review the state of the art of spatiotemporal semantic information-based video retrieval. Schoeffmann *et al*. [261] give a good review of interfaces and applications of video browsing systems.

Manuscript received March 29, 2010; revised November 2, 2010; accepted January 22, 2011. Date of publication March 10, 2011; date of current version October 19, 2011. This work was supported in part by the National Natural Science Foundation of China under Grant 60825204 and Grant 60935002 and in part by the National 863 High-Tech R&D Program of China under Grant 2009AA01Z318. This paper was recommended by Associate Editor E. Trucco.

W. Hu, N. Xie, L. Li, and X. Zeng are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wmhu@nlpr.ia.ac.cn; nhxie@nlpr.ia.ac.cn; lli@nlpr.ia.ac.cn; xlzeng@nlpr.ia.ac.cn).

S. Maybank is with the Department of Computer Science and Information Systems, Birkbeck College, London WC1E 7HX, U.K. (e-mail: sjmaybank@dcs.bbk.ac.uk).

Digital Object Identifier 10.1109/TSMCC.2011.2109710

TABLE I
SURVEYS ON VIDEO INDEXING AND RETRIEVAL

Year	Paper	Topic
2005	[256]	Video indexing
2006	[181]	Video summarization
	[183]	Video retrieval in meeting movies, news and sports
	[260]	Multimedia information retrieval
2007	[16]	Shot boundary detection
	[39]	Video abstraction
	[229]	Text and image retrieval for broadcast news video
	[257]	Semantic image/video search
2008	[245]	Video classification
2009	[262]	Concept-based video retrieval
	[266]	Concept detection in TRECVid
	[278]	Spatio-temporal information-based video retrieval
2010	[261]	Video browsing interfaces and applications
	[263]	Video shot boundary detection in TREVID

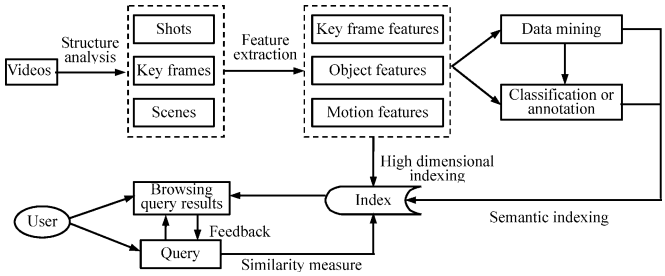


Fig. 1. Generic framework for visual content-based video indexing and retrieval.

Unlike previous reviews, we give a more general overview on the overall process of a video indexing and retrieval framework which is outlined in Fig. 1. The framework includes the following: 1) structure analysis: to detect shot boundaries, extract key frames, and segment scenes; 2) feature extraction from segmented video units (shots or scenes): These features include static features in key frames, object features, motion features, etc.; 3) video data mining using the extracted features; 4) video annotation: using extracted features and mined knowledge to build a semantic video index. The semantic index together with the high-dimensional index of video feature vectors constitutes the total index for video sequences that are stored in the database; 5) query: the video database is searched for the desired videos using the index and the video similarity measures; 6) video browsing and feedback: The videos found in response to a query are returned to the user to browse in the form of a video summary, and subsequent search results are optimized through relevance feedback. In this paper, we review recent developments and analyze future open directions in visual content-based video indexing and retrieval. The main contributions of this survey are as follows.

1) Video indexing and retrieval components are discussed in a clearly organized hierarchical manner, and interlinks between these components are shown.
2) To examine the state of the art, each task involved in visual content-based video indexing and retrieval is divided into subprocesses and various categories of approaches to the subprocesses are discussed. The merits and limitations of the different approaches are summarized. For the tasks for which there exist surveys, we focus on reviewing recent papers as a supplement to the previous surveys. For the tasks that have not yet been specially surveyed, detailed reviews are given.
3) We discuss in detail future directions in visual content-based video indexing and retrieval.

The aforesaid contributions clearly distinguish our survey from the existing surveys on video indexing and retrieval. To our knowledge, our survey is the broadest.

The remainder of this paper is organized as follows: Section II briefly reviews the work related to video structure analysis. Section III addresses feature extraction. Section IV discusses video data mining, classification, and annotation. Section V describes the approaches for video query and retrieval. Section VI presents video summarization for browsing. Section VII analyzes possible directions for future research. Section VIII summarizes this paper.

II. VIDEO STRUCTURE ANALYSIS

Generally, videos are structured according to a descending hierarchy of video clips, scenes, shots, and frames. Video structure analysis aims at segmenting a video into a number of structural elements that have semantic contents, including shot boundary detection, key frame extraction, and scene segmentation.

A. Shot Boundary Detection

A shot is a consecutive sequence of frames captured by a camera action that takes place between start and stop operations, which mark the shot boundaries [10]. There are strong content correlations between frames in a shot. Therefore, shots are considered to be the fundamental units to organize the contents of video sequences and the primitives for higher level semantic annotation and retrieval tasks. Generally, 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. Cut detection is easier than gradual transition detection.

The research on shot boundary detection has a long history, and there exist specific surveys on video shot boundary detection [16], [263]. For completeness, we only briefly introduce the basic categories of methods for shot boundary detection and their merits and limitations, and review some recent papers as a supplement to [16] and [263].

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. In the following, we discuss the main three steps in shot boundary detection: feature extraction, similarity measurement [113], and detection.

The features used for shot boundary detection include color histogram [87] or block color histogram, edge change ratio, motion vectors [85], [163], together with more novel features such as scale invariant feature transform [83], corner points [82], information saliency map [77], etc. Color histograms are robust to small camera motion, but they are not able to differentiate the shots within the same scene, and they are sensitive to large camera motions. Edge features are more invariant to illumination changes and motion than color histograms, and motion features can effectively handle the influence of object and camera motion. However, edge features and motion features as well as more complicated features cannot in general outperform the simple color histograms [16].

To measure similarity between frames using the extracted features is the second step required for shot boundary detection. Current similarity metrics for extracted feature vectors include the 1-norm cosine dissimilarity, the Euclidean distance, the histogram intersection, and the chi-squared similarity [11], [12], [191], as well as some novel similarity measures such as the earth mover's distance [87] and mutual information [68], [72], [207]. The similarity measures include pair-wise similarity measures that measure the similarities between consecutive frames and window similarity measures that measure similarities between frames within a window [191]. Window-based similarity measures incorporate contextual information to reduce the influence of local noises or disturbances, but they need more computation than the pair-wise similarity measures.

Using the measured similarities between frames, shot boundaries can be detected. Current shot boundary detection approaches can be classified into threshold-based and statistical learning-based.

1) Threshold-Based Approach: The threshold-based approach detects shot boundaries by comparing the measured pair-wise similarities between frames with a predefined threshold [47], [180]: When a similarity is less than the threshold, a boundary is detected. The threshold can be global, adaptive, or global and adaptive combined. 1) The global threshold-based algorithms use the same threshold, which is generally set empirically, over the whole video, as in [180]. The major limitation of the global threshold-based algorithms is that local content variations are not effectively incorporated into the estimation of the global threshold, therefore influencing the boundary detection accuracy. 2) The adaptive threshold-based algorithms [77], [87], [207] compute the threshold locally within a sliding window. Detection performance is often improved when an adaptive threshold is used instead of a global threshold [65]. However, estimation of the adaptive threshold is more difficult than estimation of the global threshold and users are required to be more familiar with characteristics of videos in order to choose parameters such as the size of the sliding window. 3) Global and adaptive combined algorithms adjust local thresholds, taking into account the values of the global thresholds. Quenot *et al*. [264] define the thresholds for cut transition detection, dissolve transition detection, and flash detection as the functions of two global thresholds that are obtained from a trade-off between recall and precision. Although this algorithm only needs to tune two global thresholds, the values of the functions are changed locally. The limitation of this algorithm is that the functional relations between the two global thresholds and the locally adaptive thresholds are not easy to determine.

2) Statistical Learning-Based Approach: The statistical learning-based approach regards shot boundary detection as a classification task in which frames are classified as shot change or no shot change depending on the features that they contain. Supervised learning and unsupervised learning are both used.

a) Supervised learning-based classifiers: The most commonly used supervised classifiers for shot boundary detection are the support vector machine (SVM) and Adaboost.

1) *SVM* [11], [21]: Chavez *et al*. [84] use the SVM as a two-class classifier to separate cuts from noncuts. A kernel function is used to map the features into a high-dimensional space in order to overcome the influence of changes in illumination and fast movement of objects. Zhao *et al*. [61] exploit two SVM classifiers, in a sliding window, to detect cuts and gradual transitions, respectively. Ling *et al*. [58] first extract several features from each frame, and then use the SVM to classify the frames using these features into three categories: cut, gradual transition, and others. Yuan *et al*. [16] and Liu *et al*. [72] combine the threshold-based method with an SVM-based classifier. First, the candidate boundaries are selected using the threshold-based method, and then the SVM classifier is used to verify the boundaries. The SVM-based algorithms are widely used for shot boundary detection [265] because of their following merits.
   a) They can fully utilize the training information and maintain good generalization.
   b) They can deal efficiently with a large number of features by the use of kernel functions.
   c) Many good SVM codes are readily available.
2) *Adaboost:* Herout *et al.* [63] make cut detection a pattern recognition task to which the Adaboost algorithm is applied. Zhao and Cai [85] apply the Adaboost algorithm to shot boundary detection in the compressed domain. The color and motion features are roughly classified first using a fuzzy classifier, and then each frame is classified as a cut, gradual, or no change frame using the Adaboost classifier. The main merit of the Adaboost boundary classifiers is that a large number of features can be handled: These classifiers select a part of features for boundary classification.
3) *Others:* Other supervised learning algorithms have been employed for shot boundary detection. For instance, Cooper *et al*. [191] use the binary k nearest-neighbor (kNN) classifier, where the similarities between frames within the particular temporal interval are used as its input. Boreczky and Wilcox [121] apply hidden Markov (HMM) models with separate states to model shot cuts, fades, dissolves, pans, and zooms.

The merits of the aforementioned supervised-learning approaches are that there is no need to set the thresholds used

in the threshold-based approaches, and different types of features can be combined to improve the detection accuracy. The limitation is their heavy reliance on a well-chosen training set containing both positive and negative examples.

b) Unsupervised learning-based algorithms: The unsupervised learning-based shot boundary detection algorithms are classified into frame similarity-based and frame-based. The frame similarity-based algorithms cluster the measurements of similarity between pairs of frames into two clusters: the cluster with lower values of the similarities corresponds to shot boundaries and the cluster with higher values of the similarities corresponds to nonboundaries. Clustering algorithms such as K-means and fuzzy K-means [64] have been used. The frame-based algorithms treat each shot as a cluster of frames that have similar visual content. Chang *et al*. [83] use clustering ensembles to group different frames into their corresponding shots. Lu *et al.* [12] use K-means clustering, and Damnjanovic *et al*. [57] use spectral clustering to cluster frames to detect the different shots. The merit of clustering-based approaches is that the training dataset is not needed. Their limitations are that temporal sequence progression information is not preserved, and they are inefficient in recognizing the different types of gradual transition.

Shot boundary detection approaches can be classified into uncompressed domain-based and compressed domain-based. To avoid time-consuming video decompression, the features available in the compressed domain such as discrete cosine transform coefficients, DC image and MB types, and motion vectors can be directly employed for shot boundary detection [40], [60], [85]. However, the compressed domain-based approach is highly dependent on the compression standards, and it is less accurate than the uncompressed domain-based approach.

Recently, the detection of gradual transitions has received more attention. Ngo [41] detects dissolves based on multiresolution analysis. Yoo *et al*. [131] detect gradual transitions according to the variance distribution curve of edge information in frame sequences.

B. 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 [15], [39], [170], [193] to succinctly represent the shot. The extracted key frames should contain as much salient content of the shot as possible and avoid as much redundancy as possible. The features used for key frame extraction include colors (particularly the color histogram), edges, shapes, optical flow, MPEG-7 motion descriptors such as temporal motion intensity and spatial distribution of motion activity [206], MPEG discrete cosine coefficient and motion vectors [202], camera activity, and features derived from image variations caused by camera motion [161], [208].

Referring to [39], 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.

1) Sequential Comparison Between Frames: In these algorithms, frames subsequent to a previously extracted key frame are sequentially compared with the key frame until a frame which is very different from the key frame is obtained. This frame is selected as the next key frame. For instance, Zhang *et al*. [209] used the color histogram difference between the current frame and the previous key frame to extract key frames. Zhang *et al*. [210] use the accumulated energy function computed from image-block displacements across two successive frames to measure the distance between frames to extract key frames. The merits of the sequential comparison-based algorithms include their simplicity, intuitiveness, low computational complexity, and adaptation of the number of key frames to the length of the shot. The limitations of these algorithms include the following. 1) The key frames represent local properties of the shot rather than the global properties. b) The irregular distribution and uncontrolled number of key frames make these algorithms unsuitable for applications that need an even distribution or a fixed number of key frames. c) Redundancy can occur when there are contents appearing repeatedly in the same shot.

2) Global Comparison Between Frames: The algorithms based on global differences between frames in a shot distribute key frames by minimizing a predefined objective function that depends on the application. In general, the objective function has one of the following four forms [39].

1) *Even temporal variance:* These algorithms select key frames in a shot such that the shot segments, each of which is represented by a key frame, have equal temporal variance. The objective function can be chosen as the sum of differences between temporal variances of all the segments. The temporal variance in a segment can be approximated by the cumulative change of contents across consecutive frames in the segment [208] or by the difference between the first and last frames in the segment. For instance, Divakaran *et al*. [211] obtain key frames by dividing the shot into segments with equal cumulative motion activity using the MPEG-7 motion activity descriptor, and then, the frame located at the halfway point of each segment is selected as a key frame.
2) *Maximum coverage:* These algorithms extract key frames by maximizing their representation coverage, which is the number of frames that the key frames can represent [39]. If the number of key frames is not fixed, then these algorithms minimize the number of key frames subject to a predefined fidelity criterion; alternatively, if the number of key frames is fixed, the algorithms maximize the number of frames that the key frames can represent [212], [213]. For instance, Chang *et al*. [214] specify the coverage of a key frame as the number of the frames that are visually similar to the key frame. A greedy algorithm is used iteratively to find key frames.
3) *Minimum correlation:* These algorithms extract key frames to minimize the sum of correlations between key frames (especially successive key frames), making key frames as uncorrelated with each other as possible. For instance, Porter *et al*. [215] represent frames in a shot and

their correlations using a directed weighted graph. The shortest path in the graph is found and the vertices in the shortest path which corresponds to minimum correlation between frames designate the key frames.

4) *Minimum reconstruction error:* These algorithms extract key frames to minimize the sum of the differences between each frame and its corresponding predicted frame reconstructed from the set of key frames using interpolation. These algorithms are useful for certain applications, such as animation. Lee and Kim [216] use an iterative procedure to select a predetermined number of key frames, in order to reduce the shot reconstruction error as much as possible. Liu *et al*. [217] propose a key frame selection algorithm based on the extent to which key frames record the motion during the shot. In the algorithm, an inertia-based frame interpolation algorithm is used to interpolate frames.

The merits of the aforesaid global comparison-based algorithms include the following. 1) The key frames reflect the global characteristics of the shot. 2) The number of key frames is controllable. 3) The set of key frames is more concise and less redundant than that produced by the sequential comparison-based algorithms. The limitation of the global comparison-based algorithms is that they are more computationally expensive than the sequential comparison-based algorithms.

3) Reference Frame: These algorithms generate a reference frame and then extract key frames by comparing the frames in the shot with the reference frame. For instance, Ferman and Tekalp [204] construct an alpha-trimmed average histogram describing the color distribution of the frames in a shot. Then, the distance between the histogram of each frame in the shot and the alpha-trimmed average histogram is calculated. Key frames are located using the distribution of the distance curve. Sun *et al*. [205] construct a maximum occurrence frame for a shot. Then, a weighted distance is calculated between each frame in the shot and the constructed frame. Key frames are extracted at the peaks of the distance curve. The merit of the reference frame-based algorithms is that they are easy to understand and implement. The limitation of these algorithms is that they depend on the reference frame: If the reference frame does not adequately represent the shot, some salient contents in the shot may be missing from the key frames.

4) Clustering: These algorithms cluster frames and then choose frames closest to the cluster centers as the key frames. Girgensohn and Boreczky [199] select key frames using the complete link method of hierarchical agglomerative clustering in the color feature space. Yu *et al*. [200] extract key frames using the fuzzy K-means clustering in the color feature subspace. Gibson *et al*. [201] use Gaussian mixture models (GMM) in the eigenspace of the image, in which the number of GMM components is the required number of clusters. The merits of the clustering-based algorithms are that they can use generic clustering algorithms, and the global characteristics of a video can be reflected in the extracted key frames. The limitations of these algorithms are as follows: First, they are dependent on the clustering results, but successful acquisition of semantic meaningful clusters is very difficult, especially for large data, and second, the sequential nature of the video cannot be naturally utilized: Usually, clumsy tricks are used to ensure that adjacent frames are likely to be assigned to the same cluster.

5) Curve Simplification: These algorithms represent each frame in a shot as a point in the feature space. The points are linked in the sequential order to form a trajectory curve and then searched to find a set of points which best represent the shape of the curve. Calic and Izquierdo [218] generate the frame difference metrics by analyzing statistics of the macroblock features extracted from the MPEG compressed stream. The key frame extraction method is implemented using difference metrics curve simplification by the discrete contour evolution algorithm. The merit of the curve simplification-based algorithms is that the sequential information is kept during the key frame extraction. Their limitation is that optimization of the best representation of the curve has a high computational complexity.

6) Objects/Events: These algorithms [192] jointly consider key frame extraction and object/event detection in order to ensure that the extracted key frames contain information about objects or events. Calic and Thomas [196] use the positions of regions obtained using frame segmentation to extract key frames where objects merge. Kim and Hwang [197] use shape features to extract key frames that can represent changes of human gestures. Liu and Fan [194] select initial key frames based on the color histogram and use the selected key frames to estimate a GMM for object segmentation. The segmentation results and the trained GMM are further used to refine the initial key frames. Song and Fan [195] propose a joint key frame extraction and object segmentation method by constructing a unified feature space for both processes, where key frame extraction is formulated as a feature selection process for object segmentation in the context of GMM-based video modeling. Liu *et al*. [203] propose a triangle model of perceived motion energy for motion patterns in videos. The frames at the turning points of the motion acceleration and motion deceleration are selected as key frames. Han and Kweon [220] extract key frames by the maximum curvature of camera motion at each temporal scale. The key frames provide temporal interest points for classification of video events. The merit of the object/event-based algorithms is that the extracted key frames are semantically important, reflecting objects or the motion patterns of objects. The limitation of these algorithms is that object/event detection strongly relies on heuristic rules specified according to the application. As a result, these algorithms are efficient only when the experimental settings are carefully chosen.

Because of the subjectivity of the key frame definition, there is no uniform evaluation method for key frame extraction. In general, the error rate and the video compression ratio are used as measures to evaluate the result of key frame extraction. Key frames giving low error rates and high compression rates are preferred. In general, a low error rate is associated with a low compression rate. The error rate depends on the parameters in the key frame extraction algorithms. Examples of these parameters are the thresholds in sequential comparison-based, global comparison-based, reference frame-based, and clustering-based algorithms, as well as the parameters to fit the curve in the curve

simplification-based algorithms. Users choose the parameters according to the error rate that can be tolerated.

C. Scene Segmentation

Scene segmentation is also known as story unit segmentation. In general, a scene is a group of contiguous shots that are coherent with a certain subject or theme. Scenes have higher level semantics than shots. Scenes are identified or segmented out by grouping successive shots with similar content into a meaningful semantic unit. The grouping may be based on information from texts, images, or the audio track in the video.

According to shot representation, scene segmentation approaches can be classified into three categories: key frame-based, audio and visual information integration-based, and background-based.

1) Key Frame-Based Approach: This approach [145] represents each video shot by a set of key frames from which features are extracted. Temporally close shots with similar features are grouped into a scene. For instance, Hanjalic *et al*. [140] compute similarities between shots using block matching of the key frames. Similar shots are linked, and scenes are segmented by connecting the overlapping links. Ngo *et al*. [144] extract and analyze the motion trajectories encoded in the temporal slices of image volumes. A motion-based key frame selection strategy is, thus, used to compactly represent shot contents. Scene changes are detected by measuring the similarity of the key frames in the neighboring shots. The limitation of the key frame-based approach is that 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.

2) Audio and Vision Integration-Based Approach: This approach selects a shot boundary where the visual and audio contents change simultaneously as a scene boundary. For instance, Sundaram and Chang [147] detect audio scenes and video scenes separately. A time-constrained nearest neighbor algorithm is used to determine the correspondences between these two sets of scenes. The limitation of the audio and visual integration-based approach is that it is difficult to determine the relation between audio segments and visual shots.

3) Background-Based Approach: This approach segments scenes under the assumption that shots belonging to the same scene often have similar backgrounds. For instance, Chen *et al.* [139] use a mosaic technique to reconstruct the background of each video frame. Then, the 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. The limitation of the background-based approach is the assumption that shots in the same scene have similar backgrounds: sometimes the backgrounds in shots in a scene are different.

According to the processing method, current scene segmentation approaches can be divided into four categories: merging-based, splitting-based, statistical model-based, and shot boundary classification-based.

a) Merging-based approach: This approach gradually merges similar shots to form a scene in a bottom-up style. Rasheed and Shah [133] propose a two-pass scene segmentation algorithm. In the first pass, oversegmentation of scenes is carried out using backward shot coherence. In the second pass, the oversegmented scenes are identified using motion analysis and then merged. Zhao *et al*. [134] propose a best first model-merging algorithm for scene segmentation. The algorithm takes each shot as a hidden state and loops upon the boundaries between consecutive shots by a left-right HMM.

b) Splitting-based approach: This approach splits the whole video into separate coherent scenes using a top-down style. For instance, Rasheed and Shah [136] construct a shot similarity graph for a video and partition the graph using normalized cuts. The subgraphs represent individual scenes in the video. Tavanapong and Zhou [138] introduce a scene definition for narrative films and present a technique to cluster relevant shots into a scene using this definition.

c) Statistical model-based approach: This approach constructs statistical models of shots to segment scenes. Zhai and Shah [132] use the stochastic Monte Carlo sampling to simulate the generation of scenes. The scene boundaries are updated by diffusing, merging, and splitting the scene boundaries estimated in the previous step. Tan and Lu [137] use the GMM to cluster video shots into scenes according to the features of individual shots. Each scene is modeled with a Gaussian density. Gu *et al*. [149] define a unified energy minimization framework in which the global content constraint between individual shots and the local temporal constraint between adjacent shots are both represented. A boundary voting procedure decides the optimal scene boundaries.

d) Shot boundary classification-based approach: In this approach, features of shot boundaries are extracted and then used to classify shot boundaries into scene boundaries and non-scene boundaries. Goela *et al*. [148] present a genre-independent method to detect scene boundaries in broadcast videos. In their method, scene segmentation is based on a classification with the two classes of "scene change" and "nonscene change." An SVM is used to classify the shot boundaries. Hand-labeled video scene boundaries from a variety of broadcast genres are used to generate positive and negative training samples for the SVM.

The common point in the merging-based, splitting-based, and statistical model-based approaches is that the similarities between different shots are used to combine similar shots into scenes. This is simple and intuitive. However, in these approaches, shots are usually represented by a set of selected key frames, which often fail to represent the dynamic contents of the shots. As a result, two shots are regarded as similar, if their key frames are in the same environment rather than if they are visually similar. The shot boundary classification-based approach takes advantage of the local information about shot boundaries. This ensures that algorithms with low computational complexities are easy to obtain. However, lack of global information about shots inevitably reduces the accuracy of scene segmentation.

It is noted that most current approaches for scene segmentation exploit the characteristics of specific video domains such

as movies, TVs, and news broadcasts [150], [152], [153], for example, using the production rules by which movies or TV shows are composed. The accuracy of scene segmentation is improved, but it is necessary to construct *a priori* model for each application.

III. Feature Extraction

To extract features according to video structural analysis results is the base of video indexing and retrieval. We focus on the visual features suitable for video indexing and retrieval. These mainly include features of key frames, objects, and motions. Auditory features and text features are not covered.

A. Static Features of Key Frames

The key frames of a video reflect the characteristics of the video to some extent. Traditional image retrieval techniques can be applied to key frames to achieve video retrieval. The static key frame features useful for video indexing and retrieval are mainly classified as color-based, texture-based, and shape-based.

1) Color-Based Features: Color-based features include color histograms, color moments, color correlograms, a mixture of Gaussian models, etc. The exaction of color-based features depends on color spaces such as RGB, HSV, YCbCr and normalized r-g, YUV, and HVC. The choice of color space depends on the applications. Color features can be extracted from the entire image or from image blocks into which the entire image is partitioned. Color-based features are the most effective image features for video indexing and retrieval. In particular, color histogram and color moments are simple but efficient descriptors. Amir *et al*. [222] compute color histogram and color moments for video retrieval and concept detection. Yan and Hauptmann [229] first split the image into 5×5 blocks to capture local color information. Then in each block, color histogram and color moments are extracted for video retrieval. Adcock *et al*. [226] use color correlograms to implement a video search engine. The merits of color-based features are that they reflect human visual perception, they are easy to extract, and their extraction has low computational complexity. The limitation of color-based features is that they do not directly describe texture, shape, etc., and are, thus, ineffective for the applications in which texture or shape is important.

2) Texture-Based Features: Texture-based features are object surface-owned intrinsic visual features that are independent of color or intensity and reflect homogenous phenomena in images. They contain crucial information about the organization of object surfaces, as well as their correlations with the surrounding environment. Texture features in common use include Tamura features, simultaneous autoregressive models, orientation features, wavelet transformation-based texture features, co-occurrence matrices, etc. Amir *et al*. [222] use co-occurrence texture and Tamura features including coarseness, contrast and directionality for the TRECVid-2003 video retrieval task. Hauptmann *et al*. [223] use Gabor wavelet filters to capture texture information for a video search engine. They design 12 oriented energy filters. The mean and variance of the filtered outputs are concatenated into a texture feature vector. Hauptmann *et al*. [228] divide the image into 5×5 blocks and compute texture features using Gabor-wavelet filters in each block. The merit of texture-based features is that they can be effectively applied to applications in which texture information is salient in videos. However, these features are unavailable in nontexture video images.

3) Shape-Based Features: Shape-based features that describe object shapes in the image can be extracted from object contours or regions. A common approach is to detect edges in images and then describe the distribution of the edges using a histogram. Hauptmann *et al*. [223] use the edge histogram descriptor (EHD) to capture the spatial distribution of edges for the video search task in TRECVid-2005. The EHD is computed by counting the number of pixels that contribute to the edge according to their quantized directions. To capture local shape features, Foley *et al*. [224] and Cooke *et al*. [225] first divide the image into 4×4 blocks and then extract a edge histogram for each block. Shape-based features are effective for applications in which shape information is salient in videos. However, they are much more difficult to extract than color- or texture-based features.

B. Object Features

Object features include the dominant color, texture,, size, etc., of the image regions corresponding to the objects. These features can be used to retrieve videos likely to contain similar objects [17]. Faces are useful objects in many video retrieval systems. For example, Sivic *et al*. [18] 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. Le *et al*. [19] propose a method to retrieve faces in broadcast news videos by integrating temporal information into facial intensity information. Texts in a video are extracted as one type of object to help understand video contents. Li and Doermann [20] implement 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. The limitation of object-based features is that identification of objects in videos is difficult and time-consuming. Current algorithms focus on identifying specific types of objects, such as faces, rather than various objects in various scenes.

C. Motion Features

Motion is the essential characteristic distinguishing dynamic videos from still images. Motion information represents the visual content with temporal variation. Motion features are closer to semantic concepts than static key frame features and object features. Video motion includes background motion caused by camera motion and foreground motion caused by moving objects. Thus, 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. Video retrieval using only camera-based features has the limitation that they cannot describe motions of key objects.

Object-based motion features have attracted much more interest in recent work. Object-based motion features can be further classified into statistics-based, trajectory-based, and objects' spatial relationships-based.

1) Statistics-Based: Statistical features of the motions of points in frames in a video are extracted to model the distribution of global or local motions in the video. For instance, Fablet *et al*. [233] use causal Gibbs models to represent the spatiotemporal distribution of appropriate local motion-related measurements computed after compensating for the estimated dominant image motions in the original sequence. Then, a general statistical framework is developed for video indexing and retrieval. Ma and Zhang [234] transform the motion vector field to a number of directional slices according to the energy of the motion. These slices yield a set of moments that form a multidimensional vector called motion texture. The motion texture is used for motion-based shot retrieval. The merit of statistics-based features is that their extraction has low computational complexity. The limitation of these features is they cannot represent object actions accurately and cannot characterize the relations between objects.

2) Trajectory-Based: Trajectory-based features [22] are extracted by modeling the motion trajectories of objects in videos. Chang *et al*. [236] propose an online video retrieval system supporting automatic object-based indexing and spatiotemporal queries. The system includes algorithms for automated video object segmentation and tracking. Bashir *et al*. [237] present a motion trajectory-based compact indexing and efficient retrieval mechanism for video sequences. Trajectories are represented by temporal orderings of subtrajectories. The subtrajectories are then represented by their principal component analysis coefficients. Chen and Chang [238] use wavelet decomposition to segment each trajectory and produce an index based on velocity features. Jung *et al*. [25] base their motion model on polynomial curve fitting. The motion model is used as an indexing key to access individual objects. Su *et al*. [26] construct motion flows from motion vectors embedded in MPEG bitstreams to generate continual motion information in the form of a trajectory. Given a trajectory, the system retrieves a set of trajectories that are similar to it. Hsieh *et al*. [27] divide trajectories into several small segments, and each segment is described by a semantic symbol. A distance measure combining an edit distance and a visual distance is exploited to match trajectories for video retrieval. The merit of trajectory-based features is that they can describe object actions. The limitation of these features is that their extraction depends on correct object segmentation and tracking and automatic recording of trajectories, all of which are still very challenging tasks.

3) Objects' Relationship-Based: These features describe spatial relationships between objects. Bimbo *et al.* [235] describe relationships between objects using a symbolic representation scheme which is applied to video retrieval. Yajima *et al*. [24] query the movements of multiple moving objects and specify the spatiotemporal relationships between objects by expressing each object's trace on a timeline. The merit of objects' relationship-based features is that they can intuitively represent relationships between multiple objects in the temporal domain. The limitation of these features is that it is difficult to label each object and its position.

IV. Video Data Mining, Classification, and Annotation

Video data mining, classification, and annotation rely heavily on video structure analysis and the extracted video features. There are no boundaries between video data mining, video classification, and video annotation. In particular, the concepts of video classification and annotation are very similar. In this section, we review the basic concepts and approaches for video data mining, classification, and annotation. The annotation is the basis for the detection of video's semantic concepts and the construction of semantic indices for videos.

A. Video Data Mining

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 [230], [232] and their associations, and other video semantic knowledge [45], in order to achieve video intelligent applications, such as video retrieval [118]. The choice of a strategy for video data mining depends on the application. Current strategies include the following.

1) Object Mining: Object mining is the grouping of different instances of the same object that appears in different parts in a video. It is very hard because the appearance of an object can change a great deal from one instance to another. Sivic and Zisserman [86] use a spatial neighborhood technique to cluster the features in the spatial domain of the frames. These clusters are used to mine frequently appearing objects in key frames. Anjulan and Canagarajah [81] extract stable tracks from shots. These stable tracks are combined into meaningful object clusters, which are used to mine similar objects. Quack *et al*. [28] present a method for mining frequently occurring objects and scenes from videos. Object candidates are detected by finding recurring spatial arrangements of affine covariant regions.

2) Special Pattern Detection: Special pattern detection applies to actions or events for which there are *a priori* models, such as human actions, sporting events [127], traffic events, or crime patterns. Laptev *et al*. [124] propose an appearance-based method that recognizes eight human actions in movies, e.g., answer phone, get out of a car, handshake, hug person, kiss. They extract local space-time features in space-time pyramids, build a spatial-temporal bag-of-features, and employ multichannel nonlinear SVMs for recognition. Ke *et al*. [125] propose a template-based method that recognizes human actions, such as picking up a dropped object or waving in a crowd. They oversegment the video to obtain spatial-temporal patches, and combine shape and optical flow cues to match testing patches and templates. Liu *et al*. [126] detect events in a football match, including penalty kicks, free kicks near the penalty box, and corner kicks in football games. Li and Porikli [128] detect six traffic patterns using a Gaussian mixture HMM framework, and Xie *et al*. [129] extract traffic jam events by analyzing the road background features. Nath [130] detects crime patterns using a clustering algorithm.

3) Pattern Discovery: Pattern discovery is the automatic discovery of unknown patterns in videos using unsupervised or semisupervised 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. The discovered patterns have the following applications: 1) detecting unusual events [230] that are often defined by their dissimilarity to discovered patterns; 2) associating clusters or patterns with words for video retrieval, etc; 3) building supervised classifiers based on the mined clusters for video classification or annotation, etc. Burl [105] describes an algorithm for mining motion trajectories to detect trigger events, determine typical or anomalous patterns of activities, classify activities into named categories, cluster activities, determine interactions between entities, etc. Hamid *et al*. [2] use *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 events. Turaga *et al*. [1] use a generative model to capture and represent a diverse class of activities, and build affine and view invariance of the activity into the distance metric for clustering. The clusters correspond to semantically meaningful activities. Cutler and Davis [14] compute an object's self-similarity as it evolves in time, and apply time–frequency analysis to detect and characterize the periodic motion. The periodicity is analyzed using the 2-D lattice structures inherent in similarity matrices.

4) Video Association Mining: Video association mining is mainly used to discover inherent relations between different events or the most frequent association patterns for different objects, such as the simultaneous occurrence of two objects, frequency of shot switches, and association between video types [118]. Video association mining also includes the deduction of interassociations between semantic concepts in the same shot from existing annotations or the inference of a semantic concept for the current shot from detection results of neighboring shots, etc. Pan and Faloutsos [102] propose an algorithm to find correlations between different events in news programs, such as those between "earthquake" and "volcano" or "tourism" and "wine." Zhu *et al*. [100] propose explicit definitions and evaluation measures for video associations by integrating distinct feature of the video data. Their algorithm introduces multilevel sequential association mining to explore associations between audio and visual cues, classifies the associations by assigning each of them a class label, and uses their appearances in the video to construct video indices. Yan *et al*. [13] describe various multiconcept relational learning algorithms based on a unified probabilistic graphical model representation and use graphical models to mine the relationship between video concepts. Liu *et al*. [231] use association-mining techniques to discover interconcept associations in the detected concepts, and mine intershot temporal dependence, in order to improve the accuracy of semantic concept detection.

5) Tendency Mining: Tendency mining is the detection and analysis of trends of certain events by tracking current events [118]. Xie *et al*. [103] propose a news video mining method, which involves two visualization graphs: the time-tendency graph and the time-space distribution graph. The time-tendency graph records the tendencies of events, while the time-space distribution graph records the spatial-temporal relations between various events. Oh and Bandi [104] mine the tendency of a traffic jam by analyzing the spatial-temporal relations between objects in videos.

6) Preference Mining: For news videos, movies, etc., the user's preferences can be mined [118]. For instance, Kules *et al*. [101] propose a personalized multimedia news portal to provide a personalized news service by mining the user's preferences.

B. Video Classification

The task of video classification [106], [245] 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.

Video content includes semantic content and editing effects. Referring to [23], semantic content classification can be performed on three levels: video genres, video events, and objects in the video, where genres have rougher and wider detection range; and events and objects have thinner and limited detection range. In the following, we discuss edit effect classification, genre classification, event classification, and object classification, respectively.

1) Edit Effect Classification: Editing effects depend on the ways for editing videos, such as camera motion and the composition of scenes and shots. Editing effects themselves are not a part of video content, but they influence the understanding of video content; therefore, they may be used in video semantic classification. For instance, Ekin *et al.* [165] classify shots of soccer videos into long, in-field medium, close-up, and out-of-field views using cinematic features and further detect events such as play, break, and replay. Xu *et al.* [246] use the domain-specific feature of grass-area-ratio to classify frames of soccer videos into global, zoom-in, and close-up views and obtain play/break statuses of games from the sequences of labeled frames. Tan *et al.* [247] estimate camera motion using data from the MPEG stream, and further classify basketball shots into wide-angle and close-up views and detect events such as fast breaks, shots at the basket, etc.

2) Video Genre Classification: Video genre classification is the classification of videos into different genres such as "movie," "news," "sports," and "cartoon." Approaches to classify video genres can be classified into statistic-based, rule- or knowledge-based, and machine learning-based [23].

a) Statistic-based approach: This approach classifies videos by statistically modeling various video genres. Fisher *et al.* [89] classify videos as news, car race, tennis, animated cartoon, and commercials. 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. Based on characteristics of films, Rasheed *et al.* [123] only use four visual features, namely average shot length, color variance, motion content, and lighting key, to classify films into comedies, actions, dramas, or horror films. The classification is achieved using mean shift clustering.

Some methods only utilize dynamic features to classify video genres. Roach *et al.* [122] propose a cartoon video classification method that uses motion features of foreground objects to distinguish between cartoons and noncartoons. Roach *et al.* [108] classify videos based on the dynamic content of short video sequences, where foreground object motion and background camera motion are extracted from videos. The classified videos include sports, cartoons, and news.

b) Rule- or knowledge-based approach: This approach applies heuristic rules from domain knowledge to low-level features to classify videos. Chen and Wong [109] develop a knowledge-based video classification method, in which the relevant knowledge is coded in the form of generative rules with confidences to form a rule-base. The Clip language is used to compile a video content classification system using the rule-base. Zhou *et al*. [110] propose a supervised rule-based video classification system, in which higher semantics are derived from a joint use of low-level features along with classification rules that are derived through a supervised learning process. Snoek *et al*. [93] propose a video classification and indexing method, combining video creation knowledge to extract semantic concepts from videos by exploring different paths through three consecutive analysis steps: the multimodel video content analysis step, the video style analysis step, and the context analysis step. Zhou *et al*. [107] propose a rule-based video classification system that applies video content analysis, feature extraction and clustering techniques to the semantic clustering of videos. Experiments on basketball videos are reported.

c) Machine learning-based approach: This approach uses labeled samples with low-level features to train a classifier or a set of classifiers for videos. Mittal and Cheong [112] use the Bayesian network to classify videos. The association between a continuous and nonparametric descriptor space and the classes is learned and the minimum Bayes error classifier is deduced. Qi *et al*. [97] propose a video classification framework using SVMs-based active learning. The results of clustering all the videos in the dataset are used as the input to the framework. The accuracy of the classifiers is improved gradually during the active-learning process. Fan *et al*. [98] use multiple levels of concepts of video contents to achieve hierarchical semantic classification of videos to enable highly efficient access to video contents. Truong *et al*. [90] classify videos into the genres of cartoons, commercials, music, news, and sports. The features used include the average shot length, the percentage of each type of transition, etc. The C4.5 decision tree is used to build the classifier for genre labeling. Yuan *et al*. [240] present an automatic video genre classification method based on a hierarchical ontology of video genres. A series of SVM classifiers united in a binary-tree form assign each video to its genre. Wu *et al*. [154] propose an online video semantic classification framework, in which local and global sets of optimized classification models are online trained by sufficiently exploiting both local and global statistic characteristics of videos. Yuan *et al*. [155] learn concepts from a large-scale imbalanced dataset using support cluster machines.

From the aforesaid video genres classification approaches, the following conclusions can be drawn [23]. 1) These approaches either use static features only, dynamic features only, or combine them both. 2) All the approaches preferably employ global statistical low-level features. This is because such features are robust to video diversity, making them appropriate for video genre classification. Many algorithms attempt to add some semantic features on the basis of these low-level features. 3) Prior domain knowledge is widely used in video genres classification. To use knowledge or rules can improve the classification efficiency for special domains, but the corresponding algorithms cannot be generalized to videos from other domains.

3) Event Classification: 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 subevents. To determine the classes of events in a video is an important component of content-based video classification [3], and it is connected with event detection in video data mining. There is a great deal of published work on event classification. Yu *et al.* [115] detect and track balls in broadcast soccer videos and extract ball trajectories, which are used to detect events such as hand ball and ball possession by a team. Chang *et al*. [111] detect and classify highlights in baseball game videos using HMM models that are learned from special shots identified as highlights. Duan *et al.* [116] propose a visual feature representation model for sports videos. This model is combined with supervised learning to perform a top-down semantic shot classification. These semantic shot classes are further used as a midlevel representation for high-level semantic analysis. Xu *et al*. [94] present an HMM-based framework for video semantic analysis. Semantics in different granularities are mapped to a hierarchical model in which a complex analysis problem is decomposed into subproblems. The framework is applied to basketball event detection. Osadchy and Keren [119] offer a natural extension of the "antiface" method to event detection, in both the gray-level and feature domains. Xie *et al*. [151] employ HMM and dynamic programming to detect the sports video concepts of "play," "no play," etc. Pan *et al.* [114] extract visual features and then use an HMM to detect slow-motion replays in sports videos.

From the aforesaid event classification algorithms, the following conclusions can be drawn [23]. 1) In contrast with genre classification, event classification needs more complex feature extraction. 2) Complicated motion measures are often attached to event classifiers. Some event classification methods employ only dynamic features, involving the accurate tracking of moving objects or rough region-based motion measures, and then classify the object motions in order to recognize motion events.

4) Object Classification: Video object classification which is connected with object detection in video data mining is conceptually the lowest grade of video classification. The most common detected and classified object is the face [120]. Object detection often requires the extraction of structural features of objects and classification of these features. Prior knowledge

such as an object appearance model is often incorporated into the process of object feature extraction and classification. Hong *et al.* [92] propose 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. A shot is classified by finding the best matching cluster. Dimitrova *et al.* [91] propose a method to classify four types of TV programs. Faces and texts are detected and tracked, and the number of faces and texts is used to label each frame of a video segment. An HMM is trained for each type using the frame labels as the observation symbols. The limitation of object classification for video indexing is that it is not generic; video object classification only works in specific environments.

C. Video Annotation

Video annotation [4], [117], [241] is the allocation of video shots or video segments to different predefined semantic concepts, such as person, car, sky, people walking. Video annotation is similar to video classification, except for two differences [239]: 1) Video classification has a different category/concept ontology compared with video annotation, although some of the concepts could be applied to both; and 2) video classification applies to complete videos, while video annotation applies to video shots or video segments. Video annotation and video classification share similar methodologies: First, low-level features are extracted, and then certain classifiers are trained and employed to map the features to the concept/category labels.

Corresponding to the fact that a video may be annotated with multiple concepts, the approaches for video annotation can be classified as isolated concept-based annotation, context-based annotation, and integrated-based annotation [244].

1) Isolated Concept-Based Annotation: This annotation method trains a statistical detector for each of the concepts in a visual lexicon, and the isolated binary classifiers are used individually and independently to detect multiple semantic concepts—correlations between the concepts are not considered. Feng *et al.* [8] use the multiple-Bernoulli distribution to model image and video annotation. The multiple-Bernoulli model explicitly focuses on the presence or absence of words in the annotation, based on the assumption that each word in an annotation is independent of the other words. Naphade and Smith [69] investigate the efficiencies of a large variety of classifiers, including GMM, HMM, kNN, and Adaboost, for each concept. Song *et al.* [9] introduce active learning together with semisupervised learning to perform semantic video annotation. In this method, a number of two-class classifiers are used to carry out the classification with multiple classes. Duan *et al.* [116] employ supervised learning algorithms based on the construction of effective midlevel representations to perform video semantic shot classification for sports videos. Shen *et al.* [73] propose a cross-training strategy to stack concept detectors into a single discriminative classifier and to handle the classification errors that occur when the classes overlap in the feature space. The limitation of isolated concept-based annotation is that the associations between the different concepts are not modeled.

2) Context-Based Annotation: To use contexts for different concepts [71] can improve concept detection performance. The task of context-based annotation is to refine the detection results of the individual binary classifiers or infer higher level concepts from detected lower level concepts using a context-based concept fusion strategy. For instance, Wu *et al*. [248] use an ontology-based learning method to detect video concepts. An ontology hierarchy is used to improve the detection accuracy of the individual binary classifiers. Smith and Naphade [249] construct model vectors based on the detection scores of individual classifiers to mine the unknown or indirect correlations between specific concepts and then train an SVM to refine the individual detection results. Jiang *et al*. [250] propose an active-learning method to annotate videos. In the method, users annotate a few concepts for a number of videos, and the manual annotations are then used to infer and improve detections of other concepts. Bertini *et al*. [251] propose an algorithm that uses pictorially enriched ontologies that are created by an unsupervised clustering method to perform automatic soccer video annotation. Occurrences of events or entities are automatically associated with higher level concepts, by checking their proximity to visual concepts that are hierarchically linked to higher level semantics. Fan *et al*. [32], [253] propose a hierarchical boosting scheme, which incorporates concept ontology and multitask learning, to train a hierarchical video classifier that exploits the strong correlations between video concepts. The limitation of context-based annotation is that the improvement of contextual correlations to individual detections is not always stable because the detection errors of the individual classifiers can propagate to the fusion step, and partitioning of the training samples into two parts for individual detections and conceptual fusion, respectively, causes that there are no sufficient samples for the conceptual fusion because of usual complexity of the correlations between the concepts.

3) Integration-Based Annotation: This annotation method simultaneously models both the individual concepts and their correlations: The learning and optimization are done simultaneously. The entire set of samples is used simultaneously to model the individual concepts and their correlations. Qi *et al*. [244] propose a correlative multilabel algorithm, which constructs a new feature vector that captures both the characteristics of concepts and the correlations between concepts. The limitation of the integration-based annotation is its high computational complexity.

The learning of a robust and effective detector for each concept requires a sufficiently large number of accurately labeled training samples, and the number required increases exponentially with the feature dimension. Recently, some approaches have been proposed to incorporate unlabeled data into the supervised learning process in order to reduce the labeling burden. Such approaches can be classified into semisupervised-based and active-learning-based.

a) Semisupervised learning: This approach uses unlabeled samples to augment the information in the available labeled examples. Yan and Naphade [74], [146] present semisupervised cross feature learning for cotraining-based video concept detection and investigate different labeling strategies in

cotraining involving unlabeled data and a small number of labeled videos. Yuan *et al*. [143] propose a feature selection-based manifold-ranking algorithm to learn concepts using a small number of samples. The algorithm consists of three major components: feature pool construction, prefiltering, and manifold ranking. Wang *et al*. [141] propose a video annotation algorithm, based on semisupervised learning by the kernel density estimation. Wang *et al*. [135], [279] propose an optimized multigraph-based semisupervised learning algorithm to deal with the insufficiency of training data in video annotation. Ewerth and Freisleben [167] propose a semisupervised learning method to adaptively learn the appearances of certain objects or events for a particular video. Adaboost and SVM are incorporated for feature selection and ensemble classification.

b) Active learning: Active learning is also an effective way to handle the lack of labeled samples. Song *et al*. [6] propose an active-learning algorithm for video annotation based on multiple complementary predictors and incremental model adaptation. Furthermore, Song *et al*. [7] propose a video annotation framework based on an active learning and semisupervised ensemble method, which is specially designed for personal video databases.

V. Query and Retrieval

Once video indices are obtained, content-based video retrieval [5] 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, etc. In the following, we review query types, similarity matching, and relevance feedback.

A. Query Types

Nonsemantic-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.

1) Query by Example: This query extracts low-level features from given example videos or images and similar videos are found by measuring feature similarity. The static features of key frames are suitable for query by example, as the key frames extracted from the example videos or exemplar images can be matched with the stored key frames.

2) 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. Hu *et al*. [36] propose a method of query by sketch, where trajectories drawn by users are matched to trajectories extracted from videos.

3) 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 [267]. In contrast with query by example and query by sketch, the search results of query by objects are the locations of the query object in the videos.

4) 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. Keywords can refer to video metadata, visual concepts, transcripts, etc. In this paper, we mainly consider visual concepts.

5) Query by Natural Language: This is the most natural and convenient way of making a query. Aytar *et al*. [255] use semantic word similarity to retrieve the most relevant videos and rank them, given a search query specified in the natural language (English). The most difficult part of a natural language interface is the parsing of natural language and the acquisition of accurate semantics.

6) Combination-Based Query: This query combines different types of queries such as text-based queries and video example-based queries. The combination-based query is adaptable to multimodel search. Kennedy *et al*. [259] develop a framework to automatically discovere useful query classes by clustering queries in a training set according to the performance of various unimodal search methods. Yan *et al*. [258] propose an adaptive method to fuse different search tools to implement query-class-dependent video retrieval, where the query-class association weights of the different search tools are automatically determined. Yuan *et al*. [219] classify the query space into person and nonperson queries in their multimedia retrieval system. Yan and Hauptmann [198] consider the classification of queries and the determination of combination weights in a probabilistic framework by treating query classes as latent variables.

The following query interfaces are among the most famous so far.

1) The Informedia interface [31], [70]: This interface supports filtering based on visual semantic concepts. The visual concept filters are applied after a keyword-based search is carried out.
2) The MediaMill query interface [30], [99]: This interface combines query-by-visual concept, query by example, and query by textual keyword.

B. Similarity Measure

Video similarity measures play an important role in content-based video retrieval. Methods 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.

1) Feature Matching: The most direct measure of similarity between two videos is the average distance between the features of the corresponding frames [34]. Query by example usually uses low-level feature matching to find relevant videos. However, video similarity can be considered in different levels of resolution or granularity [35].

According to different user' demands, static features of key frames [59], object features [81], and motion features [36] all can be used to measure video similarity. For example, Sivic *et al*. [18] extract face features from an example shot containing the queried face and match the extracted features with the stored face features. Then, shots containing the queried face are

retrieved. Lie and Hsiao [37] extract trajectory features of major objects in a given set of videos and match the extracted trajectory features with stored trajectory features to retrieve videos.

The merit of feature matching is that the video similarity can be conveniently measured in the feature space. Its limitation is that semantic similarity cannot be represented because of the gap between sets of feature vectors and the semantic categories familiar to people.

2) Text Matching: Matching the name of each concept with query terms is the simplest way of finding the videos that satisfy the query. Snoek *et al*. [242] 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. Finally, the concepts with the highest similarity are selected. The merits of the text-matching approach are its intuitiveness and simplicity of implementation. The limitation of this approach is that all related concepts must be explicitly included in the query text in order to obtain satisfactory search results.

3) Ontology-Based Matching: This approach achieves similarity matching using the ontology between semantic concepts or semantic relations between keywords. Query descriptions are enriched from knowledge sources, such as ontology of concepts or keywords. Snoek *et al*. [242] perform the syntactic disambiguation of the words in the text query and then translate the nouns and noun chunks extracted from the text to ontological concepts by looking up each noun in Wordnet. As the concepts are also linked to Wordnet, the ontology is used to determine which concepts are mostly related to the original query text. Based on the fact that the semantic word similarity is a good approximation for visual co-occurrence. Aytar *et al*. [255] utilize semantic word similarity measures to measure the similarity between text annotated videos and users' queries. Videos are retrieved based on their relevance to a user-defined text query. The merit of the ontology-based matching approach is that extra concepts from knowledge sources are used to improve retrieval results [221], [227]. The limitation of this approach is that irrelevant concepts are also likely to be brought in, perhaps leading to unexpected deterioration of search results.

4) Combination-Based Matching: This approach "leverages semantic concepts by learning the combination strategies from a training collection, e.g., learning query-independent combination models [222] and query-class-dependent combination models [258]" [229]. It is useful for combination-based queries that are adaptable to multimodal searches. The merits of the combination-based matching approach are that concept weights can be automatically determined and hidden semantic concepts can be handled to some extent. The limitation of this approach is that it is difficult to learn query combination models.

C. Relevance Feedback

In relevance feedback, 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. The refinement methods include query point optimization, feature weight adjustment, and information embedding. Relevance feedback bridges the gap between semantic notions of search relevance and the low-level representation of video content. Relevance feedback also reflects user's preferences by taking into account user feedback on the previously searched results. Like relevance feedback for image retrieval, relevance feedback for video retrieval can be divided into three categories: explicit, implicit, and pseudofeedback.

1) Explicit Relevance Feedback: This feedback asks the user to actively select relevant videos from the previously retrieved videos. Thi *et al*. [49] propose an interface for image and video retrieval. Users are required to choose positive samples and retrieval results are improved by modifying the query point toward the positive examples. Chen *et al*. [51] adjust the weights embedded in the similarity measure to reflect the user's feedback. The user can label sample videos as "highly relevant," "relevant," "no-opinion," "nonrelevant," or "highly nonrelevant," Sudha *et al*. [42] employ a simultaneous perturbation stochastic approximation-based algorithm to compute the optimal feature weights according to user's feedback. Aksoy and Cavus [55] describe a relevance feedback response technique that can adjust the weights of different features and different spatial locations in key frames according to the user's feedback. Browne and Smeaton [52] describe ostensive relevance feedback that takes into account the changes in the user's requirements that occur while users search for information. A decay function is used to weight the contribution of a previously viewed and relevant object to reflect the evolvement of the user's interest. Sav *et al*. [54], [56] present an interactive object-based video retrieval system that uses relevance feedback to refine an underlying model of the search object. The user can directly select the features important for the user and the image sections that the user wants to search for. The merit of explicit feedback is that it can obtain better results than implicit feedback or the pseudofeedback discussed later as it uses the user feedback directly. Its limitation is that it needs more user interaction, which requires more user patience and cooperation.

2) Implicit Relevance Feedback: This 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 [43]. Ghosh *et al*. [53] present a method to personalize video retrieval results based on click-through data analysis. The parameters of a Bayesian network that establishes uncertainties between concepts and video features are learned using the implicit user feedback. Hopfgartner *et al*. [62] propose a simple model to adapt retrieval results based on simulated implicit relevance feedback. The merit of implicit feedback is that it does not require the conscious cooperation of the user, making it more acceptable, available, and practicable than explicit feedback. The limitation of implicit feedback is that the information gathered from the user is less accurate than in explicit feedback.

3) Pseudorelevance Feedback: This feedback selects positive and negative samples from the previous retrieval results without the participation of the user. The positive samples are the ones near to the query sample in the feature space, and the

negative samples are far from the query sample. This way, the user's feedback is simulated. These samples are returned to the system for the second search. Muneesawang and Guan [50] present a self-training neural network-based relevance feedback that can obtain good retrieval performance with no user input. Forward and backward signal propagation is used to simulate the user's feedback. Yan *et al*. [33] propose an automatic retrieval method that learns an adaptive similarity space by automatically feeding back the bottom-ranked examples for negative feedback. Hauptmann *et al*. [252] develop a robust pseudorelevance feedback method called probabilistic local feedback based on a discriminative probabilistic retrieval framework. The proposed method is effective to improve retrieval accuracy without assuming that most of the top-ranked documents are relevant. The merit of pseudorelevance feedback is the substantial reduction in user interaction. It is limited in applications because of the semantic gap between low-level and high-level features: the similarities of low-level features obtained from different videos do not always coincide with the similarities between the videos defined by the user.

Active learning has been applied to relevance feedback for video retrieval. For example, Luan *et al*. [269] iteratively select videos that are the most relevant to the query until the number of videos labeled as relevant by users in an iteration step becomes very small. Then, the videos closest to the classifier boundary are returned to users for identification and the system is updated using the identified videos. Nguyen *et al*. [270] also use active learning in the interaction process to choose videos close to the classifier boundary. In contrast with the aforementioned algorithm that selects videos in the feature space, they choose videos in the dissimilarity space represented by a number of prototypes. Bruno *et al*. [271] design multimodal dissimilarity spaces for fast and efficient video retrieval. Different prototypes are learned for each modality. Videos are selected in the multimodal dissimilarity spaces based on the multimodal characteristics. In contrast with the traditional relevance feedback algorithms which select the most relevant videos which are ranked for further search, active-learning-based relevance feedback algorithms [268] usually return the videos closest to the classifier boundary to users for identification. Then, these most informative videos are used to improve the precision of the retrieval system. It has been verified that active-learning-based relevance feedback algorithms have a better performance than traditional relevance feedback algorithms [268].

VI. Video Summarization and Browsing

Video summarization [39], [156], [157], [181] removes the redundant data in videos and makes an abstract representation or summary of the contents, which is exhibited to users in a readable fashion to facilitate browsing. Video summarization complements video retrieval [183], by making browsing of retrieved videos faster, especially when the total size of the retrieved videos is large: The user can browse through the abstract representations to locate the desired videos. A detailed review on video browsing interfaces and applications can be found in [261].

There are two basic strategies for video summarization.

1) *Static video abstracts:* each of which consists of a collection of key frames extracted from the source video.
2) *Dynamic video skims:* each of which consists of a collection of video segments (and corresponding audio segments) that are extracted from the original video and then concatenated to form a video clip which is much shorter than the original video.

These two strategies can be combined to form hierarchical video summarizations. In the following, the different methods for video summarization are briefly reviewed. As video summarization is a research topic which is as large as video retrieval, we focus on reviewing papers published in the last four years, as a supplement to previous surveys [39], [181] on video summarization.

A. Key Frame-Based Static Abstracts

In recent years, many approaches [182], [189] have been proposed to organize extracted key frames into static video abstracts. These approaches include video table of contents [142], storyboard, and pictorial video summary. For instance, Xie and Wu [169] propose an algorithm to automatically generate the video summary for broadcast news videos. An affinity propagation-based clustering algorithm is used to group the extracted key frames into clusters, aiming to keep the pertinent key frames that distinguish one scene from the others and remove redundant key frames. Li *et al*. [164] propose a MinMax rate distortion optimization algorithm to find key frames for an optimal video summary. Optimal algorithms are developed to solve both the rate minimization and the distortion minimization formulations with different summarization rates. Calic *et al*. [175] propose a video key frame summarization and browsing algorithm that produces a comic-like representation of a video. The algorithm creates visual summaries in a user centered way. Guironnet *et al*. [161] propose a method for video summarization using camera motion analysis, based on rules to avoid temporal redundancy between the selected frames. Aner *et al*. [186] compose mosaics of scene backgrounds in sitcom programs. The mosaic images provide a compact static visual summary of the physical settings of scenes. Choudary and Liu [190] summarize the visual content in instructional videos, using extracted texts and figures. They match and mosaick the extracted key frames to reduce content redundancy and to build compact visual summaries. Christel *et al*. [276] have constructed a baseline rushes summarization system at TRECVid 2008 [272]. This baseline method simply presents the entire video at 50$\times$normal speed.

The merits of key frame-based static abstracts include the following. 1) The video content is displayed in a rapid and compact way, with no timing or synchronization issues, for browsing and navigation purposes. 2) Nonlinear browsing of video content is possible. 3) The total video content can be covered. The limitations of key frame-based static abstracts include the following. 1) Audio content in the original video is missing. 2) The dynamic visual content of videos cannot be described.

3) The abstracts are unnatural and hard to understand when the video is complex.

B. Dynamic Video Skimming

Dynamic video skimming [166], [168], [172], [173] condenses the original video into a much shorter version that consists of important segments selected from the original video. This shorter version can be used to browse or to guide the editing of the original video.

The merits of dynamic video skimming include the following. 1) It preserves the time-evolving nature of the original video. 2) Audio can be included in skims. 3) It is often more entertaining and interesting to watch a skim rather than a slide show of key frames. The limitations of dynamic video skimming include the following. 1) The sequential display of video skims is time-consuming. 2) The content integrity is sacrificed, while video highlights are emphasized.

There are three main approaches to video skimming: redundancy removal, object or event detection, and multimodal integration.

1) Redundancy Removal: This approach removes uninformative or redundant video segments from the original video and retains the most informative video segments that are concatenated to form a skim. For example, Xiao *et al*. [158] extract repeating patterns from a video. A video shot importance evaluation model is used to select the most informative video shots to construct the video summary. Ngo *et al*. [160] represent a video as a complete undirected graph and use the normalized cut algorithm to optimally partition the graph into video clusters. At most one shot is retained from each cluster of visually similar shots in order to eliminate redundant shots. Gao *et al*. [174] propose a video summarization algorithm suitable for personal video recorders. In the algorithm, according to the defined impact factors of scenes and key frames, parts of shots are selected to generate an initial video summary. Then, repetitive frame segment detection is applied to remove redundant information from the initial video summary. Wang and Ngo [176] have proposed an algorithm for rushes summarization task in TRECVid-2007 [177]–[179]. In the algorithm, undesirable shots are filtered out and the intershot redundancy is minimized by detecting and then removing repetitive shots. The most representative video segments are selected for summarization using techniques such as object detection, camera motion estimation, key-point matching, and tracking. Liu *et al*. [274] have devised an algorithm for the rushes summarization task in TRECVid-2008 [272]. They first segment a video into shots and then use a clustering algorithm to find and remove similar shots. Then, saliency detection is applied to detect the most informative shots to be included in the summary. Chasanis *et al*. [275] measure similarities between shots based on the similarities between key frames and employ them to remove repeated shots.

2) Object or Event Detection: Semantic primitives in videos, such as relevant objects, actions, and events, can be used in highlight-preserving video skims. In the object or event detection-based approach, video segments are selected according to the results of video segment classification and object or event detection. Detected objects and events are ranked to create the video skim. For example, in skimming sports videos, goals, fouls, touchdowns, etc. are detected as important events. Ekin *et al*. [165] propose a framework to skim soccer videos through dominant color region detection, robust shot boundary detection, shot classification, goal detection, referee detection, and penalty-box detection. As an example of object-based skimming, Peker *et al*. [188] propose a video skimming algorithm using face detection on broadcast video programs. In the algorithm, faces are the primary targets, as they constitute the focus of most consumer video programs. Naci *et al*. [277] extract features using face detection, camera motion, and MPEG-7 color layout descriptors of each frame. A clustering algorithm is employed to find and then remove repeated shots. Bailer and Thallinger [273] compare two content selection methods for skimming rush videos. One approach is rule-based, and the other is HMM-based. A face detection module is employed to help select important segments to be included in the summary.

3) Multimodal Integration: For videos whose content is largely contained in the audio, such as news programs and documentaries, the spoken texts can assist video summarization. Once caption texts or speech transcripts in a video are available, a text summary can be integrated with the visual summary into the video skim, or the video sections corresponding to the selected texts can be concatenated to generate the video skim. For instance, Taskiran *et al*. [184] divide a video into segments by pause detection, and derive a score for each segment according to the frequencies of the words in the audio track for the segment. A summary is produced by selecting the segments with the highest scores while maximizing the coverage of the summary over the full video. Gong [159] summarizes the audio and visual content of a source video separately and then integrates the two summaries using a bipartite graph. The audio content summarization is achieved by selecting representative spoken sentences from the audio track, while the visual content summarization is achieved by preserving visually distinct contents from the image track.

C. Hierarchical Summarization

Hierarchical video summaries can be obtained from key frames or from static video abstracts combined with video skimming. For instance, Taskiran *et al*. [185] cluster key frames extracted from shots using color, edge, and texture features and present them in a hierarchical fashion using a similarity pyramid. Geng *et al*. [187] propose a hierarchical video summarization algorithm based on video structure and highlights. In the algorithm, video structure units (frame, shot, and scene) are ranked using visual and audio attention models. According to the measured ranks, the skim ratio and the key frame ratio of the different video structure units are calculated and used to construct summaries at different levels in a hierarchical video summary. Ciocca and Schettini [171] remove meaningless key frames using supervised classification on the basis of pictorial features derived directly from the frames, together with other features derived from the processing of the frames by a visual attention model. Then, the key frames are grouped into clusters

to allow multilevel summary using both low-level and high-level features.

VII. Future Developments

Although a large amount of work has been done in visual content-based video indexing and retrieval, many issues are still open and deserve further research, especially in the following areas.

1) Motion Feature Analysis. 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.

2) Hierarchical Analysis of Video Contents. One video may contain different meanings at different semantic levels. Hierarchical organization of video concepts is required for semantic-based video indexing and retrieval. Hierarchical analysis requires the decomposition of high-level semantic concepts into a series of low-level basic semantic concepts and their constraints. Low-level basic semantic concepts can be directly associated with low-level features, and high-level semantic concepts can be deduced from low-level basic semantic concepts by statistical analysis. In addition, building hierarchical semantic relations between scenes, shots, and key frames, on the basis of video structural analysis; establishing links between classifications with the three different levels: genres, event and object; and hierarchically organizing and visualizing retrieval results are all interesting research issues.

3) Hierarchical Video Indices. Corresponding to hierarchical video analysis, hierarchical video indices can be utilized in video indexing. The lowest layer in the hierarchy is the index store model corresponding to the high-dimensional feature index structure. The highest layer is the semantic index model describing the semantic concepts and their correlations in the videos to be retrieved. The middle layer is the index context model that links the semantic concept model and the store model. 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.

4) Fusion of Multimodels. The semantic content of a video is usually an integrated expression of multiple models. Fusion of information from multiple models can be useful in content-based video retrieval [38], [95]. Description of temporal relations between different kinds of information from multiple models, dynamic weighting of features of different models, fusion of information from multiple models that express the same theme, and fusion of multiple model information in multiple levels are all difficult issues in the fusion analysis of integrated models.

5) Semantic-Based Video Indexing and Retrieval. Current approaches for semantic-based video indexing and retrieval usually utilize a set of texts to describe the visual contents of videos. Although many automatic semantic concept detectors have been developed, there are many unanswered questions: How to select the features that are the most representative of semantic concepts? How should large-scale concept ontology for videos [76] be constructed? How to choose useful generic concept detectors with high retrieval utility? How many useful concepts are needed [243]? How can high-level concepts be automatically incorporated into video retrieval? How can ontology [79], [80] be constructed for translating the query into terms that a concept detector set can handle? How can inconsistent annotations resulting from different people's interpretations of the same visual data be reconciled? How can elaborate ontology be established between the detector lexica? How can multimodality fusion be used to detect concepts more accurately? How can different machine learning approaches be fused to obtain more accurate concepts?

6) Extensible Video Indexing. 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. Feature extraction with less domain knowledge and dynamic construction of classification rules using rule-mining techniques may eliminate this dependence.

7) Multimodel Human–Computer Interface. A multimodel human–computer interface can convey the query intentions more accurately and improve the accuracy of the retrieval results. Furthermore, the video output with multimodel representation is more visual and vivid. 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.

8) Combination of Perception with Video Retrieval. It is interesting to simulate human perception to exploit new video retrieval approaches. The research in visual perception shows that the human vision system quickly identifies the salient image regions [46]. This ability extends from salient objects (i.e., faces) in video scenes to significant behaviors in video sequences. New video retrieval schemes could be based on the detection and extraction of image regions or video behaviors that attract users' attention [46], [162].

9) Affective Computing-Based Video Retrieval. 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. To combine affective semantics from the spoken language in the audio track with visual affective semantics [75], to utilize cognitive models, cultural backgrounds, aesthetic criteria, cold and warm tones, psychology, or photography [88], and to understand affective semantics in videos are very interesting research issues.

10) Distributed Network Video Retrieval. Network video retrieval should adopt distributed frameworks [48] rather than conventional centralized frameworks. Distributed network retrieval is composed of two stages: video gathering and video content analysis. In video gathering, the bandwidth of

transportation and storage is kept low by using only real-time online methods to analyze the videos. In video content analysis, more elaborate methods are used to search for specific objects and to categorize the videos. The video content analysis uses key data extracted during video gathering. 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.

11) Social Network-Based Video Retrieval. Besides optimizing the performance of each local video indexing and retrieval system, local systems can be integrated into a network characterized by distributed and collaborative intelligence to achieve more accurate retrieval. It is very interesting to implement such integration in the context of a social network in which users label video contents which they recommend to each other through collaborative filtering or in the context of a video content inference network in which video content inference systems learn from each other and complement each other [29]. Such integration by a social network can even take the social or cultural context into consideration during video retrieval.

VIII. Conclusion

We have presented a review on recent developments in visual content-based video indexing and retrieval. The state of the art of existing approaches in each major issue has been described with the focus on the following tasks: video structure analysis including shot boundary detection, key frame extraction and scene segmentation, extraction of features of static key frames, objects and motions, video data mining, video classification and annotation, video search including interface, similarity measure and relevance feedback, and video summarization and browsing. At the end of this survey, we have discussed future directions such as affective computing-based video retrieval and distributed network video retrieval.

References

[1] P. Turaga, A. Veeraraghavan, and R. Chellappa, "From videos to verbs: Mining videos for activities using a cascade of dynamical systems," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2007, pp. 1–8.

[2] R. Hamid, S. Maddi, A. Bobick, and M. Essa, "Structure from statistics—Unsupervised activity analysis using suffix trees," in *Proc. IEEE Int. Conf Comput. Vis.*, Oct., 2007, pp. 1–8.

[3] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 39, no. 5, pp. 489–504, Sep. 2009.

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

[5] X. Chen, C. Zhang, S. C. Chen, and S. Rubin, "A human-centered multiple instance learning framework for semantic video retrieval," *IEEE Trans. Syst, Man, Cybern., C: Appl. Rev.*, vol. 39, no. 2, pp. 228–233, Mar. 2009.

[6] Y. Song, X.-S. Hua, L. Dai, and M. Wang, "Semi-automatic video annotation based on active learning with multiple complementary predictors," in *Proc. ACM Int. Workshop Multimedia Inf. Retrieval*, Singapore, 2005, pp. 97–104.

[7] Y. Song, X.-S. Hua, G.-J. Qi, L.-R. Dai, M. Wang, and H.-J. Zhang, "Efficient semantic annotation method for indexing large personal video database," in *Proc. ACM Int. Workshop Multimedia Inf. Retrieval*, Santa Barbara, CA, 2006, pp. 289–296.

[8] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun./Jul. 2004, vol. 2, pp. 1002–1009.

[9] Y. Song, G.-J. Qi, X.-S. Hua, L.-R. Dai, and R.-H. Wang, "Video annotation by active learning and semi-supervised ensembling," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2006, pp. 933–936.

[10] C. H Yeo, Y. W. Zhu, Q. B. Sun, and S. F Chang, "A Framework for sub-window shot detection," in *Proc. Int. Multimedia Modelling Conf.*, Jan. 2005, pp. 84–91.

[11] G. Camara-Chavez, F. Precioso, M. Cord, S. Phillip-Foliguet, and A. de A. Araujo, "Shot boundary detection by a hierarchical supervised approach," in *Proc. Int. Conf. Syst., Signals Image Process.*, Jun. 2007, pp. 197–200.

[12] H. Lu, Y.-P. Tan, X. Xue, and L. Wu, "Shot boundary detection using unsupervised clustering and hypothesis testing," in *Proc. Int. Conf. Commun. Circuits Syst.*, Jun. 2004, vol. 2, pp. 932–936.

[13] R. Yan, M.-Y. Chen, and A. G. Hauptmann, "Mining relationship between video concepts using probabilistic graphical model," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2006, pp. 301–304.

[14] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 781–796, Aug. 2000.

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

[16] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang, "A formal study of shot boundary detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 2, pp. 168–186, Feb. 2007.

[17] R. Visser, N. Sebe, and E. M. Bakker, "Object recognition for video retrieval," in *Proc. Int. Conf. Image Video Retrieval*, London, U.K., Jul. 2002, pp. 262–270.

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

[19] D.-D. Le, S. Satoh, and M. E. Houle, "Face retrieval in broadcasting news video by fusing temporal and intensity information," in *Proc. Int. Conf. Image Video Retrieval*, (Lect. Notes Comput. Sci.), 4071, Jul. 2006, pp. 391–400.

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

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

[22] M. S. Dao, F. G. B. DeNatale, and A. Massa, "Video retrieval using video object-trajectory and edge potential function," in *Proc. Int. Symp. Intell. Multimedia, Video Speech Process.*, Oct. 2004, pp. 454–457.

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

[24] C. Yajima, Y. Nakanishi, and K. Tanaka, "Querying video data by spatio-temporal relationships of moving object traces," in *Proc. Int. Federation Inform. Process. TC2/WG2.6 Working Conf. Visual Database Syst.*, Brisbane, Australia, May 2002, pp. 357–371.

[25] Y. K. Jung, K. W. Lee, and Y. S. Ho, "Content-based event retrieval using semantic scene interpretation for automated traffic surveillance," *IEEE Trans. Intell. Transp. Syst.*, vol. 2, no. 3, pp. 151–163, Sep. 2001.

[26] C.-W. Su, H.-Y. M. Liao, H.-R. Tyan, C.-W. Lin, D.-Y. Chen, and K.-C. Fan, "Motion flow-based video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 6, pp. 1193–1201, Oct. 2007.

[27] J.-W. Hsieh, S.-L. Yu, and Y.-S. Chen, "Motion-based video retrieval by trajectory matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 3, pp. 396–409, Mar. 2006.

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

[29] A. Hanjalic, R. Lienhart, W.-Y. Ma, and J. R. Smith, "The holy grail of multimedia information retrieval: So close or yet so far away?" *Proc. IEEE*, vol. 96, no. 4, pp. 541–547, Apr. 2008.

[30] M. Worring, C. Snoek, O. de Rooij, G. P. Nguyen, and A. Smeulders, "The mediamill semantic video search engine," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2007, vol. 4, pp. IV.1213–IV.1216.

[31] M. G. Christel and R. M. Conescu, "Mining novice user activity with TRECVID interactive retrieval tasks," in *Proc. Int. Conf. Image Video Retrieval*, Tempe, AZ, Jul. 2006, pp. 21–30.

[32] J. Fan, H. Luo, Y. Gao, and R. Jain, "Incorporating concept ontology to boost hierarchical classifier training for automatic multi-level annotation," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 939–957, Aug. 2007.

[33] R. Yan, A. G. Hauptmann, and R. Jin, "Negative pseudo-relevance feedback in content-based video retrieval," in *Proc. ACM Int. Conf. Multimedia*, Berkeley, CA, Nov. 2003, pp. 343–346.

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

[35] R. Lienhart, "A system for effortless content annotation to unfold the semantics in videos," in *Proc. IEEE Workshop Content-Based Access Image Video Libraries*, Jun. 2000, pp. 45–49.

[36] W. M. Hu, D. Xie, Z. Y. Fu, W. R. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *IEEE Trans. Image Process.*, vol. 16, no. 4, pp. 1168–1181, Apr. 2007.

[37] W. N. Lie and W. C. Hsiao, "Content-based video retrieval based on object motion trajectory," in *Proc. IEEE Workshop Multimedia Signal Process.*, Dec. 2002, pp. 237–240.

[38] C. Snoek, M. Worring, and A. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. ACM Int. Conf. Multimedia*, Singapore, 2005, pp. 399–402.

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

[40] S. Bruyne, D. Deursen, J. Cock, W. Neve, P. Lambert, and R. Walle, "A compressed-domain approach for shot boundary detection on H.264/AVC bit streams," *J. Signal Process.: Image Commun.*, vol. 23, no. 7, pp. 473–489, 2008.

[41] C.-W. Ngo, "A robust dissolve detector by support vector machine," in *Proc ACM Int. Conf. Multimedia*, 2003, pp. 283–286.

[42] V. Sudha, B. Shalabh, S. V. Basavaraja, and V. Sridhar, "SPSA-based feature relevance estimation for video retrieval," in *Proc. IEEE Workshop Multimedia Signal Process.*, Cairns, Qld., Oct. 2008, pp. 598–603.

[43] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. ACM Conf. Knowl. Discovery Data Mining*, Edmonton, AB, 2002, pp. 133–142.

[44] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 2, no. 1, pp. 1–19, Feb. 2006.

[45] T. Mei, X. S. Hua, H. Q. Zhou, and S. P. Li, "Modeling and mining of users' capture intention for home videos," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 66–76, Jan. 2007.

[46] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[47] Choi, K.-C. Ko, Y.- M. Cheon, G-Y. Kim, H-Il, S.-Y. Shin, and Y.-W. Rhee, "Video shot boundary detection algorithm," *Comput. Vis., Graph. Image Process.*, (Lect. Notes Comput. Sci.), 4338, pp. 388–396, 2006.

[48] C.-Y. Chen, J.-C. Wang, and J.-F. Wang, "Efficient news video querying and browsing based on distributed news video servers," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 257–269, Apr. 2006.

[49] L. L. Thi, A. Boucher, and M. Thonnat, "An interface for image retrieval and its extension to video retrieval," in *Proc. Nat. Symp. Res., Develop. Appl. Inform. Commun. Technol.*, May 2006, pp. 278–285.

[50] P. Muneesawang and L. Guan, "Automatic relevance feedback for video retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2003, vol. 3, pp. III.1–III.4.

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

[52] P. Browne and A. F. Smeaton, "Video information retrieval using objects and ostensive relevance feedback," in *Proc. ACM Symp. Appl. Comput.*, Nicosia, Cyprus, Mar. 2004, pp. 1084–1090.

[53] H. Ghosh, P. Poornachander, A. Mallik, and S. Chaudhury, "Learning ontology for personalized video retrieval," in *Proc. ACM Workshop Multimedia Inform., Retrieval*, Augsburg, Germany, Sep. 2007, pp. 39–46.

[54] S. Sav, H. Lee, A. F. Smeaton, N. O'Connor, and N. Murphy, "Using video objects and relevance feedback in video retrieval," in *Proc. SPIE—Internet Multimedia Manag. Syst. VI*, Boston, MA, Oct. 2005, vol. 6015, pp. 1–12.

[55] S. Aksoy and O. Cavus, "A relevance feedback technique for multimodal retrieval of news videos," in *Proc. Int. Conf. Comput. Tool*, Nov. 2005, vol. 1, pp. 139–142.

[56] S. Sav, H. Lee, N. O'Connor, and A. F. Smeaton, "Interactive object-based retrieval using relevance feedback," in *Proc. Adv. Concepts Intell. Vis. Syst.*, (Lect. Notes Comput. Sci.), 3708, Oct. 2005, pp. 260–267.

[57] U. Damnjanovic, E. Izquierdo, and M. Grzegorzek, "Shot boundary detection using spectral clustering," in *Proc. Eur. Signal Process. Conf.*, Poznan, Poland, Sep. 2007, pp. 1779–1783.

[58] X. Ling, L. Chao, H. Li, and X. Zhang, "A general method for shot boundary detection," in *Proc. Int. Conf. Multimedia Ubiquitous Eng.*, 2008, pp. 394–397.

[59] Y. Wu, Y. T. Zhuang, and Y. H. Pan, "Content-based video similarity model," in *Proc. ACM Int. Conf. Multimedia*, 2000, pp. 465–467.

[60] H. Koumaras, G. Gardikis, G. Xilouris, E. Pallis, and A. Kourtis, "Shot boundary detection without threshold parameters," *J. Electron. Imag.*, vol. 15, no. 2, pp. 020503-1–020503-3, May 2006.

[61] Z.-C. Zhao, X. Zeng, T. Liu, and A.-N. Cai, "BUPT at TRECVID 2007: Shot boundary detection," in *Proc. TREC Video Retrieval Eval.*, 2007, Available: http://www-nlpir.nist.gov/projects/tvpubs/tv7.papers/bupt.pdf.

[62] F. Hopfgartner, J. Urban, R. Villa, and J. Jose, "Simulated testing of an adaptive multimedia information retrieval system," in *Proc. Int. Workshop Content-Based Multimedia Indexing*, Bordeaux, France, Jun. 2007, pp. 328–335.

[63] A. Herout, V. Beran, M. Hradis, I. Potucek, P. Zemcík, and P. Chmelar, "TRECVID 2007 by the Brno Group," in *Proc. TREC Video Retrieval Eval.*, 2007. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv7.papers/brno.pdf

[64] C. Lo and S.-J. Wang, "Video segmentation using a histogram-based fuzzy C-means clustering algorithm," in *Proc. IEEE Int. Fuzzy Syst. Conf.*, Dec. 2001, pp. 920–923.

[65] A. Hanjalic, "Shot-boundary detection: Unraveled and resolved?," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 2, pp. 90–105, Feb. 2002.

[66] 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, 2007.

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

[68] L. Bai, S.-Y. Lao, H.-T. Liu, and J. Bu, "Video shot boundary detection using petri-net," in *Proc. Int. Conf. Mach. Learning Cybern.*, 2008, pp. 3047–3051.

[69] M. R. Naphade and J. R. Smith, "On the detection of semantic concepts at TRECVID," in *Proc. ACM Int. Conf. Multimedia*, New York, 2004, pp. 660–667.

[70] M. Christel, C. Huang, N. Moraveji, and N. Papernick, "Exploiting multiple modalities for interactive video retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, QC, Canada, 2004, vol. 3, pp. 1032–1035.

[71] L. Hollink, M. Worring, and A. T. Schreiber, "Building a visual ontology for video retrieval," in *Proc. ACM Int. Conf. Multimedia*, Singapore, 2005, pp. 479–482.

[72] C. Liu, H. Liu, S. Jiang, Q. Huang, Y. Zheng, and W. Zhang, "JDL at TRECVID 2006 shot boundary detection," in *Proc. TREC Video Retrieval Eval. Workshop*, 2006. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv6.papers/cas_jdl.pdf

[73] X. Shen, M. Boutell, J. Luo, and C. Brown, "Multi-label machine learning and its application to semantic scene classification," in *Proc. Int. Symp. Electron. Imag.*, Jan. 2004, pp. 188–199.

[74] R. Yan and M. Naphade, "Co-training non-robust classifiers for video semantic concept detection," in *Proc. IEEE Int. Conf. Image Process.*, Singapore, 2005, vol. 1, pp. 1205–1208.

[75] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, Feb. 2005.

[76] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, Jul./Sep. 2006.

[77] X. Wu, P. C. Yuan, C. Liu, and J. Huang, "Shot boundary detection: An information saliency approach," in *Proc. Congr. Image Signal Process.*, 2008, vol. 2, pp. 808–812.

[78] R. V. Babu and K. R. Ramakrishnan, "Compressed domain video retrieval using object and global motion descriptors," *Multimedia Tools Appl.*, vol. 32, no. 1, pp. 93–113, 2007.

[79] J. Fan, H. Luo, Y. Gao, and R. Jain, "Incorporating concept ontology for hierarchical video classification, annotation, and visualization," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 939–957, 2007.

[80] D. Vallet, P. Castells, M. Fernandez, P. Mylonas, and Y. Avrithis, "Personalized content retrieval in context using ontological knowledge," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 336–345, Mar. 2007.

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

[82] X. B. Gao, J. Li, and Y. Shi, "A video shot boundary detection algorithm based on feature tracking," in *Proc. Int. Conf. Rough Sets Knowl. Technol.*, (Lect. Notes Comput. Sci.), 4062, 2006, pp. 651–658.

[83] Y. Chang, D. J. Lee, Y. Hong, and J. Archibald, "Unsupervised video shot detection using clustering ensemble with a color global scale-invariant feature transform descriptor," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, 2008.

[84] G. C. Chavez, F. Precioso, M. Cord, S. P. -Foliguet, and A. de A. Araujo, "Shot boundary detection at TRECVID 2006," in *Proc. TREC Video Retrieval Eval.*, 2006. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv6.papers/dokuz.pdf

[85] Z.-C. Zhao and A.-N. Cai, "Shot boundary detection algorithm in compressed domain based on adaboost and fuzzy theory," in *Proc. Int. Conf. Nat. Comput.*, 2006, pp. 617–626.

[86] J. Sivic and A. Zisserman, "Video data mining using configurations of viewpoint invariant regions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2004, vol. 1, pp. I-488–I-495.

[87] C. H. Hoi, L. S. Wong, and A. Lyu, "Chinese university of Hong Kong at TRECVID 2006: Shot boundary detection and video search," in *Proc. TREC Video Retrieval Eval.*, 2006. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv6.papers/chinese_uhk.pdf

[88] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 689–704, Jun. 2006.

[89] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," in *Proc. ACM Int. Conf. Multimedia*, 1995, pp. 367–368.

[90] B. T. Truong, C. Dorai, and S. Venkatesh, "Automatic genre identification for content-based video categorization," in *Proc. IEEE Int. Conf. Pattern Recog.*, vol. 4, Barcelona, Spain, 2000, pp. 230–233.

[91] N. Dimitrova, L. Agnihotri, and G. Wei, "Video classification based on HMM using text and faces," in *Proc. Eur. Signal Process. Conf.*, Tampere, Finland, 2000, pp. 1373–1376.

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

[93] C. G. M. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders, "The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1678–1689, Oct. 2006.

[94] G. Xu,Y.-F. Ma, H.-J. Zhang, and Sh.-Q. Yang, "An HMM-based framework for video semantic analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 11, pp. 1422–1433, Nov. 2005.

[95] Y. T. Zhuang, C. M. Wu, F. Wu, and X. Liu, "Improving web-based learning: Automatic annotation of multimedia semantics and cross-media indexing," in *Proc. Adv. Web-Based Learning – ICWL*, (Lect. Notes Comput. Sci.), 3143, 2004, pp. 255–262.

[96] Y. X. Peng and C.-W. Ngo, "Hot event detection and summarization by graph modeling and matching," in *Proc. Int. Conf. Image Video Retrieval*, Singapore, Jul. 2005, pp. 257–266.

[97] G.-J. Qi, Y. Song, X.-S. Hua, H.-J. Zhang, and L.-R. Dai, "Video annotation by active learning and cluster tuning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshop*, Jun. 2006, pp. 114–121.

[98] J. P. Fan, A. K. Elmagarmid, X. Q. Zhu, W. G. Aref, and L. D. Wu, "ClassView: Hierarchical video shot classification, indexing, and accessing," *IEEE Trans. Multimedia*, vol. 6, no. 1, pp. 70–86, Feb. 2004.

[99] C. G. M. Snoek, M. Worring, D. C. Koelma, and A. W. M. Smeulders, "A learned lexicon-driven paradigm for interactive video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 280–292, Feb. 2007.

[100] X. Q. Zhu, X. D. Wu, A. K. Elmagarmid, Z. Feng, and L. D. Wu, "Video data mining: Semantic indexing and event detection from the association perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 5, pp. 665–677, May 2005.

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

[102] J. Y. Pan and C. Faloutsos, "GeoPlot: Spatial data mining on video libraries," in *Proc. Int. Conf. Inform. Knowl. Manag.*, 2002, pp. 405–412.

[103] Y.-X. Xie, X.-D. Luan, S.-Y. Lao, L.-D. Wu, X. Peng, and Z.-G. Han, "A news video mining method based on statistical analysis and visualization," in *Proc. Int. Conf. Image Video Retrieva1*, Jul. 2004, pp. 115–122.

[104] J.-H. Oh and B. Bandi, "Multimedia data mining framework for raw video sequences," in *Proc. ACM Int. Workshop Multimedia Data Mining*, Edmonton, AB, Canada, 2002, pp. 18–35.

[105] M. C. Burl, "Mining patterns of activity from video data," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2004, pp. 532–536.

[106] M. Roach, J. Mason, L.-Q. Xu, and F. Stentiford, "Recent trends in video analysis: A taxonomy of video classification problems," in *Proc. Int. Assoc. Sci. Technol. Develop. Int. Conf. Internet Multimedia Syst. Appl.*, Honolulu, HI, Aug. 2002, pp. 348–354.

[107] W. S. Zhou, A. Vellaikal, and C.-C. J. Kuo, "Rule-based video classification system for basketball video indexing," in *Proc. ACM Workshops Multimedia*, 2000, pp. 213–216.

[108] M. J. Roach, J. D. Mason, and M. Pawlewski, "Video genre classification using dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 3, pp. 1557–1560.

[109] Y. Chen and E. K. Wong, "A knowledge-based approach to video content classification," in *Proc. SPIE Vol. 4315: Storage and Retrieval for Media Databases*, Jan. 2001, pp. 292–300.

[110] W. S. Zhou, S. Dao, and C. C. J. Kuo, "On-line knowledge- and rule-based video classification system for video indexing and dissemination," *Inform. Syst.*, vol. 27, no. 8, pp. 559–586, Dec. 2002.

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

[112] A. Mittal and L. F. Cheong, "Addressing the problems of Bayesian network classification of video using high dimensional features," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 2, pp. 230–244, Feb. 2004.

[113] S. Chantamunee and Y. Gotoh, "University of Sheffield at TRECVID 2007: Shot boundary detection and rushes summarisation," in *Proc. TREC Video Retrieval Eval.*, 2007. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv7.papers/sheffield_university.pdf

[114] H. Pan, P. Van Beek, and M. I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, May 2001, pp. 1649–1652.

[115] X. Yu, C. Xu, H. W. Leong, Q. Tian, Q. Tang, and K. Wan, "Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video," in *Proc. ACM Int. Conf. Multimedia*, Berkeley, CA, 2003, pp. 11–20.

[116] L. Y. Duan, M. Xu, Q. Tian, and C. Xu, "A unified framework for semantic shot classification in sports video," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1066–1083, Dec. 2005.

[117] C. S. Xu, J. J. Wang, H. Q. Lu, and Y. F. Zhang, "A novel framework for semantic annotation and personalized retrieval of sports video," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 421–436, Apr. 2008.

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

[119] M. Osadchy and D. Keren, "A rejection-based method for event detection in video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 4, pp. 534–541, Apr. 2004.

[120] U. Park, H. Chen, and A. K. Jain, "3D model-assisted face recognition in video," in *Proc. Workshop Face Process. Video*, Victoria, BC, Canada, May 2005, pp. 322–329.

[121] J. S. Boreczky and L. D. Wilcox, "A hidden Markov model framework for video segmentation using audio and image features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1998, vol. 6, pp. 3741–3744.

[122] M. J. Roach, J. S. D. Mason, and M. Pawlewski, "Motion-based classification of cartoons," in *Proc. Int. Symp. Intell. Multimedia*, 2001, pp. 146–149.

[123] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 52–64, Jan. 2005.

[124] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[125] Y. Ke, R. Sukthankar, and M. Hebert, "Event detection in crowded videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[126] H.-Y. Liu, T. T He, and Z. Hui, "Event detection in sports video based on multiple feature fusion," in *Proc. Int. Conf. Fuzzy Syst. Knowl. Discovery*, 2007, vol. 2, pp. 446–450.

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

[128] X. K. Li and F. M. Porikli, "A hidden Markov model framework for traffic event detection using video features," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2004, vol. 5, pp. 2901–2904.
[129] L. Xie, Q. Wu, X. M Chu, J. Wang, and P. Cao, "Traffic jam detection based on corner feature of background scene in video-based ITS," in *Proc. IEEE Int. Conf. Netw., Sens. Control*, Apr. 2008, pp. 614–619.
[130] S. V. Nath, "Crime pattern detection using data mining," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. Workshops*, 2006, pp. 41–44.
[131] H.-W. Yoo, H.-J. Ryoo, and D.-S. Jang, "Gradual shot boundary detection using localized edge blocks," *Multimedia Tools*, vol. 28, no. 3, pp. 283–300, Mar. 2006.
[132] Y. Zhai and M. Shah, "Video scene segmentation using Markov chain Monte Carlo," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 686–697, Aug. 2006.
[133] Z. Rasheed and M. Shah, "Scene detection in Hollywood movies and TV shows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2003, vol. 2, pp. 343–350.
[134] L. Zhao, W. Qi, Y.-J. Wang, S.-Q. Yang, and H.-J. Zhang, "Video shot grouping using best first model merging," in *Proc. Storage Retrieval Media Database*, 2001, pp. 262–269.
[135] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L. R. Dai, "Optimizing multi-graph learning: Towards a unified video annotation scheme," in *Proc. ACM Int. Conf. Multimedia*, Augsburg, Germany, 2007, pp. 862–871.
[136] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1097–1105, Dec. 2005.
[137] Y.-P. Tan and H. Lu, "Model-based clustering and analysis of video scenes," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2002, vol. 1, pp. 617–620.
[138] W. Tavanapong and J. Zhou, "Shot clustering techniques for story browsing," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 517–527, Aug. 2004.
[139] L.-H. Chen, Y.-C. Lai, and H.-Y. M. Liao, "Movie scene segmentation using background information," *Pattern Recognit.*, vol. 41, no. 3, pp. 1056–1065, Mar. 2008.
[140] A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automated high-level movie segmentation for advanced video-retrieval systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 4, pp. 580–588, Jun. 1999.
[141] M. Wang, X.-S. Hua, Y. Song, X. Yuan, S. Li, and H.-J. Zhang, "Automatic video annotation by semi-supervised learning with kernel density estimation," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, CA, 2006, pp. 967–976.
[142] Y. Rui, T. S. Huang, and S. Mehrotra, "Constructing table-of-content for video," *Multimedia Syst.*, vol. 7, no. 5, pp. 359–368, 1999.
[143] X. Yuan, X.-S. Hua, M. Wang, and X. Wu, "Manifold-ranking based video concept detection on large database and feature pool," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, CA, 2006, pp. 623–626.
[144] C.-W. Ngo, T.-C. Pong, H.-J. Zhang, and R. T. Chin, "Motion-based video representation for scene change detection," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 127–142, 2002.
[145] B. T. Truong, S. Venkatesh, and C. Dorai, "Scene extraction in motion pictures," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 5–15, Jan. 2003.
[146] R. Yan and M. Naphade, "Semi-supervised cross feature learning for semantic concept detection in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jul. 2005, vol. 1, pp. 657–663.
[147] H. Sundaram and S.-F. Chang, "Video scene segmentation using video and audio features," in *Proc. IEEE Int. Conf. Multimedia Expo.*, New York, 2000, pp. 1145–1148.
[148] N. Goela, K. Wilson, F. Niu, A. Divakaran, and I. Otsuka, "An SVM framework for genre-independent scene change detection," in *Proc. IEEE Int. Conf. Multimedia Expo.*, vol. 3, New York, Jul. 2007, pp. 532–535.
[149] Z. W. Gu, T. Mei, X. S. Hua, X. Q. Wu, and S. P. Li, "EMS: Energy minimization based video scene segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2007, pp. 520–523.
[150] Y. Ariki, M. Kumano, and K. Tsukada, "Highlight scene extraction in real time from baseball live video," in *Proc. ACM Int. Workshop Multimedia Inform. Retrieval*, Berkeley, CA, Nov. 2003, pp. 209–214.
[151] L. Xie, P. Xu, S.-F. Chang, A. Dirakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden Markov models," *Pattern Recognit. Lett.*, vol. 25, no. 7, pp. 767–775, 2004.
[152] Y. Zhai, A. Yilmaz, and M. Shah, "Story segmentation in news using visual and text cues," in *Proc. Int. Conf. Image Video Retrieval*, Singapore, Jul. 2005, pp. 92–102.
[153] W. H.-M. Hsu and S.-F. Chang, "Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jun. 2004, vol. 2, pp. 1091–1094.
[154] J. Wu, X.-S. Hua, and H.-J. Zhang, "An online-optimized incremental learning framework for video semantic classification," in *Proc. ACM Int. Conf. Multimedia*, New York, Oct. 2004, pp. 320–323.
[155] J. Yuan, J. Li, and B. Zhang, "Learning concepts from large scale imbalanced data sets using support cluster machines," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, CA, 2006, pp. 441–450.
[156] I. Otsuka, K. Nakane, A. Divakaran, K. Hatanaka, and M. Ogawa, "A highlight scene detection and video summarization system using audio feature for a personal video recorder," *IEEE Trans. Consum. Electron*, vol. 51, no. 1, pp. 112–116, Feb. 2005.
[157] K. Wan, X. Yan, and C. Xu, "Automatic mobile sports highlights," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2005, pp. 638–641.
[158] R.-G. Xiao, Y.-Y. Wang, H. Pan, and F. Wu, "Automatic video summarization by spatio-temporal analysis and non-trivial repeating pattern detection," in *Proc. Congr. Image Signal Process.*, May 2008, vol. 4, pp. 555–559.
[159] Y. H. Gong, "Summarizing audio–visual contents of a video program," *EURASIP J. Appl. Signal Process., Special Issue Unstructured Inform. Manag. Multimedia Data Sources*, vol. 2003, no. 2, pp. 160–169, Feb. 2003.
[160] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 296–305, Feb. 2005.
[161] M. Guironnet, D. Pellerin, N. Guyader, and P. Ladret, "Video summarization based on camera motion and a subjective evaluation method," *EURASIP J. Image Video Process.*, vol. 2007, pp. 1–12, 2007.
[162] J. Y. You, G. Z. Liu, L. Sun, and H. L Li, "A multiple visual models based perceptive analysis framework for multilevel video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 273–285, Mar. 2007.
[163] S. V. Porter, "Video segmentation and indexing using motion estimation," Ph.D. dissertation, Dept. Comput. Sci., Univ. Bristol, Bristol, U.K., 2004.
[164] Z. Li, G. Schuster, and A. Katsaggelos, "Minmax optimal video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 10, pp. 1245–1256, Oct. 2005.
[165] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Process.*, vol. 12, no. 7, pp. 796–807, Jul. 2003.
[166] P. M. Fonseca and F. Pereira, "Automatic video summarization based on MPEG-7 descriptions," *Signal Process.: Image Commun.*, vol. 19, no. 8, pp. 685–699, Sep. 2004.
[167] R. Ewerth and B. Freisleben, "Semi-supervised learning for semantic video retrieval," in *Proc. ACM Int. Conf. Image Video Retrieval*, Amsterdam, The Netherlands, Jul.2007, pp. 154–161.
[168] W.-N. Lie and K.-C. Hsu, "Video summarization based on semantic feature analysis and user preference," in *Proc. IEEE Int. Conf. Sens. Netw., Ubiquitous Trustworthy Comput.*, Jun. 2008, pp. 486–491.
[169] X.-N. Xie and F. Wu, "Automatic video summarization by affinity propagation clustering and semantic content mining," in *Proc. Int. Symp. Electron. Commerce Security*, Aug. 2008, pp. 203–208.
[170] D. Besiris, F. Fotopoulou, N. Laskaris, and G. Economou, "Key frame extraction in video sequences: A vantage points approach," in *Proc. IEEE Workshop Multimedia Signal Process.*, Athens, Greece, Oct. 2007, pp. 434–437.
[171] G. Ciocca and R. Schettini, "Supervised and unsupervised classification post-processing for visual video summaries," *IEEE Trans. Consum. Electron.*, vol. 52, no. 2, pp. 630–638, May 2006.
[172] Z. Li, G. M. Schuster, A. K. Katsaggelos, and B. Gandhi, "Rate-distortion optimal video summary generation," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1550–1560, Oct. 2005.
[173] I. Otsuka, K. Nakane, and A. Divakaran, "A highlight scene detection and video summarization system using audio feature for a personal video recorder," *IEEE Trans. Consum. Electron.*, vol. 51, no. 1, pp. 112–116, 2005.
[174] Y. Gao, W.-B. Wang, and J.-H. Yong, "A video summarization tool using two-level redundancy detection for personal video recorders," *IEEE Trans. Consum. Electron.*, vol. 54, no. 2, pp. 521–526, May 2008.
[175] J. Calic, D. Gibson, and N. Campbell, "Efficient layout of comic-like video summaries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 7, pp. 931–936, Jul. 2007.
[176] F. Wang and C.-W. Ngo, "Rushes video summarization by object and event understanding," in *Proc. Int. Workshop TREC Video Retrieval Eval. Video Summarization*, Augsburg, Bavaria, Germany, 2007, pp. 25–29.

[177] V. Valdes and J. M. Martinez, "On-line video summarization based on signature-based junk and redundancy filtering," in *Proc. Int. Workshop Image Anal. Multimedia Interactive Services*, 2008, pp. 88–91.

[178] J. Kleban, A. Sarkar, E. Moxley, S. Mangiat, S. Joshi, T. Kuo, and B. S. Manjunath, "Feature fusion and redundancy pruning for rush video summarization," in *Proc. Int. Workshop TREC Video Retrieval Eval. Video Summarization*, Augsburg, Bavaria, Germany, 2007, pp. 84–88.

[179] P. Over, A. F. Smeaton, and P. Kelly, "The TRECVID BBC rushes summarization evaluation pilot," in *Proc. Int. Workshop TREC-VID Video Summarization*, Augsburg, Bavaria, Germany, Sep. 2007.

[180] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 82–90, Jan. 2006.

[181] Y. Li, S.-H. Lee, C.-H. Yeh, and C.-C. J. Kuo, "Techniques for movie content analysis and skimming: Tutorial and overview on video abstraction techniques," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 79–89, Mar. 2006.

[182] P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-based video summarization using Delaunay clustering," *Int. J. Digital Libraries*, vol. 6, no. 2, pp. 219–232, Apr. 2006.

[183] Z. Xiong, X. S. Zhou, Q. Tian, Y. Rui, and T. S. Huang, "Semantic retrieval of videoReview of research on video retrieval in meetings, movies and broadcast news, and sports," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 18–27, Mar. 2006.

[184] C. M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E. Delp, "Automated video program summarization using speech transcripts," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 775–790, Aug. 2006.

[185] C. Taskiran, J.-Y. Chen, A. Albiol, L. Torres, C. A. Bouman, and E. J. Delp, "Vibe: A compressed video database structured for active browsing and search," *IEEE Trans. Multimedia*, vol. 6, no. 1, pp. 103–118, Feb. 2004.

[186] A. Aner, L. Tang, and J. R. Kender, "A method and browser for cross referenced video summaries," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Lausanne, Switzerland, Aug. 2002, vol. 2, pp. 237–240.

[187] Y. L. Geng, D. Xu, and S. H Feng, "Hierarchical video summarization based on video structure and highlight," in *Lecture Notes in Computer Science*, vol. 4109. Berlin, Germany: Springer, 2006, pp. 226–234.

[188] K. A. Peker, I. Otsuka, and A. Divakaran, "Broadcast video program summarization using face tracks," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2006, pp. 1053–1056.

[189] C. Gianluigi and S. Raimondo, "An innovative algorithm for key frame extraction in video summarization," *J. Real-Time Image Process.*, vol. 1, no. 1, pp. 69–88, Oct. 2006.

[190] C. Choudary and T. C. Liu, "Summarization of visual content in instructional videos," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1443–1455, Nov. 2007.

[191] M. Cooper, T. Liu, and E. Rieffel, "Video segmentation via temporal pattern classification," *IEEE Trans. Multimedia*, vol. 9, no. 3, pp. 610–618, Apr. 2007.

[192] H.-W. Kang and X.-S. Hua, "To learn representativeness of video frames," in *Proc. ACM Int. Conf. Multimedia*, Singapore, 2005, pp. 423–426.

[193] D. P. Mukherjee, S. K. Das, and S. Saha, "Key frame estimation in video using randomness measure of feature point pattern," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 5, pp. 612–620, May 2007.

[194] L. J. Liu and G. L. Fan, "Combined key-frame extraction and object-based video segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 7, pp. 869–884, Jul. 2005.

[195] X. M. Song and G. L. Fan, "Joint key-frame extraction and object segmentation for content-based video analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 904–914, Jul. 2006.

[196] J. Calic and B. Thomas, "Spatial analysis in key-frame extraction using video segmentation," in *Proc. Workshop Image Anal. Multimedia Interactive Services*, Lisbon, Portugal, Apr. 2004.

[197] C. Kim and J. Hwang, "Object-based video abstraction using cluster analysis," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2001, vol. 2, pp. 657–660.

[198] R. Yan and A. G. Hauptmann, "Probabilistic latent query analysis for combining multiple retrieval sources," in *Proc. Ann. Int. ACM SIGIR Conf. Inform. Retrieval*, Seattle, WA, 2006, pp. 324–331.

[199] A. Girgensohn and J. Boreczky, "Time-constrained keyframe selection technique," *Multimedia Tools Appl.*, vol. 11, no. 3, pp. 347–358, 2000.

[200] X. D. Yu, L. Wang, Q. Tian, and P. Xue, "Multilevel video representation with application to keyframe extraction," in *Proc. Int. Multimedia Modelling Conf.*, 2004, pp. 117–123.

[201] D. Gibson, N. Campbell, and B. Thomas, "Visual abstraction of wildlife footage using Gaussian mixture models and the minimum description length criterion," in *Proc. IEEE Int. Conf. Pattern Recog.*, Dec. 2002, vol. 2, pp. 814–817.

[202] T. Wang, Y. Wu, and L. Chen, "An approach to video key-frame extraction based on rough set," in *Proc. Int. Conf. Multimedia Ubiquitous Eng.*, 2007.

[203] T. M Liu, H.-J. Zhang, and F. H. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 10, pp. 1006–1013, Oct. 2003.

[204] A. M. Ferman and A. M. Tekalp, "Two-stage hierarchical video summary extraction to match low-level user browsing preferences," *IEEE Trans. Multimedia*, vol. 5, no. 2, pp. 244–256, Jun. 2003.

[205] Z. H. Sun, K. B. Jia, and H. X. Chen, "Video key frame extraction based on spatial-temporal color distribution," in *Proc. Int. Conf. Intell. Inform. Hiding Multimedia Signal Process.*, 2008, p. 196-199.

[206] R. Narasimha, A. Savakis, R. M. Rao, and R. De Queiroz, "Key frame extraction using MPEG-7 motion descriptors," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, vol. 2, pp. 1575–1579.

[207] D. Xia, X. Deng, and Q. Zeng, "Shot boundary detection based on difference sequences of mutual information," in *Proc. Int. Conf. Image Graph.*, Aug. 2007, pp. 389–394.

[208] B. Fauvet, P. Bouthemy, P. Gros, and F. Spindler, "A geometrical key-frame selection method exploiting dominant motion estimation in video," in *Proc. Int. Conf. Image Video Retrieval*, Jul. 2004, pp. 419–427.

[209] H. J. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognit.*, vol. 30, no. 4, pp. 643–658, 1997.

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

[211] A. Divakaran, R. Radhakrishnan, and K. A Peker, "Motion activity-based extraction of key-frames from video shots," in *Proc. IEEE Int. Conf. Image Process.*, 2002, vol. 1, Rochester, NY, pp. 932–935.

[212] J. Rong, W. Jin, and L. Wu, "Key frame extraction using inter-shot information," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jun. 2004, pp. 571–574.

[213] M. Cooper and J. Foote, "Discriminative techniques for keyframe selection," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2005, pp. 502–505.

[214] H. S. Chang, S. Sull, and S. U. Lee, "Efficient video indexing scheme for content-based retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1269–1279, Dec. 1999.

[215] S. V. Porter, M. Mirmehdi, and B. T. Thomas, "A shortest path representation for video summarization," in *Proc. Int. Conf. Image Anal. Process.*, Sep. 2003, pp. 460–465.

[216] H.-C. Lee and S.-D. Kim, "Iterative key frame selection in the rate-constraint environment," *Signal Process. Image Commun.*, vol. 18, no. 1, pp. 1–15, 2003.

[217] T. Liu, X. Zhang, J. Feng, and K. Lo, "Shot reconstruction degree: A novel criterion for key frame selection," *Pattern Recognit. Lett.*, vol. 25, no. 12, pp. 1451–1457, Sep. 2004.

[218] J. Calic and E. Izquierdo, "Efficient key-frame extraction and video analysis," in *Proc. Int. Conf. Inf. Technol.: Coding Comput.*, Apr. 2002, pp. 28–33.

[219] J. Yuan, L. Xiao, D. Wang, D. Ding, Y. Zuo, Z. Tong, X. Liu, S. Xu, W. Zheng, X. Li, Z. Si, J. Li, F. Lin, and B. Zhang, "Tsinghua University at TRECVID 2005," in *Proc. TREC Video Retrieval Eval.*, Gaithersburg, MD, 2005. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/tsinghua.pdf

[220] S.-H. Han and I.-S. Kweon, "Scalable temporal interest points for abstraction and classification of video events," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2005, pp. 1–4.

[221] S. Y. Neo, J. Zhao, M. Y. Kan, and T. S. Chua, "Video retrieval using high level features: Exploiting query matching and confidence-based weighting," in *Proc. Conf. Image Video Retrieval*, Singapore, 2006, pp. 370–379.

[222] A. Amir, W. Hsu, G. Iyengar, C. Y. Lin, M. Naphade, A. Natsev, C. Neti, H. J. Nock, J. R. Smith, B. L. Tseng, Y. Wu, and D. Zhang, "IBM research TRECVID-2003 video retrieval system," in *Proc. TREC Video Retrieval Eval.*, Gaithersburg, MD, 2003. Available: http://www-nlpir.nist.gov/projects/tvpubs/tvpapers03/ibm.smith.paper.final2.pdf

[223] 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, "Informedia at TRECVID 2003: Analyzing and searching broadcast news

video," in *Proc. TREC Video Retrieval Eval.*, Gaithersburg, MD, 2003. Available: http://www-nlpir.nist.gov/projects/tvpubs/tvpapers03/cmu.final.paper.pdf

[224] C. Foley, C. Gurrin, G. Jones, H. Lee, S. McGivney, N. E. O'Connor, S. Sav, A. F. Smeaton, and P. Wilkins, "TRECVID 2005 experiments at Dublin city university," in *Proc. TREC Video Retrieval Eval.*, Gaithersburg, MD, 2005. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/dcu.pdf

[225] E. Cooke, P. Ferguson, G. Gaughan, C. Gurrin, G. Jones, H. L. Borgue, H. Lee, S. Marlow, K. McDonald, M. McHugh, N. Murphy, N. O'Connor, N. O'Hare, S. Rothwell, A. Smeaton, and P. Wilkins, "TRECVID 2004 experiments in Dublin city university," in *Proc. TREC Video Retrieval Eval.*, Gaithersburg, MD, 2004. Available: http://www-nlpir.nist.gov/projects/tvpubs/tvpapers04/dcu.pdf

[226] J. Adcock, A. Girgensohn, M. Cooper, T. Liu, L. Wilcox, and E. Rieffel, "FXPAL experiments for TRECVID 2004," in *Proc. TREC Video Retrieval Eval., Gaithersburg*, MD, 2004. Available: http://www-nlpir.nist.gov/projects/tvpubs/tvpapers04/fxpal.pdf

[227] T. Volkmer and A. Natsev, "Exploring automatic query refinement for text-based video retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Toronto, 2006, pp. 765–768.

[228] A. Hauptmann, M. Y. Chen, M. Christel, C. Huang, W. H. Lin, T. Ng, N. Papernick, A. Velivelli, J. Yang, R. Yan, H. Yang, and H. D. Wactlar, "Confounded expectations: Informedia at TRECVID 2004," in *Proc. TREC Video Retrieval Eval.*, Gaithersburg, MD, 2004. Available: http://www-nlpir.nist.gov/projects/tvpubs/tvpapers04/cmu.pdf

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

[230] L. X. Xie, H. Sundaram, and M. Campbell, "Event mining in multimedia streams," *Proc. IEEE*, vol. 96, no. 4, pp. 623–646, Apr. 2008.

[231] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, and M.-S. Chen, "Association and temporal rule mining for post-filtering of semantic concept detection in video," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 240–251, Feb. 2008.

[232] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 252–259, Feb. 2008.

[233] R. Fablet, P. Bouthemy, and P. Perez, "Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 393–407, Apr. 2002.

[234] Y.-F. Ma and H.-J. Zhang, "Motion texture: A new motion based video representation," in *Proc. Int. Conf. Pattern Recog.*, Aug. 2002, vol. 2, pp. 548–551.

[235] A. D. Bimbo, E. Vicario, and D. Zingoni, "Symbolic description and visual querying of image sequences using spatiotemporal logic," *IEEE Trans. Knowl. Data Eng.*, vol. 7, pp. 609–622, Aug. 1995.

[236] S. F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "A fully automated content-based video search engine supporting spatiotemporal queries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 602–615, Sep. 1998.

[237] F. I. Bashir, A. A. Khokhar, and D. Schonfeld, "Real-time motion trajectory-based indexing and retrieval of video sequences," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 58–65, Jan. 2007.

[238] W. Chen and S. F. Chang, "Motion trajectory matching of video objects," in *Proc. SPIE vol. 3972: Storage and Retrieval for Media Databases*, Jan. 2000, pp. 544–553.

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

[240] X. Yuan, W. Lai, T. Mei, X.-S. Hua, and X.-Q. Wu, "Automatic video genre categorization using hierarchical SVM," in *Proc. IEEE Int. Conf. Image Process.*, Atlanta, GA, Oct. 2006, pp. 2905–2908.

[241] C. G. M. Snoek, M. Worring, J. C. van Gemert, J. M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, CA, 2006, p. 421-430.

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

[243] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 958–966, Aug. 2007.

[244] G.-J. Qi, X.-S. Hua, Y. Rui, J. H. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *Proc. ACM Int. Conf. Multimedia*, Augsburg, Germany, 2007, pp. 17–26.

[245] D. Brezeale and D. J. Cook, "Automatic video classification: A survey of the literature," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 38, no. 3, pp. 416–430, May 2008.

[246] P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and system for segmentation and structure analysis in soccer video," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Tokyo, Japan, 2001, pp. 928–931.

[247] Y. P. Tan, D. D. Saur, S. R. Kulkarni, and P. J. Ramadge, "Rapid estimation of camera motion from compressed video with applications to video annotation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 1, pp. 133–146, Feb. 2000.

[248] Y. Wu, B. L. Tseng, and J. R. Smith, "Ontology-based multi-classification learning for video concept detection," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jun. 2004, vol. 2, pp. 1003–1006.

[249] J. R. Smith and M. Naphade, "Multimedia semantic indexing using model vectors," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2003, vol. 2, pp. 445–448.

[250] W. Jiang, S.-F. Chang, and A. Loui, "Active concept-based concept fusion with partial user labels," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2006, pp. 2917–2920.

[251] M. Bertini, A. Del Bimbo, and C. Torniai, "Automatic video annotation using ontologies extended with visual information," in *Proc. ACM Int. Conf. Multimedia*, Singapore, Nov. 2005, pp. 395–398.

[252] A. G. Hauptmann, M. Christel, and R. Yan, "Video retrieval based on semantic concepts," *Proc. IEEE*, vol. 96, no. 4, pp. 602–622, Apr. 2008.

[253] J. Fan, H. Luo, and A. K. Elmagarmid, "Concept-oriented indexing of video databases: Towards semantic sensitive retrieval and browsing," *IEEE Trans. Image Process.*, vol. 13, no. 7, pp. 974–992, Jul. 2004.

[254] F. Pereira, A. Vetro, and T. Sikora, "Multimedia retrieval and delivery: Essential metadata challenges and standards," *Proc. IEEE*, vol. 96, no. 4, pp. 721–744, Apr. 2008.

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

[256] C. G. M. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools Appl.*, vol. 25, no. 1, pp. 5–35, Jan. 2005.

[257] S.-F. Chang, W.-Y. Ma, and A. Smeulders, "Recent advances and challenges of semantic image/video search," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2007, vol. 4, pp. IV-1205–IV-1208.

[258] R. Yan, J. Yang, and A. G. Hauptmann, "Learning query-class dependent weights in automatic video retrieval," in *Proc. ACM Int. Conf. Multimedia*, New York, Oct. 2004, pp. 548–555.

[259] L. Kennedy, P. Natsev, and S.-F. Chang, "Automatic discovery of query class dependent models for multimodal search," in *Proc. ACM Int. Conf. Multimedia*, Singapore, Nov. 2005, pp. 882–891.

[260] P. Over, G. Awad, J. Fiscus, and A. F. Smeaton. (2010). "TRECVID 2009–Goals, tasks, data, evaluation mechanisms and metrics," [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

[261] K. Schoeffmann, F. Hopfgartner, O. Marques, L. Boeszoermenyi, and J. M. Jose, "Video browsing interfaces and applications: A review," *SPIE Rev.*, vol. 1, no. 1, pp. 018004.1–018004.35, May 2010.

[262] C. G. M. Snoek and M. Worring, "Concept-based video retrieval," *Foundations Trends Inform. Retrieval*, vol. 2, no. 4, pp. 215–322, 2009.

[263] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of TRECVid activity," *Comput. Vis. Image Understanding*, vol. 114, no. 4, pp. 411–418, 2010.

[264] G. Quenot, D. Moraru, and L. Besacier. (2003). "CLIPS at TRECVID: Shot boundary detection and feature detection," in *Proc. TREC Video Retrieval Eval. Workshop Notebook Papers* [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2003.

[265] P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton. (2005). "TRECVID 2005—An overview," in *Proc. TREC Video Retrieval Eval. Workshop.* [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2005.

[266] A. F. Smeaton, P. Over, and W. Kraaij, "High-level feature detection from video in TRECVID: A 5-year retrospective of achievements," *Multimedia Content Analysis: Theory and Applications* (Springer Series on

Signals and Communication Technology) Berlin, Germany: Springer, 2009, pp. 151–174.
[267] J. Sivic and A. Zisserman, "Video Google: Efficient visual search of videos," in *Toward Category-Level Object Recognition.*. Berlin, Germany: Springer, 2006, pp. 127–144.
[268] M. Chen, M. Christel, A. Hauptmann, and H. Wactlar, "Putting active learning into multimedia applications: Dynamic definition and refinement of concept classifiers," in *Proc. ACM Int. Conf. Multimedia*, 2005, pp. 902–911.
[269] H. B. Luan, S. Y. Neo, H. K. Goh, Y. D. Zhang, S. X. Lin, and T. S. Chua, "Segregated feedback with performance-based adaptive sampling for interactive news video retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 293–296.
[270] G. P. Nguyen, M. Worring, and A. W. M. Smeulders, "Interactive search by direct manipulation of dissimilarity space," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1404–1415, Jun. 2007.
[271] E. Bruno, N. Moenne-Loccoz, and S. Marchand-Maillet, "Design of multimodal dissimilarity spaces for retrieval of video documents," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1520–1533, Sep. 2008.
[272] P. Over, A. F. Smeaton, and G. Awad, "The TRECVid 2008 BBC rushes summarization evaluation," in *Proc. 2nd ACM TREC Video Retrieval Eval. Video Summarization Workshop*, 2008, pp. 1–20.
[273] W. Bailer and G. Thallinger, "Comparison of content selection methods for skimming rushes video," in *Proc. 2nd ACM TREC Video Retrieval Eval. Video Summarization Workshop*, 2008, pp. 85–89.
[274] Z. Liu, E. Zavesky, B. Shahraray, D. Gibbon, and A. Basso, "Brief and high-interest video summary generation: Evaluating the AT&T labs rushes summarizations," in *Proc. 2nd ACM TREC Video Retrieval Eval. Video Summarization Workshop*, 2008, pp. 21–25.
[275] V. Chasanis, A. Likas, and N. Galatsanos, "Video rushes summarization using spectral clustering and sequence alignment," in *Proc. 2nd ACM TREC Video Retrieval Eval. Video Summarization Workshop*, 2008, pp. 75–79.
[276] M. G. Christel, A. G. Hauptmann, W.-H. Lin, M.-Y. Chen, B. Maher, and R. V. Baron, "Exploring the utility of fast-forward surrogates for BBC rushes," in *Proc. 2nd ACM TREC Video Retrieval Eval. Video Summarization Workshop*, 2008, pp. 35–39.
[277] S. Naci, U. Damnjanovic, B. Mansencal, J. Benois-Pineau, C. Kaes, and M. Corvaglia, "The COST292 experimental framework for RUSHES task in TRECVID 2008," in *Proc. 2nd ACM TREC Video Retrieval Eval. Video Summarization Workshop*, 2008, pp. 40–44.
[278] W. Ren, S. Singh, M. Singh, and Y. S. Zhu, "State-of-the-art on spatio-temporal information-based video retrieval," *Pattern Recognit.*, vol. 42, no. 2, pp. 267–282, Feb. 2009.
[279] M. Wang, X. S. Hua, J. Tang, and R. Hong, "Beyond distance measurement: Constructing neighborhood similarity for video annotation," *IEEE Trans. Multimedia*, vol. 11, no. 3, pp. 465–476, Apr. 2009.

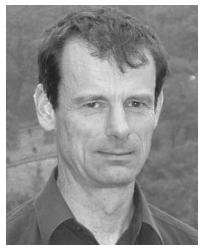
Weiming Hu (SM'09) received the Ph.D. degree from the Department of Computer Science and Engineering, Zhejiang University, Zhejiang, China.

From April 1998 to March 2000, he was a Postdoctoral Research Fellow with the Institute of Computer Science and Technology, Founder Research and Design Center, Peking University, Beijing, China. Since April 1998, he has been with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, where he is currently a Professor and a Ph.D. Student Supervisor with the laboratory. He has published more than 100 papers in national and international journals and international conferences. His research interests include visual surveillance, filtering of Internet objectionable information, retrieval of multimedia, and understanding of Internet behaviors.

Nianhua Xie, photograph and biography not available at the time of publication.

Li Li, photograph and biography not available at the time of publication.

Xianglin Zeng, photograph and biography not available at the time of publication.



Stephen Maybank received the B.A. degree in mathematics from King's College, Cambridge, U.K., in 1976 and the Ph.D. degree in computer science from Birkbeck College, University of London, London, U.K., in 1988.

He joined the Pattern Recognition Group at Marconi Command and Control Systems, Frimley, U.K., in 1980 and moved to the GEC Hirst Research Centre, Wembley, U.K., in 1989. During 1993–1995, he was a Royal Society/Engineering and Physical Sciences Research Council Industrial Fellow with the Department of Engineering Science, University of Oxford, Oxford, U.K. In 1995, he joined the University of Reading, Reading, U.K., as a Lecturer within the Department of Computer Science. In 2004, he became a Professor with the School of Computer Science and Information Systems, Birkbeck College. His research interests include the geometry of multiple images, camera calibration, visual surveillance, information geometry, and the applications of statistics to computer vision.