International Journal of Computer and Information Technology (ISSN: 2279 – 0764)
Volume 03 – Issue 01, January 2014

Video Indexing: A Survey

Muhammad Nabeel Asghar, Fiaz Hussain, Rob Manton
Department of Computer Science and Technology
University of Bedfordshire
Park Square Luton LU1, 3JU, United Kingdom
Email: nabeel209 {at} @gmail.com

Abstract—**there is a wide range of applications in video retrieval and indexing grabbing researcher's attention. The following paper includes tutorial as well as a description of the background to common approaches in visual content-based video retrieval and indexing. The focal point is on the mechanisms of video shot boundary detection, key frame extraction, structure analysis, and scene segmentation, along with extraction of key frame features, static features, motion features, object features, video data mining, video annotation, video retrieval, including similarity measure, query interfaces, relevance feedback, and video browsing. Ultimately, directions for future work will be proposed.**

I. Introduction

The basic multimedia information is required for dynamic video indexing and retrieval. There are necessary components for storing, sorting and accessing multimedia contents. They also help in finding the desired components to form a multimedia repository with an ease [1]. Besides many multimedia resources' video is a key component which comprises mainly upon three major parts. The first one is that the vigorous video provides rich contents then that of images. The second is an enormous quantity of raw data. Lastly, the structure of video is very minute. These features have made the retrieval and indexing of video relatively complex. Previously, database video was diminutive, along with that the retrieval and indexing were also based on manual keyword annotation. In recent times, the database is getting enormous, the content-based retrieval and indexing are needed with less human interaction to analyse videos automatically.

There are a lot of applications for video indexing and content-based recovery. For instance, visual electronic commerce scrutiny (analysis of interest trends for user elections and orderings), examination of correlations in digital institutions, instant browsing of video folders, news event analysis [2], intellectual administration of web videos (useful video search and harmful video tracing), and video inspection that has motivated the interests of researchers. Two main examples of research are predominately significant: 1) The National Institute of Standards and Technology (the supporter of the annual Text Retrieval Conference) TREC Video Retrieval Evaluation (TRECVid) to endorse improvement in video retrieval and analysis since 2001. Substantial amount of experimented video is provided by TRECVid. There are many contributors who submitted their algorithms on content based video retrieval to the collected work [3] [4], and [5]; 2) for standard video, the main objective is to certify suitability between the depiction of interfaces for the contents of video to improve and help in the progress of fast and perfect retrieval of video algorithms. The TV-Anytime Standard and moving picture experts group (MPEG) are the key values for videos [6]. There are many investigations using MPEG-7 for the classification of video contents to extort characteristics or for the explanations of the video objects, condensed domain [7]. There can be an acoustic channel along with that optical channel in a video. From videos the obtainable information includes subsequently [8] [9]: 1) video metadata is the marked texts entrenched in date, actors, videos, title, producer, summary, as well as the duration of broadcast, video format, file size, copyright and so on; 2) the auditory channel provides audio information; 3) by speech recognition speech transcription can be obtained and by character recognition techniques, caption texts can be examined. 4) As it is known that in the images, the visual information is contained from the visual channels. The web page texts are linked with video when it is uploaded on web page. In the paper, the focus will be on optical contents of videos and surveys and will be conducted on the visual content-based video retrieval and indexing.

Many surveys have been conducted due to the popularity and significance of video retrieval and indexing. Broadly, every paper focuses on a part of video retrieval and indexing. For instance, video shot boundary detection a good review given by Smeaton *et al*. [4] in the seven years of TRECVid activity. A complete review given by Snoek and Worring [10] on concept based video retrieval. Their main focus was on evaluation of algorithms using the databases of TRECVid, concept detection, and video search by means of semantic concepts. High quality review stated by Schoeffmann *et al.* [11] on application and interfaces of video browsing systems. On the art of the state spatio-temporal semantic information based video retrieval a review has been done by Ren *et al.* [12].

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

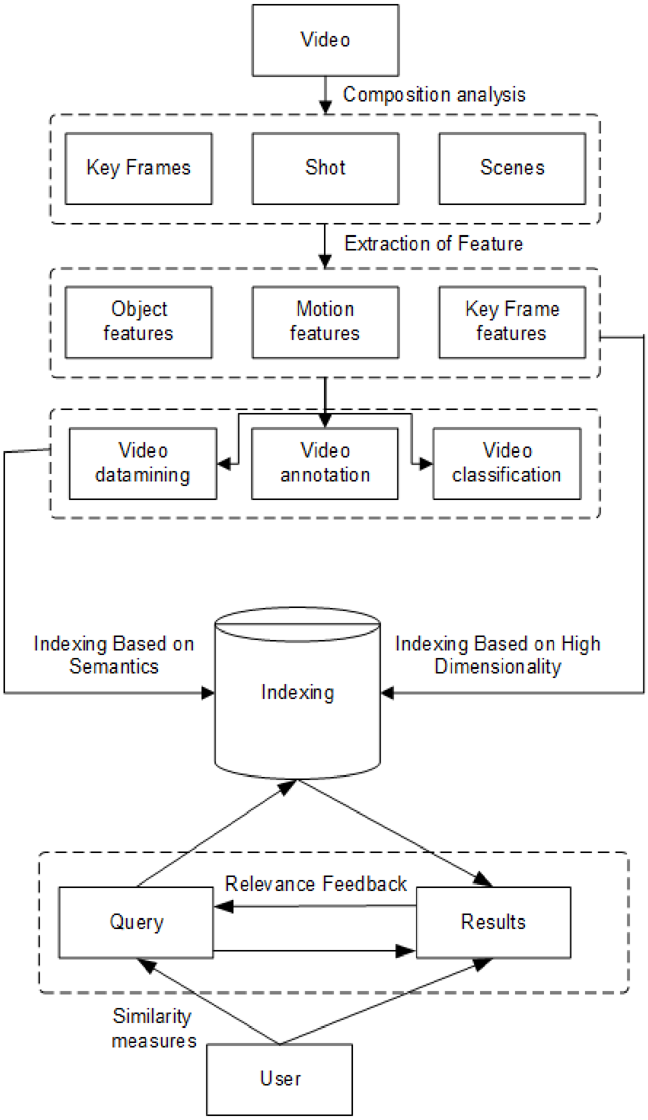

Figure. 1. Video indexing and retrieval framework

In the paper, the recent developments are reviewed as well as the upcoming open directions in ocular content-based video retrieval and indexing is analysed. Following is the core features of this survey: 1. in a lucid and clear way, the video retrieval and indexing components are discussed and the links between these components are shown in hierarchy. 2. For the inspection of the state of the art, every task consists of ocular content-based video indexing; different categories in the sub process are discussed and retrieval is fragmented into sub processes. Different approaches' benefits and limitations are summarized. The main focus is on reviewing current papers as an addition to the prior surveys. Thorough reviews have been given for those tasks which are not been surveyed up until now. 3. A detail discussion is done in prospect to the future of visual based content video indexing and retrieval.

The after mentioned points evidently discern this survey as compare to the previous and existing surveys on video indexing and retrieval. As this survey is the broadest according to our survey. The order of the remaining part of the paper is as following: Section II a brief review on the work connected to video structure analysis. Section III is dealing with feature extractions. Discussion on classification, video data mining, and annotation is related to Section IV. Section V illustrates the video query and retrieval. Section VI is based on video summary on browsing. Section VII focuses on potential directions for further research. Section VIII is the conclusion of this paper.

II. ANALYSIS OF VIDEO COMPOSITION

Mostly, the hierarchy of video clips, scenes, shots and frames are arranged in a descending manner as shown in figure 2.

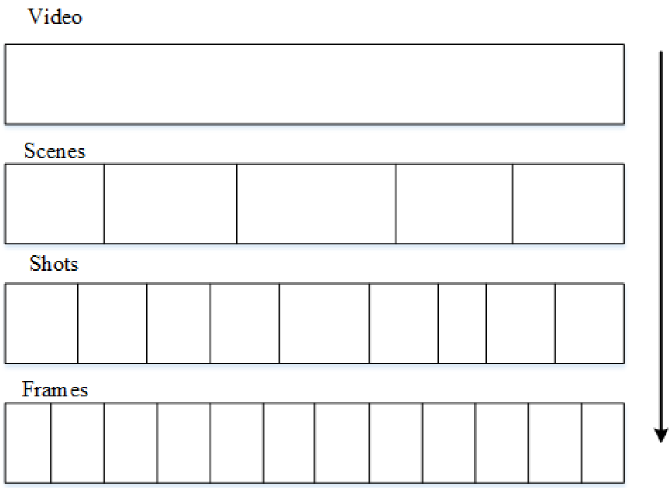

Figure. 2. General Hierarchy of Video Parsing

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

A. Boundary Detection of the Shot

In video, consecutive series of frames known as shot, which are taken by the camera. An action taken in between the start and stop process marks the boundaries of the shot [13]. In the frames of a shot, there are several well-built content relations. For advanced level of semantic annotation and retrieval tasks, the shots are regarded as the basic unit to organize the sequenced content of video and primitives. Usually, the boundary shots are categorized as cut in it the successive shot transition is rapid, and the continuing transitions consist of dissolve, wipe, fade in and out; it extends over a number of frames. The continuing transition detection is quite difficult as compare to cut detection. There is a brief history of the research and surveys of the shot boundary detection and on video shot boundary detection [14] [4]. The basic groups of approaches are introduced comprehensively for detection of shot boundary along with their reviews, qualities and disadvantages [14] [4]. Mostly, in start, the visual features are extracted from each frame for shot boundary detection, after that the similarities are measured between the frame by the use of extracted features, and at the end those shot boundaries are detected, which are not similar in the frames. Furthermore, three main steps have been discussed in shot boundary detection: resemblance measurement, extraction of the attributes, [15], and findings. For boundary detection of the shot features like motion vectors, colour histogram [16], block colour histogram and edge change ratio, [17] [18], along with the characteristics, for instance, scale invariant feature transform [19], saliency map [20], corner points [21], and so on. Small camera motions between the colour histograms are strong and sensitive as compared to large camera motion, although they cannot distinguish between the same scenes of the shots. In illumination changes and motion, the features of Edge are more invariable as compared to colour histograms, and the influence of object and camera motion can be handled efficiently by motion features. Commonly, the simple colour histograms are not outperformed by edge features, motion features and more complicated features [14]. For shot boundary detection, the extracted features are the second step needed to gauge a similarity between frames. Euclidean distance, 1-norm cosine dissimilarity, chi-squared similarity and the histogram intersection, are the contemporary similarity metrics for extracted feature vectors [22] [23] [24], along with that the earth movers distance [16], shared information are various novel similarity measures [25] [26] [27]. Pair-wise similarity which measures the resemblance among consecutive frames and similarities among frames within a window are measured by window similarity measurement [24]. To inculcate the contextual information on the reduction of the influencing local noises and disturbances the window based similarity measures are used. Although as compared to pair-wise similarity measurement, they need additional computation. Shot boundary can be identified by the usage of calculated similarities among frames. The modern advancement in shot boundary detection is that it can be classified into statistical-learning and threshold based.

1) Approach of Threshold: By comparing pair wise measured similarities among frames along the predefined threshold, the shot boundaries are detected by the threshold based approach [28] [29]. The boundary is identified when a similarity is lower than the threshold. The threshold can be of these kinds, global, adaptive or the combination of adaptive and global: 1) the similar threshold is used by the global threshold based algorithms as usually set empirically on the video [29]. Less effective inculcation of the local content in the estimation of global threshold is the major setback of the global threshold based algorithms, which are influencing the accuracy of boundary detection; 2) in a sliding window the adaptive threshold based algorithms calculate the threshold locally [20] [16] [27]. As an alternative to the global threshold when an adaptive threshold is used the performance of the detection frequently improves [30]. Although, estimation of the global threshold is easier than the adaptive threshold estimation. In adaptive threshold, users should have knowledge about the description of videos to select its parameters, for instance, sliding window size; 3) the combination of adaptive and global algorithms fine-tunes the local thresholds by having the value of the account of global threshold. Cut transition detection, dissolve transition and flash detection are defined by Quenot *et al*. [31] for the two global threshold functions which are gathered from off trade precision and recall. The value of the functions' changes locally even though that algorithm requires tuning the two global thresholds. Algorithm's limitations are that the functional relations of locally adaptive thresholds and two global thresholds are not able to be easily determined.

2) Statistical Learning-Based Approach: The detection of shot boundary is regarded as the classification task by the statistical learning-based approach. In this task, the frames are classified upon the basis of their features, depending on shot or no shot variation. In it both administered (supervised) or non-administered (non-supervised) learning is used. a) Supervised learning based classifiers: Adaboost and Support Vector Machine (SVM) are the most commonly used and are controlled or supervised classifiers for detection of shot boundary.

1) SVM [22] [32]: SVM is used as two-class categoriser to separate cuts from non-cuts by Chavez *et al*. [33]. To plot the structures in the high dimensional area, a kernel function is used to conquer the change influence in object's fast movement and illumination. Zhao *et al*. [34] in a sliding window, two SVM classifier were exploited for the detection of gradual transition and cuts correspondingly. The Ling *et al*. [35] extracted primarily, many frame features after that used the SVM for the classification of the frames in three categories: gradual transition, cut and others etc. SVM-based classifier along with the threshold based method is combined by Yuan *et al*. [14] and Liu *et al*. [26]. The candidate's boundaries are chosen primarily via threshold based method. After that to verify the boundaries, SVM classifier is used extensively [36] for the shot boundary detection. The SVM-based algorithms are used frequently as well. The merits are as following:

a) Training information and maintenance of the good generalization are fully utilized by them;
b) By using the kernel function the large numbers of features are easily handled; and
c) Already plenty of SVM codes are obtainable.

2) Adaboost: Cut detection made by Herout *et al*. [18] for the pattern detection task where the algorithm of Adaboost was used. In the compressed domain, Zhao and Cai [17] applied it for shot boundary detection. Primarily by adaptation of a fuzzy classifier the colour and motion pictures remain roughly classified. After that every frame is described as a cut, and Adaboost classifier is used for the slow, or no transformation of the frame. Adaboost boundary classifiers major plus point is that the features can be handled in a huge quantity: A part of features chosen by these classifiers for the classification of boundary.

3) Others: For shot boundary detection, some other supervised algorithms have been used. Cooper *et al*. [24] for example, uses binary (kNN) k nearest-neighbour classifier in which the frames similarities are used as input in the particular temporal interval. Hidden Markov (HMM) models are applied by Boreczky and Wilcox [37] for detection of dissolves, fades, model shot cuts, zooms with separate states and pans. Above mentioned profits about the supervised-learning approaches define that it is not needed to set thresholds in the threshold based approaches along with that for the improvement of detection accuracy different types of features can be amalgamated. The drawbacks of their profound reliance on an apt selection of training set with the examples of both negative and positive.

b) Unsupervised learning based algorithms: they have been categorized into the algorithms based on similarity of frame and simply frame based. The similarity based algorithms of the frame huddles the similarity measurements in two clusters among the frames' pair: The similarities of the cluster having lower values matches up to the superior values of the similarities and shot boundaries that match up to the non-boundaries [38]. K-means and fuzzy K-means are employed in clustering algorithms. Each shot is treated by the frame based algorithms as a cluster of frames which are alike in visual content. Clustering ensembles are used to assemblage different corresponding shots of frames by Chang *et al*. [19]. K-means clustering used by Lu *et al*. [23] as well as spectral clustering for the frames cluster for the detection of different shots used by Damnjanovic *et al*. [39]. The training assets are not needed in the value of clustering based approaches. The temporal sequence progression information is not stored. They are incompetent in identifying the different sorts of gradual change, which is its limitation. The uncompressed domain based and compressed domains based are the classifications of the shot boundary detection approach. Circumventing the time wasting decompression of the video, the compressed domain features are available, for instance cosine discrete coefficients transform. [40] [41] [17] are used. For the detection of shot boundary, motion vectors DC image and MB types can be employed. The compression standards is for the highly reliance of the compressed domain based approach. They are reduced in accuracy as compared to the domain based uncompressed approach. More attention has been received by the gradual detection, lately. Dissolves based on multi resolution analysis has been detected by Ngo [42]. In accordance to the gradual transitions, Yoo *et al*. [43] detected the variance distribution curve of edge information in frame sequences.

B. Key Frame Extraction

In the same shots there have been huge redundancies in frames that is why those frames which have the best reflection are chosen for the content of the shot as key frames [44] [45] [46] and [47] to concisely demonstrate the shot. As much as possible salient content of the shots must be contained by the extracted key frames and redundancy must be avoided. The characteristics of the key frame extraction are colour (usually the histogram colour), shapes, edges, optical flow, motion temporal intensity and spatial distribution of MPEG-7 [48] , camera activity, discrete cosine coefficient of MPEG, motion vectors [49], and derived features by the image variations because of the camera movement [50] [51]. The recent approach of key frame extraction is classified in six portions as classified by Truong and Venkatesh [45]: i) simplification of the curve based, ii) reference frame based, iii) sequential contrast-based, iv) global comparison-based, v) event/object based, vi) clustering based.

1) Simplification of the curve: Every single frame of shot is represented by point in the feature space using algorithms. To make the trajectory curve the points have been linked in the sequential order then the search is done to explore points set that remarkably matches with the curve shape. The frame difference metrics has been generated by the Calic and Izquierdo [52] by statistically analysing compressed stream of MPEG based on features

extracted from the macro block. By using different simplifications of the metrics curve of discrete contour evolution algorithms, the key frame method of extraction is implemented. The main constraint is the computational complication for the best illustration of the curve.

2) Frame Reference: A reference frame is generated by algorithms and after that key frames are selected by comparing the reference frame and the shot frames. For example, an alpha trimmed average histogram has been constructed to describe the colour distribution in the frames of the shot by Ferman and Tekalp [53]. Then the distance among the histogram of each frame in the shot and alpha trimmed average histogram are calculated. By means of the division of the distance curve, the key frames have been located. Sun *et al*. [54] constructed a maximum amount of frame occurrence for a shot. A weighted distance has been calculated between each frame in the constructed frame and the shot. At the peak of the distance curve, the key frames have been extracted. The advantages of the reference based algorithms are: a) they are understandable and; b) could easily be implemented. The primary limitations are that they are dependable upon reference frame, and some noticeable contents could be missing from the key frames of the shots, if the shot is not properly represented by the reference frame.

3) Comparison between frames sequentially: Previously extracted key frames in the algorithms of subsequent frames were consecutively contrasted with the frame until the totally different frame from the key frame is found. As a next key frame, the acquired key frame is designated. Zhang *et al*. [55] for example, the colour dissimilarity in histogram have been used to extract the key frames among current frame and previous key frame. Accumulated energy function has been computed by transversely displacing the image-block on two successive frames for the measurement of the distance between frames for the extraction of key frames by Zhang *et al*., [56]. The sequential comparison base algorithm's merits comprise of intuitiveness, fussiness, squat computational complexity and to entire shot the edition of the number of key frames. The algorithm's constrains are consisted of:

a) The key frames represent the local properties of the shot as compare to the global properties. b) The uncontrolled amount of key frames and irregular distribution formulate that the algorithms are not suitable for application that does not require levelled distribution or permanent number of key frames. c) When the contents are appeared repeatedly severance can occur.

4) Global Comparison among Frames: By the minimization of predefined objective functions are dependent on application, with the algorithms created on total alterations among frames in the distribution of key frames of the shot [45].

a) Even temporal variance: The shot algorithms select the key frames in such a way that the segment of each shot has equal temporal values represented by a key frame. Among the temporal variances of all segments, the objective functions are chosen as the sum of differences. In a segment, the temporal variance can be estimated from across consecutive contents of the frames [56] as well as the disparity between the first and the last frames. Divakaran *et al*. [57] obtained the key frames with the shot division into segments with the equivalent growing motion activity by means of the MPEG-7 motion descriptor, after that every segment is designated as a key frame of those frames located at the halfway point.

b) Maximum coverage: [45] extracted the key frames by the algorithms for maximizing the coverage representation; they are those numbers of frames, which are represented by the key frames. [58][59]. If the number of key frames are not fixed, the algorithms reduce the number of key frames subject to predefined fidelity criterion. On the other hand, the algorithms get the most out of the number of frames if the various key frames are fixed Chang *et al*. [60]. For examples, key frame coverage as those numbers of frames which are visibly same as the key frame. A greedy algorithm is used to alternatively explore the key frames.

c) Minimum correlation: The extracted key frames by these algorithms for the minimization of the correlation of the sum among key frames (specifically sequential key frames), they make maximum possibility of the uncorrelated key frames with one another e.g. Porter *et al*. [61] embodies the shot frames and by the usage of directed weight graph and their correlations. In the graph, the smallest path is founded and then in that smallest path the vertices designate the key frames which correspond to minimum correlation among frames.

d) Minimum re construction error: Key frames are extracted by these algorithms for the sum of minimization for the dissimilarity among every frame and by using interpolation their corresponding predicted frames are reconstructed from the set of key frames. For certain applications like animation etc., these algorithms are useful. An iterative procedure has been used by Lee and Kim [58] for the selection of number of the predetermined key frames for the reduction of the shot reconstruction error at its level best. A key frame selection of the algorithm has been proposed by Liu *et al*. [62] which is based on the amount of key frames which during the shot records the motion. An inertia based frame interpolation algorithm is used to insert frames in the algorithms.

The qualities of the above mentioned global contrast based algorithms are as follows. 1) The global characteristics of the shot are reflected by the key frames. 2) The key frame numbers are convenient. 3) The key frame sets are more concise and less superfluous as compared to the algorithms based on sequential comparison. The global comparison based algorithms limitations are that they are more expensive computationally as compare to the sequential comparison based algorithms.

5) Events/objects: In [63] these algorithms are considered as the extraction of key frame and detection of the event or object to ensure that the selected key frames have the object or event information. The position regions have been used for the frame segmentation for the key frames extraction in which objects can be amalgamated as proposed by Calic and Thomas [64]. Kim and Hwang use the shape features for the extraction of the key frames which can show the human gesture changes [65]. Liu and Fan selected the preliminary key frames which are based on the colour of the histogram and it uses the chosen key frames for the estimation of a GMM for the segmentation of the object [66]. The trained GMM and the segmented results were additionally used for the refinement of the preliminary key frames. Song and Fan suggested the key frame extraction joint along with that the segmented method of the object by the construction of the combined place for both processes in it the key frames formulation are extracted as the selection of feature process for the segmentation of the object with context to the GMM based modelling of a video [67]. Liu *et al*. proposed perceived motion energy's model of triangle for the purpose of videos patterns motion [68]. Motion declaration and the acceleration motion were selected as key frames in the turning points of the frames. [69] Han and Kweon extracted the key frames with the maximum bend of chronological size at the motion camera. In the classification of video proceedings temporal interest have been provided by the key frames. The motion patterns of objects or the reflecting objects are the merits of the object or event based algorithms which are semantically important for the extracted key frames. Algorithms' event or object detection firmly depends on heuristic rules precised by the application are the limitations. Thus, in return such algorithms are only efficient when the settings are selected cautiously.

6) Clustering based: the frames are clustered by the algorithms and then the chosen key frames were clustered to the centre of the closest frames. In the colour feature space Boreczky and Girgensohn [70] selects the frames by utilizing the entire link procedure of clustering base on agglomerative hierarchical. Yu *et al*. proposed the nebulous K-means clustering for the colour feature subspace to excerpt key frames [71]. Gaussian mixture model (GMM) has been used by the Gibson *et al*. [72] in the image space in which the obligatory quantities of clusters are GMM, i.e. the number of components. The advantages of clustering based algorithms could be classified as the universal features of a video which can be imitated in the extracted key frames as well as the serviceability of universal algorithms of clustering. Algorithms' restrictions are as following: primarily dependent on the clustering results, it is challenging to get semantically evocative clusters successfully, specifically for huge data; secondly, video can't be processed traditionally due to the chronological nature of it. Classic gawky techniques are used to safeguard that neighbouring frames are clustered together within the specific cluster.

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

C. Segmentation of Scene

Segmentation scenes are called story unit segmentation as well. Usually a story segment/scene is known as the continuous shots which are lucid along with the specific theme or subject. As compared to shots the scenes have a superior semantics level. The identification or segmentation of the scenes is by assembling the successive shots via the content similar to it in an evocative unit of semantics. Images, texts and the audio track in the video are that information on which the grouping is based. The approaches of scene segmentation are classified in three groups in accordance with the representation of the shots which are as following: ocular and aural information based, key frame based, approach based on the background.

1) Ocular and aural information based: A shot boundary is selected by the following approach in which the contents of visual and acoustic change happen at the same time in the form of the boundary scene. Sundaram and Chang identified the visual and acoustic scenes' separately [73]. Algorithm named time constrained nearest neighbour has been used for the determination of

the association among both scenes set. The visual and acoustic integration based approach constraints are that the establishment of the relation among audio segmentation and visual shots is very difficult.

2) Key frame based: The following approach signifies every shot of the video in a set of key frames from which features have been taken out. In a scene, close shots along with the features are grouped temporally. Hanjalic *et al*. they calculate the similarities among shots by contrasting the key frames [74]. Same shots have been linked by connecting overlapping links the scenes are segmented. The motion trajectories have been extracted and analysed then they are encoded in the form of image volumes of temporal slices. To efficiently represent the shot contents a selection of motion based key frame has been used. In the neighbouring shots of the key frame, the scene changes have been detected by measuring its similarity. The key based approach limitations are that key frames are not able to efficiently show all the dimensions of contents of the shots because in shots the scenes were usually related with the dimensions of the contents in the scene rather than in shots by frame based key similarities.

3) Background Based: The main theme about this approach is background similarity of same shots. Chen *et al*. a mosaic technique has been used by them for the reconstruction of every single frame of the video [75]. After that, all the background images of a shot have the estimation of the colour and texture distribution for the termination of the similarity of the shot as well as the film making rules have been used for the direction of the grouping shot process. The background base approach limitations are the hypothesis that the backgrounds are similar in the shots of the similar scene but sometimes backgrounds were different in single scene of the shots.

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

a) Splitting based approach: by the usage of top down style, this approach separates the entire video in the form of separate coherent scenes. The example of Rasheed and Shah [76] can explain it; they constructed for the video a shot of similarity graph and used the normalized cuts to divide the graph. Every scene of the video is the representation of the sub graphs. A system has been presented for narrative films to cluster the related shots in a scene Tavana pong and Zhou [77].

b) Merging-based approach: To make a scene in a bottom up style this approach steadily amalgamates the same shots. Two pass segmentation of the scene algorithm has been recommended by Shah and Rasheed [78]. By the usage of background shot coherence over segmentation of the scenes are taken out in the first pass. By using motion analysis in second pass the segmented scenes were identified and after that they were amalgamated. Merging algorithm has been proposed for scene segmentation the best first model by Zhao *et al*. [79]. In the successive shots by HMM left-right algorithm obtains every shot as a concealed state and on the boundaries' loops.

c) Classification based approach shot boundary: For the classification of shot boundaries in the scene and non-scene boundaries, extracted the features of shot boundaries the features of shot boundaries in this approach. A genre independent method has been presented by Goela *et al*. for the detection of boundaries scene into videos broadcasting [80]. Scene change and non-scene change on these two classes the scene segmentation is based on in their method. The shot boundaries have been issued by SVM for the classification. For the SVM, positive and negative training samples are generated by hand labelled video scene boundaries from various genres broadcasts. The similarities among the different shots are utilised to gather the identical shots into scenes.

d) Model based statistical approach: To segment the scenes, this approach builds statistical models of shots. Stochastic Monte Carlo sampling has been used by [81] Zhai and Shah for the scenes simulation. By having estimation in the previous step the scenes boundaries have been updated by diffusing, splitting, and merging the boundaries of the scene. The GMM has been used by Tan and Lu for the video shots clusters in scenes according to individual shots features [82]. With the Gaussian density, each scene is modelled. The unified energy minimized framework has been defined by Gu *et al*. in it the constraint global content among single shots, and the constrained local temporal among neighbouring shots has been presented. The optimal boundaries scenes are decided by the voting procedure boundary [83].

Main universal point in the approaches such as merging, statistical model based, and splitting based which is spontaneous and uncomplicated. In these approaches. A set of selected key frames is one of the approaches that embodies these shots, although they lack in representing the dynamic shots contents. In its consequences, both the shots are considered parallel, in that case if their key frames are in the similar environment, or if they are same visually. The advantage of the local information regarding shot boundaries has been taken by the shot boundary classification based approach. Thus, it approves that, algorithms which have low complexities, they are accessible. The accuracy of scene segmentation reduces because of the predictable lack of global knowledge about shots. It has been identified that, the characteristics of specific video domains such as movies, news broadcasts, and TV's has been used by the most current approaches for scene segmentation, for instance, the usage of the rules of production with which TV shows composition are done [84] [85] [86]. Scene segmentation accuracy has been improved, although, a priori model construction is important for each application.

III. EXTRACTION OF FEATURE

Results are the base of video retrieval and indexing, for the extraction of features with accordance to the video structural analysis. The visual features were the researcher's focus for the video indexing and retrieval. They basically consist of key frames features, motions and objects. They do not consist of the features of text and acoustic features.

A. Key Frame's Static Features

At some level, the characteristics of the video are reflected by the key frames of the video. To get key frames in video retrieval, conventional techniques for image retrieval are implemented. The main classification of the static key frames features which are constructive for video retrieval and indexing are colour based, shape based, and texture based.

1) Features based on colour: Colour histograms, a mixture of Gaussian models, colour moments, colour coral grams etc. are in the features of colour based. Colour based feature extraction are dependent on the spaces of colour for example, the HSV, RGB, YCBCR, HVC and normalized r-g and YUV.

On the applications, the choices of colour space are dependant. It is possible to extract the feature of colour from whole image or from set of images that are portioned into a full image. For video retrieval and indexing the colour based features have been the most effectual features for image. Specifically, the colour moments and colour histograms were proficient but easy descriptors. Colour histogram and colour moments for retrieval of the video and concept detection have been computed by Amir *et al*. [87]. Firstly, the images were sliced into 5 x 5 chunks to incarcerate in formation of the colour by Yan and Hauptmann [88]. After that, for video retrieval, each colour histogram, colour moments and block are extracted. Colour corral grams have been used by Adcock *et al*. [89] for the video search engine implantation. Colour based features replicate the human ocular view and they are simple in extraction as well as the extraction computational complexity is low; these are its qualities. Colour based features does not describe texture the directly and shape etc.; this is its drawback. Thus, the application where shape and texture are essential, the colour based features are not effective.

2) Texture Features Based: They are relevant to surface owned to object visual built-in features; they are autonomous in intensity or colour, as well as in images, they reflect the homogeneous phenomena. About the organization of object surfaces, they got very important information, along with that, its association with the environment around it. Simultaneous autoregressive models, co-occurrence matrices, wavelet transformation-based texture features, Tamura features, orientation features are included in frequent use of texture features. Amir *et al.* used tasks for video recovery in TRECVid 2003 including co-occurrence texture as well as Tamura features using contrast and coarseness [87]. Gabor wavelet filters have been used by Hauptmann *et al*. [90] for the video search engine by capturing the information of texture. Twelve oriented energy filters are designed by them. The filtered outputs of the mean and variance have been concatenated in a feature texture vector. The image was divided by Hauptmann *et al*. [91] in the blocks of 5 by 5 and the texture features were computed in each block, by using the Gabor wavelet filters. Texture based features benefits are that it is applied effectively on those applications which have information of texture as a salient feature in videos. Although in non-texture video images, such features are not available.

3) Features Based on Shapes: From object contours or regions, the shape based features which define the shapes of an object in the image, can be extracted. Generally, the technique of detecting edges in the images is used and after that by using histogram the edges are distributed. The Edge Histogram Descriptor (EHD) has been used for capturing the spatial edges distribution, for TRECVid-2005 the video search task, by Hauptmann *et al*. [90]. The EHD, according to its quantized directions, has been computed with counting the pixels numbers which contribute to the edge. Images were divided in 4 x 4 blocks by Cooke *et al*. and Foley *et al*. [92]; to capture features of local shape, after that, for every block, edge histogram has been extracted. The applications in which the information shape is the main feature in videos, for them the shape based features are effective. Although, as compare to colour based features the extraction of above mentioned feature is very difficult.

B. Features of Object

Texture, size and the dominant colour etc., related to the objects of the image regions, are included in features of an object. To retrieve the videos which have the similar objects, such features can be used [93]. In lots of retrieval video systems, the faces are the constructive objects. For instance, a person retrieval system has been constructed, that has the ability to get the shots of ranked list, which have a specific person and a query face in a shot has been given, by Sivic *et al*. [94]. A method has been proposed by Le *et al*. [95] with the integration of temporal information by converting into facial intensity information, in broadcast news videos for the retrieval of faces. To assist and comprehend contents of video, the texts in the video have been extracted as single object type. By increasing the semantics of a query as well as by using the Glimpse method of matching for the approximate performance of matching rather than matching it exactly, the text based video indexing as well as retrieval has been implemented by Li and Doermann [96]. Objects identification in videos has been very time taking and difficult as well, this is the drawback of object based feature. Rather than identifying various objects in various scenes, the current algorithms focal point is to identify the specific kinds of objects for instance faces.

C. Features of Motion

The distinguishing factor from the still images is the motion; it is the most important feature of the dynamic videos. By temporary variations, the visual content is represented by the motion information. As comparing to static key features and object features, the motion features were near to the concepts of semantics. In the motion of the video, the motion background is added, which is formed by camera motion as well as the foreground motion this is formed by the objects which are moving. Hence, video retrieval could be divided in two categories for motion feature based they are as following: object based as well as camera based. For video indexing, the camera based features and camera motions like: the in and out zooming, left and right panning, and up or down tilting are used. The limitation of video retrieval is by using the camera based features, that the key objects motions are not describable. In modern work, a lot of interest had been grabbed by motion features of object based. Statics based, trajectory based, and spatial relationship based objects are the further categories of object based motion features.

1) Statics Based: To model the distribution of local and global video motions, the motion's statistical features of frames points were extracted in the video. Such as, the casual Gibbs models have been used for the representation of the distribution of the spatio-temporal for the local measurements related motions, which is computed after balancing, in the original sequence, the leading motions image, by Fablet *et al*. [97]. After that, for video indexing and retrieval, the general framework statistics has been adopted. The motion vector field has been transformed by Ma and Zhang into the directional amount of a slice in accordance to the motion's energy [98]. The set of moments has been yielded by these slices, for the formation of multidimensional vector known as texture of the motion. For the motion based retrieval of the shot, the texture of the motion is used. Statics based features extraction is low in complexity of computation; this is its advantage. Although its drawback is that the object actions cannot be represented perfectly as well as the link among objects can't be illustrated.

2) Trajectory Based: In videos, with modelling the motion trajectories of objects, the trajectory features based were extracted [99]. An on-line video retrieval system has been offered by Chang *et al*. [100], it supported the spatio-temporal queries and object based automatic indexing. Algorithms for the video automated segmentation of the object and tracking are the part of the system. A trajectory based motion indexing compact as well as the proficient retrieval mechanism for the sequences of the video has been presented by Bashir *et al*. [101]. By sub trajectories of temporal orderings the trajectories are represented. With the motion model, principal analysis coefficients components of the sub trajectories were represented. For the segmentation of the trajectory and to create, on velocity features, an index based wavelet decomposition was used by Chen and Chang [102]. Motion model was introduced by Jung *et al*. [103] for curve fitting of polynomial. For the individual access, the model of motion is used as a key indexing. To generate information in the form of the trajectory, for recurrent motion information and build from motion vectors a motion flow which was embedded in MPEG bit streams by sue *et al*. [104]. A set of similar trajectories by given trajectories have been retrieved by the system. The trajectories were divided into many small segments as well as every segment was defined with a semantic symbol by Hsieh *et al*. [105]. For video retrievals; trajectories are matched by a distance measurement along with the exploitation of edit distance as well as visual distance. Object actions are describable; it is the advantage of the trajectory based feature. Its disadvantage is that, on correct object segmentation, as well as the trajectories automatic recording and tracking, its extraction is dependant and these are very difficult tasks.

3) Relationship Based Objects: among the objects such features explains the spatial relationship. For the video retrieval application, Bimbo *et al*. [106] described the link among the objects by using the symbolic representation scheme. By the expression of every object, the arrangements of several moving objects and the specification of the spatiotemporal relationships were query by Yajima *et al*. [107]. The relationship among the objects of several types in the temporal domain can be represented spontaneously this is the relationship-based features objects advantage. Object and position labelling is difficult this is the drawback of these features.

IV. Data Mining Of Video, Annotation & Classification

On extracted feature video and video analysis structure, the video data mining, annotation and classification are profoundly dependent. In video annotation, classification and data mining have no boundaries. Specifically, the video annotation and classification's concepts are quite alike. The basic approaches and concepts are reviewed for the video data mining, annotation and classification in this particular section. The foundation is annotation for the discovery of videos semantic, indices semantic, concepts and the production of video.

A. Video Mining Data

Video data mining errands are: finding the patterns of structures of video contents, to use the extracted features, scene characteristics of contents, the pattern behaviours of objects which are in motion, patterns event and its links [108][109] and knowledge of video semantics [110], as well as achieving the video retrieval and video intelligent applications [111]. On application, choosing the strategy for data mining video is dependent. These current strategies are as following:

1) Object Mining: The manifestations of different parts in a video are grouped in the same object of different instances is called object mining. Although, object mining is very difficult as from one illustration to another an entity can transform completely in appearance. Sivic and Zisserman [112] used the method of spatial neighbourhood for cluster fraternization in the domain of dimensional frames. In the key frames, the constantly appearing objects are mined with the help of those clusters. Stable tracks from the shots are extracted by Anjulan and Canagarajah [113]. Thus, in mining familiar objects, these tracks of stable are amalgamated in the clusters of significant objects, are used. A method has been presented by Quack *et al*. [111] for the objects and scenes from videos frequently occurring mining. The object candidates are discovered by exploring the periodic spatial provisions of affine covariant domain.

2) Detection of Special Pattern: There are priori models for the application of actions and events of special pattern detection; these are traffic events, sporting events, crime patterns and human actions [114]. An appearance based method has been proposed by Laptev *et al*. [115] eight human actions are recognized by it, for example, kissing, hugging, handshake, getting out from the car and answering the phone, etc. In space time pyramids, they extract space time local features and for recognition, they employ nonlinear SVMs multichannel as well as spatial temporal bag of features are constructed. A template based method has been proposed by Ke *et al*. [116] for the human action identifications like waving in the crowd or picking up the fallen things. To get the spatial temporal patches and shape combine and flow optical cue for matching test templates and patches, the videos are over segmented. In a football match Liu *et al*. [117] detected actions like penalty or corner or free kicks closer to the penalty line. Six traffic patterns are detected, which use HMM framework Gaussian mixture by Li and Porikli [118] and events of traffic jams are extracted by Xie *et al*. [119] with the road background features analyzing. Crime patterns are detected by Nath [120], clustering algorithms are used for that purpose.

3) Discovery of pattern: By using the semi supervised and unsupervised learning in videos the unknown patterns of automatic discovery is done this is known as the discovery of the pattern. For the exploration of the latest data in a set of video as well as the models initialization for additional applications, this pattern discovery of unknown patterns is valuable. By grouping different vector features in a video the emblematic unfamiliar patterns have been found. Following are the applications of the patterns which have been discovered: (1). Defined as having the dissimilarity with their discovered patterns, the detection of those unusual events [108]; (2). With words for retrieval of the video, etc., the association of the clusters or patterns; (3). For video annotation or classification, an algorithm based on trajectories mining for the trigger events detection, establishing the irregular or typical activities of patterns, in named categories classification of activities, clustered actions, among entities, determining the interaction, etc. are described by Burl [121]. Suffix trees and n-grams have been used by Hamid *et al*. [122] by analysing the sub sequence events upon various temporal scales for mining the patterns of motion. For the detection of unusual events, these mined motion patterns have been used. A generative model, for capturing and representing the various classes of activities, as well as for building the affine and invariance outlook for the activity of clustering the distance metric, has been used by Turaga *et al*. [123]. With the semantically significant activities, the cluster communicates. An object's self-similarity has been computed, while it evolves in time and it applies frequency of time analysis for the periodic motion detection and characterization, by Cutler and Davis [124]. By using the 2-D lattice inherent structures in matrices of similarity the periodicity has been analysed.

4) Video Association Mining: : To discover the inherent relations among various events as well as the most usual connotation forms for various entities, for instance, the presence of multiple entities at the same time, shot switches frequency and types of video associations [111], basically the video association mining is used. Mining of the video association comprises of the inter associations deduction, in the similar shot among the semantic concepts, from the explanation or the intrusion of the existing semantic concept by the detection of the neighbouring shots results, etc., for the recent shot. To find the connection among various events of news programs for instance, "volcano and earthquake" or "wine and tourism", an algorithm has been proposed by Pan and Faloutsos [125]. With incorporating features of the discrete data of the video for video associations, the explicit definitions and measurement evaluation has been proposed by Zhu *et al*. [126]. To construct the video indices this algorithm establishes the association of multilevel sequential mining to find the relations among audio-video cues. Various multi concept learning relational algorithms has been described by Yan *et al*. [127] which comprises of united representation of graphical probabilistic model and the graphical model has been used to mine the relationship among the concepts of video. To advance in the accuracy of semantic concept detection, mining association techniques has been used by Lie *et al*. [128] for the inter concept associations discovery in the concepts detection as well as temporal inter shot mine dependence.

5) Tendency Mining: The mining tendency of trends is the discovery and examination of different proceedings by current events tracking [111]. The video mining technique has been proposed in which the dual visual graphs take part i.e. the time space distribution and time tendency graph, by Xie *et al*. [129]. The distribution of the time space graph captures the temporal spatial link among different events and the graph of time tendency captures the events tendencies. The traffic jam tendency has been mined by Oh and Bandi [130] by the examination of the relations of spatial temporal in videos objects.

6) Mining preference: The preference of user is mineable [111], like news, movies and videos, etc. A personalized multimedia news portal has been proposed by Kules *et al*. [131] by user preferences mining to give a personalized service of news.

B. Classification of Video

The video classification can be defined as to search the knowledge or rules of video, which use extracted features or results, which are mined and after that it is assigned to the predefined categories of the videos [132]. To escalate the competency of video retrieval, video classification is the significant method. In extracted formative information, the semantic gap like texture, colour and the information interpretation of the observer formulates the video classification of the content based video very complex. Editing effects and semantic contents are included in video content. On three stages, the semantic content classifications are performable objects and events are having limited and thinner range of detection but video objects, video genres and video events genres are having wider and rougher range of detection. Subsequently, the edit effect classification, genre classification, event classification and object classification will be discussed.

1) Classification of Edit Effect: : On the methods of editing video the effects of editing are dependent, for example, the composition of shots and scenes as well as the camera motion. Although editing effects are not the video content components still it affects the video content perceptive for that reason in video semantic classification they are used. For example, the shots of soccer videos are classified by Ekin *et al*. [133] by using the features of cinematic into a medium in field views, out of field and close up along with that play, break, and replay like events are also detected. For the classification of global video into soccer frames, zooming and close up views as well as play or break game's statuses from the labelled sequences of frames used by the Xu *et al*. [134]. By using data from MPEG stream Tan *et al*. [135] estimated the motion of the camera and then classifies the shots of basketball in wide angles, close ups and identifies basket shots and fast breaks events.

2) Classification of Video Genre: : Classification of different videos genres like ”movies”, ”sports”, ”cartoons”, and ”news” is known as Video genre classification. Video genres approaches are categorized in machine learning, statistic and knowledge based.

a) Statistic based approach: By modelling statistically it categorizes the videos of numerous genres of games. Videos like car race, news, animated cartoons, commercials and tennis are classified by Fisher *et al*. [136]. Initially, motion of camera, colour statistics and motion of objects of syntactic video properties are analysed. Secondly, to get more of film style abstract attributes like camera panning and zooming, music and speech, those properties were used. At the end in the form of film genres those identified attributes of style were mapped. Films were classified into 4 categories, i.e., horror, drama, action, comedy depending on films characteristics like chief lightening, colour difference, content motion and average length of a shot by Rasheed *et al*. [137]. By means of shift clustering the classification has been gained. For the video genres classification only few techniques employ dynamic features. The method of cartoon video classifications has been given by Roach *et al*. [138] to differentiate among the non-cartoons and cartoons, for that features of motion objects of foreground are used. Motion object foreground and camera background motion were extracted from videos, classification of the video base of the dynamic content of sequences of short videos. In it, cartoons, news and sports were included.

b) Approach on Rule or Knowledge based: Heuristic rules are used in this approach, for the classification of videos from the domain knowledge to low level features. The knowledge based method of video classification has been made by Chen and Wong [70] in it to make the rule base with insurances the related knowledge has been coded in the generative rules formation. By using the rule base the language clip has been used for the compilation of video content classification system. The rule based supervised video classification system has been proposed, in it with the supervised learning process taken from the classification rules, the higher semantics are imitated from the features of mutual usage of lower level, by Zhou *et al*. [139]. With the combination of video creation knowledge for the extraction of semantic concepts by finding different paths from videos for that three continuous analysis steps have been used: the context analysis, the multi model content analysis and the style analysis step, for that a video indexing and classification method has been proposed by Snoek *et al*. [140]. For the application of video content analysis, clustering techniques and feature extraction in the clustering semantic videos, Zhour *et al*. [141] proposed the video classification system based on rules. On the videos of basketball experiments were reported.

c) Learning based machine approach: For the training of classifier or the classifiers set of videos, this approach applies the samples which are labelled samples having the low level features. The Bayesian networks were used for the video classification by Cheong and Mittal [142]. The links among the nonparametric and the continuous are construed by space descriptor as well as the error of least classifier by Bayes. By the usage of SVMs based active learning Qi *et al*. [143]

recommended the framework of video classification. As an input for the framework, the dataset of all the videos used for clustering have been utilized. Step by step, through the process of active learning, the classifiers accuracy enhanced. To get the hierarchical classification of semantic videos to permit access to the extremely competent video contents, Fan *et al*. [144] utilized the multiple video contents concepts. The genres of commercials, cartoons, music, sports and news are classified in videos by Truong *et al*. [145]. The average length of shot and each kind of transition percentile, etc. are the part of features. For the labelling of genre, the decision tree has been used to construct the classifier. Depending on the hierarchical video genres ontology, Yuan *et al*. [146] represented the automated video genre method of classification. In a binary tree, the series of SVM classifiers are combined and it allocates to the genres its video. A classification of online video semantic framework has been offered, which contains the optimized sets of global and local models classification and they are trained online by adequately exploiting local and global videos containing statistic characteristics, by Wu *et al*. [109]. By the usage of support cluster machines Yuan *et al*. [147] understood the demanding large scale dataset theory. Classification approaches of video genres which were mentioned previously the crux can be that: i). either static or dynamic or the combination of both is used by this approach. ii) For the global application statistically the minute level feature has been preferred by the approaches to be used. For the video genre classifications, those features are very strong and appropriate for video diversity. Depending on these low level features plenty of algorithms are trying to put few semantic features. iii) In genres video classifications the pre domain knowledge has been extensively used. The exceptional domains accuracy is improvable for the utilization of rules and knowledge, although from other domains, the related algorithms of videos are not possible to be comprehensive.

3) *Event Classification:* That video content that can be significantly represented by the human visible occurrence is defined as event. On number of events every video has been consisted of and on every event plenty of sub-events can be consisted of. The important content based video classification for the classes of events determination in the video it has been linked in a data mining of video with the events recognition [148]. On event classification a lot of work has been published. To find the events like ball possession and hand ball by the team, Yu *et al*. [149] in broadcast soccer videos and ball trajectories extractions detected the track balls. Those shots which were identified as highlights have been erudite by using the HMM models in the classification of baseball videos highlights, they were detected by Chang *et al*. [150]. For the videos of sports a proposal of the visual representation feature model has been given by Duan *et al*. [113]. It is a top down semantic shot classification model with the amalgamation of supervised learning for the performance. For the purpose of higher level semantic analysis, those semantic shots were used as the representative of midlevel. The HMM framework for semantic analysis of video has been presented by Xu *et al*. [151]. For the modification of complex analysis problems these problems are converted into sub problems and for that purpose semantics are granulated in different maps to the hierarchical model. Along with that for the detection of basketball event the framework is applied on it. For the event detection of the feature domains and grey levels, the extension of natural anti face method have been recommended by Osadchy and Karen [152]. For the play and no play detection of the sports concepts of videos HMM and dynamic programming has been utilized by Xie *et al*. [153]. In the videos of sports, the HMM and the visual features has been extracted by Pan *et al*. [154] to identify the slow motion replays. Event classification algorithms which have been defined previously, its crux can be defined as: 1) more complex feature extraction is needed for event classification as compared to the genre classification. 2) For the recognition of motion events, few methods of classification events require just the dynamic features in which the exact moving objects or region measuring motion measures are involved, afterwards these motions are classified. Sometimes the complex measurements of motions have been connected to event classifiers.

4) *Object Classification:* In video classification, the video object classification link with video object data mining is the considered to be at the least grade. Face is mostly identified and the object classified frequently. Objects extraction of structural features and its classification has been needed mostly for the detection of the object. In the processing of feature extraction of object and its classification, the previous information has been used, for instance the object appearance model has mostly been used in it. For the categorization of video shots object based algorithms has been proposed by Hong *et al*. [155]. By the usage of trajectory, texture, and colour the shots of the objects have been shown. For the correlative clustering shots the neural network has been used in it every cluster has been categorized into twelve parts. By getting the accurate matching of the cluster, the classification of object is completed. A method to categorize four kinds of television programs has been proposed by Dimitrova *et al*. [156]. In the video segments, faces and texts have been tracked and identified and they are used to label every frame. The symbols of observation frames are labelled by an HMM training for every type. Classification of video objects only works in particular environment and it is not generic for video indexing, and these are its restrictions.

C. Video Annotation

Various prerequisite concepts of semantics like car, person, people walking, and sky are the video segments and these are the shots for the allotment of the video annotation [157] [158]. The classification of the video is different in the ontology of category or concept than video annotation though few ideas can be applicable on both. In video annotation video shots or segments are applied where as video classification is applicable to complete videos, these are the only two differences among video annotation and classification, although they are quite similar. These are the following analogous methods of video annotation and video classification: foremost, the low level features have been extracted than different classifiers have been modified to chart the concept or category of the features labels. Video can be interpreted in different concepts with correspondence of its facts. Consequently, the reality of the video annotation with numerous concepts, they can be defined in the annotations of concept based, context based and integrated based [159].

1) Concept based Isolated Annotation: In a visual lexicon, this procedure of annotation has been used for every concept as a statistical detector trainee. For the discovery of multiple concepts of semantic, the classifiers of isolated binary have been utilized separately and autonomously although correlation among the perceptions has not been measured. The distribution of multiple Bernoulli has been used by Feng *et al*. [160] for the image and video annotation sculpturing. The focal point of this model of multiple Bernoulli is clearly on the word annotation presence and absence, although it is done with the postulation that every word is independent than other words in an annotation. For every concept, the accuracy of different classifiers like GMM, HMM, kNN, and Adaboost have been inspected by Naphade and Smith [161]. For the performance of video annotation semantic Song *et al*. [143] bring in the active learning together along with the semi supervised learning. A number of two class classifiers have been used in this method to take out of it with multiple classes of this classification. Depending on the construction of effective midlevel representations for the performance of classification of video semantic shots for sports video Duan *et al*. [162] utilized supervised learning algorithms. To load up the concept detectors in a single discriminative classifiers as well as to hold the errors of classifications which happen when in the feature space classes extend, a strategy of a cross training has been given by Shen *et al*. [163]. The link among the different concepts has not been sculptured and this is the restriction in the isolated annotation of concept based.

2) Context Based Annotation: By using different contexts for different concepts the concept detection performance can be improved [164]. By using the context based concept fusion approach the context based annotation purifies the results of detection of the binary classifiers individually or concepts of deduced higher level concepts. With the previous reference, an ontology learning based procedure has been used by Wu *et al*. [165] to find the video concepts. To make the accuracy of the detection of the individual binary classifiers up to the mark, ontology hierarchy has been used. Model vectors have been made by Smith and Naphade [166] which is dependent on the scores detection of those classifiers which are individual; it is used to mine the not known and co-relations which are not direct among the precised ideas, after that an SVM has been trained for the purpose of purgation of the detected results of an entity. For the annotation of the video Jiang *et al*. [167] introduced the active learning methodology. Users annotate some concepts in this method for few numbers videos and then these annotations are incorporated to deduce and develop other concepts of detections manually. With the help of unverified method of clustering an algorithm has been made, which utilises enhanced pictorial ontology for the execution of the automatic video annotation of the soccer. The amounts and actions have been linked by default to the upper class ideas with the help of looking at the visual concepts proximity which has been hierarchically the part of the semantics of the higher level. For the training of the hierarchical classifiers of the video which have the sturdy relation among the video concepts, an enhancing hierarchical method, which has ontology concepts and multi-task learning, has been proposed by Fan *et al*. [168]. The individual detection has never been steady due to the detection of the miscalculation of the individual classifiers which can spread the fusion steps and because of that the division of the training samples occurs, which are for individual detection and conceptual fusion, correspondingly, due to the usual complexity of the correlations among the concepts there have not been enough conceptual fusion samples, and this is the context based annotation drawback.

3) Integration Based Annotation: The following model covers the concepts of individual and its correlation at the same time. Concurrently, learning and optimization have been done. Along with that, all the samples have also been used for the individual concepts modelling and its correlation at the same time. A new feature vector is constructed, which grabs the concept's characteristics and concept's correlation, with the help a correlative algorithm of multi label proposed by Qi *et al*. [159]. The high computational complexity has been the drawback of the integration based annotation. The accurate amount of labelled training samples is required for the effective learning as well as robust detectors and with the help of feature dimensions the required numbers enhance exponentially. For the incorporation of the non-labelled data, few methodologies are proposed to convert it in the supervised process of learning to minimize the burden of labelling. The classification of these approaches can be based on semi supervised and of active learning:

1) Semi-supervised learning: The samples which are not labelled for the augmentation of the information for the available labelled models are used by this approach. To detect co-training based video concepts and for the investigation of the different strategies of labelling in co-training, which includes non-labelled data and few numbers of labelled videos, the semi- supervised cross feature for learning has been presented by Yan and Naphade [169]. By using small numbers of samples to

learn concepts, Yuan *et al*. [170] proposed the algorithm of a feature selection based manifold ranking. They comprise of three main components: pre-filtering, feature pool construction and manifold ranking. A video annotation algorithm has been proposed, which is based upon the semi-supervised learning with the help of kernel density estimation [171]. To handle the insufficiency of the training data in video annotation the semi-supervised learning algorithms based upon the optimized multi-graph has been proposed by Wang *et al*. [172]. The partially supervised learning system for the adaptive learning of different forms of objects and events for the specific video has been given by Ewerth and Freiskeben [173]. For the feature selection and total classification Adaboost and SVM has been included.

2) Active learning: it is a very useful and quick way for the handling of the lower sample brands. For the annotation of the video which comprises of the multi complementary predictors and adaptation of the increasing model, the algorithm of active learning has been proposed by Song *et al*. [174]. Along with that, it Song *et al*. [175] gave a framework of the video annotation proposal which was specifically designed for the personal databases of videos and it comprises of semi supervised and an active learning assembled process.

V. Query & Retrieval

The video retrieval which is content based starts its performance when the video indices have been attained. For the search of the user's video in accordance to the query sent by the user, the method of similarity measurement, which comprises on the indices is used. In reference to the feedback, the repossessed results have been optimized. Bellow similarity matching and the feedback relevance type queries have been reassessed.

A. Types of Queries

Those video queries which are non-semantic based, they are for instance: query by objects and query by subject. Those types of video queries which are semantic based are: query by natural language and by keywords.

a) Query by Example: From the sampled videos and images, this type sifts the slow level features and with that by calculating the similarity of the given features the exact or similar videos has been found. For the query by example, the static key frames features suits best, as the stored key frames can be matched through the sifted key frames from the sampled videos and images.

b) Query by Sketch: By making sketches this query permits the user to get their desired video. The sketched features have been extracted and are matched with the stocked videos. A query by sketch method has been given by the Hu *et al*. [176] in it the path extracted from the videos are matched with the path sketched by the users.

c) Query by Objects: With the help of this query, a user can give the object's image. In the video database, all the events of the object are returned, when the system finds them [177]. By comparing the query by example and sketch, the found results of the query by objects are at the query object in the videos.

d) Query by Keywords: It portrays the questions of users with the help of some key-words. This methodology is easy; it mostly gets the semantics and the main type of query. Video metadata, concept visuals and transcripts can be the part of keywords. The main focus of this paper is about concepts of visuals.

e) Query by Natural Language: The methodology of using natural language is the easiest way to make query. A semantic word comparison has been used by Aytar *et al*. [178] for the repossession of the pertinent videos and then it divides them with reference to the query given with the natural use of language, mostly English. Parsing the natural language and attainment of exact semantics are the hard portions for the natural language.

f) Combination-Based Query: It is the amalgamation of different kinds of queries like text and video pattern based queries. It is flexible for the multiple model findings. A frame work has been developed by Kennedy *et al*. [179] for the automatic finding of the grades of query with the help of query division in the training format with accordance to the different single model search method routine. The adaptive method has been given by Yan *et al*. [180] to mingle certain search gears for the application of the video retrieval query class dependency and in it the query class alliance weights of certain search gears are resolute automatically. Space query by person and without a person quires are differentiated by the multimedia system of retrieval. With the treatment of query classes as the latent variables Yan and Hauptmann [181] took the queries classification as well as combination weights determination into the structure of probabilistic.

Bellow mentioned query interfaces were the most known ones.

1) The informedia interface [182] [183]: Depended upon the video semantic concepts this interface endorsed filtering. When the key word search is taken out than the filters of visual concepts are applicable.

2) The query of Mediamill interface [184] [185]: This interface amalgamates the concept query visual, text keyword query, and the example query.

B. Similarity Measure

In the content based video retrieval, the video similarity measurement has the vital role. Feature matching, text matching, ontology based matching, and combination matching are the methodologies for the measurement of the video similarities. On the type of query, it is depended to choose the method.

1) Feature Matching: The general space among the corresponding frames features is the accurate similarity of measurement in between the two videos. For the findings of the similar videos, query by example mostly incorporates the low level features. In various granularities and resolutions video similarity can be taken [186]. For the measurement of the video similarity; the static features of key frames [187], motion features [176], and object features [113] all of them can be used in accordance to the user demands. Sivic *et al*. [94], sifted the features from the sample shot which had to face queried and then with the saved face features those extracted features were matched. After that the queried faces were retrieved, which were in the shots. In the set of videos, Lie and Hsiao [188] sifted the major objects trajectory features, after that for the videos retrieval it matched the sifted trajectory features with those trajectory features which were stored. In the space of the feature, the similarity of the video can easily be measured, this is the plus point of the feature matching. The negative point is that the similarity of semantic is not representable due to the spaces among the sets of vectors features as well as to people semantic categories are very similar.

2) Text matching: To find the video, name equivalence of every idea by its query is the easiest method and this method also suitable for the query. With the help of vector space query text and text description were computed to find the similarity among them, before that both the concept description and the text query were regularized by Snoek *et al*. [189]. The most similar ideas were chosen at the end. Text matching approaches instinctiveness and easy implementation are the pros of it. To get the appropriate results of the research complete matching concepts should be completely added in the query text, this is the drawback of this approach.

3) Ontology based matching: The matching of keyword semantic relation and semantic concepts resemblance ontology are done by this approach. From the knowledge sources like, concepts ontology and keywords, query descriptions have been developed. The words in the query text are disambiguated syntactically after that noun chunks and nouns are translated by finding every noun from word net; they are extracted from the concepts ontological concepts Snoek *et al*. Because of the link of the concepts to the word net. Thus, to find the concepts relation to the actual text query is determined by the usage of ontology. On the facts bases word similarity semantic is a worthy visual co-occurrence approximate. For the similarity findings of the text annotated videos and quires of users, Aytar *et al*. [178] used semantic similarity word measures. By the help of text query defined by a user, the videos are retrieved depending upon its relevance. For the improvement of the retrieved results, the additional concepts are incorporated from the knowledge sources [190], this is the benefit of the ontology based matching approach. The concepts which are inappropriate are taken, which leads to the unwanted corrosion of the findings; this is the negative point of this method.

4) Combination based matching: The combination methodologies are learned by the training collections; such as a combination of learning & query independent models, and combination of query & class models and this technique empowers the semantic concepts [88]. This approach is very much handy for quires based on combination, which has the multi modal searches flexibility. Combination based matching approach can automatically determine the concept weights as well as by far it can handle the hidden semantic concepts, these are its benefits. Query combination models are not easy to learn; this is its drawback.

C. Relevance Feedback

In this approach, those videos which are gathered in reply to the searched queries are graded automatically or by the user. For the refinement of the advance searches this ranking has been used. This method of refinement involves the optimization of the query point, weight adjustment feature, as well as embedding of information. The distance among the semantic notions of low level video content representation and search relevance, are decreased by the relevance feedback. It also has an impact on the preferences of the user; relevance feedback takes user feedback in reference of the earlier searched results. Just like relevance feedback for image relevance for the video retrieval the relevance feed is divisible into three ranks: explicit, implicit, and the pseudo feedback.

VI. FUTUREWORK

Great deal of research has been implemented in visual based video retrieval and indexing. However, following areas still need to be addressed further for its future development. They are as follows: a) hierarchical study of video contents; b) Analysis of a motion feature; c) classified video indices;

d) Human computer interface e) unity of multimodal; f) video indexing (extensible); g) semantic-based video retrieval and indexing.

A. Hierarchical Study of Video Contents

A video might have diverse connotations at certain semantic stages. Ranked group of audio-visual notions is needed for semantic-based retrieval and indexing of video. Analysis of video contents involves the breakdown of huge levelled semantic notions into sequences of lower levelled elementary semantic theories and their limitations. Lower levelled elementary semantic theories be able to be linked directly with lower levelled structures. On the other hand, higher levelled semantic notions be able to be inferred from lower levelled elementary semantic impressions through numerical examination. Additionally, creating hierarchical semantic relationships between shots, key frames on the basis of study of video structure, scenes, hierarchical visualizing and arranging results of retrieval and creating three different stages of connections between classifications, i.e. a) genres; b) object and: c) event. All of the above are important further research issues.

B. Analysis of Motion Feature

The information of analysing the motion features is significant for content-based video retrieval. In order to differentiate between fore and background motion, detection of moving events and objects, connected static and motion landscapes, and creation of motion-based indices are essential areas of research.

C. Classified Video Indices

Video directories classified video examination can be used in video indexing. Lowest level in the grading is the model of index stock matching to the upper most dimensional index of feature arrangement. Semantic index model is the upper most levels, defining the semantic notions moreover; their associations in the videos require retrieval. Mode of index context is the middle layer that associates the store model and semantic notion model. Online, adaptive modernising of the classified index model, bringing dynamic measures of video resemblance created on statistic feature assortment up to date, handling sequential video features for the period of index creation and rapid video exploration by means of hierarchical indices. Described examples are important and interesting questions for research.

D. Human Computer Interface

It has a capacity to accurately transfer the intentions of a query and precisely develop outcomes of retrieval. Additionally, output of video with representation of multi-model is more vivid and visual. The design of multi-model data in the interface of human computer, interface's success in rapidly capturing the most attracted results to the users, the interface appropriateness for the users' feedback and evaluation and efficiency of an interface for familiarising the users' habits query and personality expressions are the topics for investigation of future.

E. Unity of Multi-models

A combined integration of several models is usually termed as the semantic video content. Merging evidence from several models are useful in retrieval of content based video. Explanation of progressive associations among diverse types of evidence from models of several kinds, dynamic features weighting of certain models, combination of information from several models expressing similar story and unity of information of several models in numerous stages. All of the mentioned types are challenging issues in the overall examination of combined models.

F. Video Indexing (Extensible)

Latest approaches on video indexing mainly rely on the knowledge of prior domain, which resulted in its restriction towards new areas or domains. Absence of depending on the knowledge of domains is a research question for the future. The problem could be eliminated by using extraction of feature involving less knowledge of domain, in addition to dynamic creation of rules of classification by means of techniques called rule-mining.

G. Semantic-Based Video Indexing and Retrieval

Existing methods designed for semantic-based video retrieval and indexing typically exploit group of writings in order to pronounce the graphic video contents. Though various kinds of programmed semantic perception sensors have been established, many questions are still unanswered and require further investigation, e.g. way of choosing the features which are maximum representation of semantic notions? How to construct the large-scaled believed video ontology? By what means valuable common perception sensors with evaluated recovery to be nominated? How many useful concepts are required? How elevated ideas can be incorporated automatically into video retrieval? How ontology can be built up for interpreting the query into terms concept handling? How to reconcile the unpredictable explanations causing the different people's interpretations of the same visual data be reconciled? How to intricate the established ontology between lexica detection? How to accurately fuse the concepts of multimodality fusion? How to obtain and fuse the accurate concepts of certain approaches for machine learning?

VII. CONCLUSION

A comprehensive review on content-based visual retrieval and indexing is presented in this paper. The emphasis remains on: 1) detection of shot boundary for analysis of video structure, which is a key feature in video indexing and retrieval systems, the indicators are provided so that the holistic perspective could be seen, 2) the extraction of key frames for scene division, along with the state of the art approaches is reviewed the effort has been made to present the research progress in this area in chronological order so that it can present a clear look for the progress of research in this area 3) feature extraction from frames are then seen from the machine learning and image processing perspectives to give the understanding of the existing and future trends 4) features of motion, objects and the treatment of moving subjects in the video retrieval is looked at with the approaches in the area. 5) Data mining for video and classification is also included to give the insights of the approaches used in video retrieval domain with their overall integration of the other components of the process is given. Video indexing development throughout the years can be summarized as shown in Table 1.

TABLE 1 VIDEO INDEXING DEVELOPMENT

Theme	**Citation**	**Authors**	**Year**
Indexing with reinforcement agent	[191]	Paul *et al.*	2013
Interactive with statistical active learning	[192]	Zha *et al.*	2012
Applications and browsing interfaces	[11]	Schoeffmann *et al.*	2010
TRECVID shot boundary identification	[4]	Smeaton *et al.*	
Conceptual video retrieval	[10]	Snoek *et al.*	2009
Understanding Video Events	[148]	Lavee *et al.*	
Multimedia Retrieval and Delivery	[6]	Pereira *et al.*	2008
Shot boundary detection using petri-net	[25]	Bai *et al.*	
Semantic word similarity measures for video retrieval	[178]	Aytar *et al.*	
Semanticbased surveillance video retrieval	[176]	Hu *et al.*	2007
Statisticsunsupervised activity	[122]	Hamid *et al.*	
Retrieval approaches for broadcast news video	[88]	Yan & Hauptmann	
Video abstraction	[45]	Truong & Venkatesh	
Video mining	[111]	Ke-xue et al.	2006
Semantic concepts in multimedia	[158]	Snoek *et al.*	
Discovery of query-class-dependent	[179]	Kennedy *et al.*	2005
Video categorization engine	[155]	Hong *et al.*	

Finally, based on the extensive review some indicators to the emerging trends in video indexing and retrieval domain are presented, an effort has been made to present the latest development in the domain of video indexing and retrieval with the intension that this at least presents the perspective.

REFERENCES

[1] M. Lew, N. Sebe, C. Djeraba, and R. Jain, “Content-based multimedia information retrieval: State of the art and challenges,” *ACM Transactions on Multimedia ...*, vol. 2, no. 1, pp. 1–19, 2006.

[2] Y. Peng and C.-W. Ngo, “Hot event detection and summarization by graph modeling and matching,” in *Image and Video Retrieval*. Springer, 2005, pp. 257–266.

[3] P. Over, G. M. Awad, J. Fiscus, M. Michel, A. F. Smeaton, and W. Kraaij, “Trecvid 2009-goals, tasks, data, evaluation mechanisms and metrics,” 2010.

[4] A. F. Smeaton, P. Over, and A. R. Doherty, “Video shot boundary detection: Seven years of TRECVid activity,” *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 411–418, Apr. 2010.

[5] A. F. Smeaton, P. Over, and W. Kraaij, “High-level feature detection from video in trecvid: a 5-year retrospective of achievements,” in *Multimedia content analysis*. Springer, 2009, pp. 1–24.

[6] F. Pereira, a. Vetro, and T. Sikora, “Multimedia Retrieval and Delivery: Essential Metadata Challenges and Standards,” *Proceedings of the IEEE*, vol. 96, no. 4, pp. 721–744, Apr. 2008.

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

[8] A. F. Smeaton, “Techniques used and open challenges to the analysis, indexing and retrieval of digital video,” *Information Systems*, vol. 32, no. 4, pp. 545–559, 2007.

[9] Y. Y. Chung, W. J. Chin, X. Chen, D. Y. Shi, E. Choi, and F. Chen, “Content-based video retrieval system using wavelet transform.” *WSEAS Transactions on Circuits and Systems*, vol. 6, no. 2, pp. 259–265, 2007.

[10] C. G. M. Snoek and M. Worring, “Concept-Based Video Retrieval,” *Foundations and Trends in Information Retrieval*, vol. 2, no. 4, pp. 215–322, 2007.

[11] K. Schoeffmann, F. Hopfgartner, O. Marques, L. Boeszoermenyi, and J. M. Jose, “Video browsing interfaces and applications: a review,” *SPIE Reviews*, vol. 1, no. 1, p. 018004, 2010.

[12] W. Ren, S. Singh, M. Singh, and Y. Zhu, “State-of-the-art on spatiotemporal information-based video retrieval,” *Pattern Recognition*, vol. 42, no. 2, pp. 267–282, Feb. 2009.

[13] C. Yeo, Y.-w. Zhu, Q. Sun, and S.-f. Chang, “A Framework for Sub-Window Shot Detection,” *11th International Multimedia Modelling Conference*, pp. 84–91, 2005.

[14] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang, “A Formal Study of Shot Boundary Detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 2, pp. 168–186, Feb. 2007.

[15] S. Chantamunee and Y. Gotoh, “University of sheffield at trecvid 2007: Shot boundary detection and rushes summarisation.” in *TRECVID*. Citeseer, 2007.

[16] S. C. Hoi, L. L. Wong, and A. Lyu, “Chinese university of hongkong at trecvid 2006: Shot boundary detection and video search,” in *TRECVid 2006 Workshop*, 2006, pp. 76–86.

[17] Z.-C. Zhao and A.-N. Cai, “Shot boundary detection algorithm in compressed domain based on adaboost and fuzzy theory,” in *Advances in Natural Computation*. Springer, 2006, pp. 617–626.

[18] S. V. Porter, “Video segmentation and indexing using motion estimation,” Ph.D. dissertation, University of Bristol, 2004.

[19] 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 Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.

[20] X. Wu, P. C. Yuen, C. Liu, and J. Huang, “Shot Boundary Detection: An Information Saliency Approach,” *2008 Congress on Image and Signal Processing*, pp. 808–812, 2008.

[21] X. Gao, J. Li, and Y. Shi, “A video shot boundary detection algorithm based on feature tracking,” in *Rough Sets and Knowledge Technology*. Springer, 2006, pp. 651–658.

[22] G. Camara-Chavez, F. Precioso, M. Cord, S. Phillip-Foliguet, and A. de A Araujo, “Shot boundary detection by a hierarchical supervised

approach," in *Systems, Signals and Image Processing, 2007 and 6 th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on*. IEEE, 2007, pp. 197–200.

[23] H. Lu, Y.-P. Tan, X. Xue, and L. Wu, "Shot boundary detection using unsupervised clustering and hypothesis testing," in *Communications, Circuits and Systems, 2004. ICCCAS 2004. 2004 International Conference on*, vol. 2. IEEE, 2004, pp. 932–936.

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

[25] L. Bai, S.-Y. Lao, H.-T. Liu, and J. Bu, "Video shot boundary detection using petri-net," in *Machine Learning and Cybernetics, 2008 International Conference on*, vol. 5. IEEE, 2008, pp. 3047–3051.

[26] C. Liu, H. Liu, S. Jiang, Q. Huang, Y. Zheng, and W. Zhang, "Jdl at trecvid 2006 shot boundary detection," in *TRECVID 2006 Workshop*, 2006.

[27] D. Xia, X. Deng, and Q. Zeng, "Shot Boundary Detection Based on Difference Sequences of Mutual Information," *Fourth International Conference on Image and Graphics (ICIG 2007)*, pp. 389–394, Aug. 2007.

[28] K.-C. Ko, Y. M. Cheon, G.-Y. Kim, H.-I. Choi, S.-Y. Shin, and Y.-W. Rhee, "Video shot boundary detection algorithm," in *Computer Vision, Graphics and Image Processing*. Springer, 2006, pp. 388–396.

[29] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, no. 1, pp. 82 –91, 2006.

[30] a. Hanjalic, "Shot-boundary detection: unraveled and resolved?" *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 2, pp. 90–105, 2002.

[31] G. M. Qusenot,´ D. Moraru, and L. Besacier, "Clips at trecvid: Shot boundary detection and feature detection," 2003.

[32] K. Matsumoto, M. Naito, K. Hoashi, and F. Sugaya, "Svm-based shot boundary detection with a novel feature," in *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 2006, pp. 1837–1840.

[33] C. M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E. J. Delp, "Automated video program summarization using speech transcripts," *Multimedia, IEEE Transactions on*, vol. 8, no. 4, pp. 775–791, 2006.

[34] Z.-C. Zhao, X. Zeng, T. Liu, and A.-N. Cai, "Bupt at trecvid 2007: Shot boundary detection." in *TRECVID*. Citeseer, 2007.

[35] X. Ling, L. Chao, L. Huan, and X. Zhang, "A General Method for Shot Boundary Detection," *2008 International Conference on Multimedia and Ubiquitous Engineering (mue 2008)*, pp. 394–397, 2008.

[36] P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton, "Trecvid 2005- an overview," 2005.

[37] J. S. Boreczky and L. D. Wilcox, "A hidden markov model framework for video segmentation using audio and image features," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 6. IEEE, 1998, pp. 3741–3744.

[38] C.-c. Lo and S.-j. Wang, "Video segmentation using a histogrambased fuzzy c-means clustering algorithm," *10th IEEE International Conference on Fuzzy Systems. (Cat. No.01CH37297)*, vol. 3, pp. 920–923, 2001.

[39] U. Damnjanovic, E. Izquierdo, and M. Grzegorzek, "Shot boundary detection using spectral clustering," in *15th European Signal Processing Conference*, 2007, p. 1779.

[40] S. De Bruyne, D. Van Deursen, J. De Cock, W. De Neve, P. Lambert, and R. Van de Walle, "A compresseddomain approach for shot boundary detection on H.264/AVC bit streams," *Signal Processing: Image Communication*, vol. 23, no. 7, pp. 473–489, Aug. 2008.

[41] H. Koumaras, G. Gardikis, G. Xilouris, E. Pallis, and a. Kourtis, "Shot boundary detection without threshold parameters," *Journal of Electronic Imaging*, vol. 15, no. 2, p. 020503, 2006.

[42] C.-W. Ngo, "A robust dissolve detector by support vector machine," *Proceedings of the eleventh ACM international conference on Multimedia - MULTIMEDIA '03*, no. 3, p. 283, 2003.

[43] H.-W. Yoo, H.-J. Ryoo, and D.-S. Jang, "Gradual shot boundary detection using localized edge blocks," *Multimedia Tools and Applications*, vol. 28, no. 3, pp. 283–300, 2006.

[44] K.-W. Sze, K.-M. Lam, and G. Qiu, "A new key frame representation for video segment retrieval," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 9, pp. 1148–1155, 2005.

[45] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 3, no. 1, p. 3, 2007.

[46] D. Besiris, N. Laskaris, F. Fotopoulou, and G. Economou, "Key frame extraction in video sequences: a vantage points approach," *2007 IEEE 9th Workshop on Multimedia Signal Processing*, pp. 434–437, 2007.

[47] D. P. Mukherjee, S. K. Das, and S. Saha, "Key frame estimation in video using randomness measure of feature point pattern," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 5, pp. 612–620, 2007.

[48] R. Narasimha, A. Savakis, R. Rao, and R. De Queiroz, "Key frame extraction using mpeg-7 motion descriptors," in *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. IEEE, 2003, pp. 1575–1579.

[49] T. Wang, Y. Wu, and L. Chen, "An Approach to Video Key-frame Extraction Based on Rough Set," *2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*, pp. 590–596, 2007.

[50] M. Guironnet, D. Pellerin, N. Guyader, P. Ladret*et al.*, "Video summarization based on camera motion and a subjective evaluation method," *EURASIP Journal on Image and Video Processing*, vol. 2007, 2007.

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

[52] J. Calic and E. Izuierdo, "Efficient key-frame extraction and video analysis," in *Information Technology: Coding and Computing, 2002. Proceedings. International Conference on*. IEEE, 2002, pp. 28–33.

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

[54] Z. Sun, K. Jia, and H. Chen, "Video Key Frame Extraction Based on Spatial-Temporal Color Distribution," *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 196–199, Aug. 2008.

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

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

[57] A. Divakaran, R. Radhakrishnan, and K. A. Peker, "Motion activitybased extraction of key-frames from video shots," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1. IEEE, 2002, pp. I–932.

[58] H.-C. Lee and S.-D. Kim, "Iterative key frame selection in the rateconstraint environment," *Signal Processing: Image Communication*, vol. 18, no. 1, pp. 1–15, 2003.

[59] M. Cooper and J. Foote, "Discriminative techniques for keyframe selection," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 2005, pp. 4–pp.

[60] H. S. Chang, S. Sull, and S. U. Lee, "Efficient video indexing scheme for content-based retrieval," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 9, no. 8, pp. 1269–1279, 1999.

[61] S. Porter, M. Mirmehdi, and B. Thomas, "A shortest path representation for video summarisation," *12th International Conference on Image Analysis and Processing, 2003.Proceedings*, pp. 460–465.

[62] T. Liu, X. Zhang, J. Feng, and K.-T. Lo, "Shot reconstruction degree: a novel criterion for key frame selection," *Pattern recognition letters*, vol. 25, no. 12, pp. 1451–1457, 2004.

[63] H.-W. Kang and X.-S. Hua, "To learn representativeness of video frames," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 423–426.

[64] J. Calic and B. Thomas, "Spatial analysis in key-frame extraction using video segmentation," in *Workshop on Image Analysis for Multimedia Interactive Services*, 2004.

[65] C. Kim, "Object-based video abstraction using cluster analysis," *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, vol. 2, pp. 657–660.

[66] L. Liu and G. Fan, "Combined key-frame extraction and object-based video segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 7, pp. 869–884, Jul. 2005.

[67] X. Song and G. Fan, "Joint Key-Frame Extraction and Object Segmentation for Content-Based Video Analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 7, pp. 904–914, Jul. 2006.

[68] T. Liu, H.-j. Zhang, and F. Qi, "A novel video key-frameextraction algorithm based on perceived motion energy model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13 , no. 10, pp. 1006–1013, Oct. 2003.

[69] S.-H. Han *et al.*, "Scalable temporal interest points for abstraction and classification of video events," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 2005, pp. 4 –pp.

[70] A. Girgensohn and J. Boreczky, "Time-constrained keyframe selection technique," in *Multimedia Computing and Systems, 1999. IEEE International Conference on*, vol. 1. IEEE, 1999, pp. 756–761.

[71] X.-D. Yu, L. Wang, Q. Tian, and P. Xue, "Multilevel video representation with application to keyframe extraction," in *Multimedia Modelling Conference, 2004. Proceedings. 10th International*. IEEE, 2004, pp. 117–123.

[72] D. Gibson, N. Campbell, and B. Thomas, "Visual abstraction of wildlife footage using gaussian mixture models and the minimum description length criterion," in *Pattern Recognition, 2002. Proceedings. 16 th International Conference on*, vol. 2. IEEE, 2002, pp. 814–817.

[73] H. Sundaram and S.-F. Chang, "Video scene segmentation using video and audio features," in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, vol. 2. IEEE, 2000, pp. 1145 – 1148.

[74] A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automated high-level movie segmentation for advanced video-retrieval systems," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 9, no. 4, pp. 580–588, 1999.

[75] L.-H. Chen, Y.-C. Lai, and H.-Y. Mark Liao, "Movie scene segmentation using background information," *Pattern Recognition*, vol. 41, no. 3 , pp. 1056–1065, 2008.

[76] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," *IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1097–1105, Dec. 2005.

[77] W. Tavanapong and J. Zhou, "Shot Clustering Techniques for Story Browsing," *IEEE Transactions on Multimedia*, vol. 6, no. 4, pp. 517–527, Aug. 2004.

[78] Z. Rasheed and M. Shah, "Scene detection in Hollywood movies and TV shows," *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*, vol. 2, pp. II–343–8.

[79] L. Zhao, W. Qi, Y.-J. Wang, S.-Q. Yang, and H. Zhang, "Video shot grouping using best-first model merging," in *Proceedings of SPIE*, vol. 4315, 2001, p. 262.

[80] N. Goela, K. Wilson, F. Niu, A. Divakaran, and I. Otsuka, "An svm framework for genre-independent scene change detection," in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007, pp. 532–535.

[81] Y. Zhai and M. Shah, "Video scene segmentation using markov chain montecarlo," *Multimedia, IEEE Transactions on*, vol. 8, no. 4, pp. 686–697, 2006.

[82] Y.-P. Tan and H. Lu, "Model-based clustering and analysis of video scenes," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1. IEEE, 2002, pp. I–617.

[83] Z. Gu, T. Mei, X.-S. Hua, X. Wu, and S. Li, "EMS: Energy Minimization Based Video Scene Segmentation," *Multimedia and Expo, 2007 IEEE International Conference on*, pp. 520–523, Jul. 2007.

[84] Y. Ariki, M. Kumano, and K. Tsukada, "Highlight scene extraction in real time from baseball live video," *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval - MIR '03*, p. 209, 2003.

[85] Y. Zhai, A. Yilmaz, and M. Shah, "Story segmentation in news videos using visual and text cues," in *Image and Video Retrieval*. Springer, 2005, pp. 92–102.

[86] W.-M. Hsu and S.-F. Chang, "Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation," in *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, vol. 2. IEEE, 2004, pp. 1091–1094.

[87] A. Amir, M. Berg, S.-F. Chang, W. Hsu, G. Iyengar, C.-Y. Lin, M. Naphade, A. Natsev, C. Neti, H. Nock *et al.*, "Ibm research trecvid2003 video retrieval system," *NIST TRECVID-2003*, 2003.

[88] R. Yan and A. G. Hauptmann, "A review of text and image retrieval approaches for broadcast news video," *Information Retrieval*, vol. 10, no. 4-5, pp. 445–484, 2007.

[89] J. Adcock, A. Girgensohn, M. Cooper, T. Liu, L. Wilcox, and E. Rieffel, "Fxpal experiments for trecvid 2004," *Proceedings of the TREC Video Retrieval Evaluation (TRECVID)*, pp. 70–81, 2004.

[90] A. Hauptmann, R. V. Baron, M.-y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W.-H. Lin, T. Ng, and N. Moraveji, "Informedia at trecvid 2003: Analyzing and searching broadcast news video," DTIC Document, Tech. Rep., 2004.

[91] A. Hauptmann, M. Chen, M. Christel, C. Huang, W.-H. Lin, T. Ng, N. Papernick, A. Velivelli, J. Yang, R. Yan *et al.*, "Confounded expectations: Informedia at trecvid 2004," in *Proc. of TRECVID*, 2004.

[92] C. Foley, C. Gurrin, G. J. Jones, H. Lee, S. McGivney, N. E. O'Connor, S. Sav, A. F. Smeaton, and P. Wilkins, "Trecvid 2005 experiments at dublin city university," 2005.

[93] R. Visser, N. Sebe, and E. Bakker, "Object recognition for video retrieval," in *Image and Video Retrieval*. Springer, 2002, pp. 262– 270.

[94] J. Sivic, M. Everingham, and A. Zisserman, "Person spotting: video shot retrieval for face sets," in *Image and Video Retrieval*. Springer, 2005, pp. 226–236.

[95] D.-D. Le, S. Satoh, and M. E. Houle, "Face retrieval in broadcasting news video by fusing temporal and intensity information," in *Image and Video Retrieval*. Springer, 2006, pp. 391–400.

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

[97] R. Fablet, P. Bouthemy, and P. Perez,´ "Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval," *Image Processing, IEEE Transactions on*, vol. 11, no. 4, pp. 393–407, 2002.

[98] Y.-F. Ma and H.-J. Zhang, "Motion texture: a new motion based video representation," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 2. IEEE, 2002, pp. 548–551.

[99] T. Quack, V. Ferrari, and L. Van Gool, "Video mining with frequent itemset configurations," in *Image and Video Retrieval*. Springer, 2006, pp. 360–369.

[100] W. Chen, H. Meng, and H. Sundaram, "A fully automated contentbased video search engine supporting spatiotemporal queries," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 602–615, 1998.

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

[102] W. Chen and S.-F. Chang, "Motion trajectory matching of video objects," in *Electronic Imaging*. International Society for Optics and Photonics, 1999, pp. 544–553.

[103] Y.-k. Jung, K.-w. Lee, and Y.-s. Ho, "Content-based event retrieval using semantic scene interpretation for automated traffic surveillance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 2, no. 3, pp. 151–163, 2001.

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

[105] J.-w. Hsieh, S.-l. Yu, and Y.-s. Chen, "Motion-based video retrieval by trajectory matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 3, pp. 396–409, Mar. 2006.

[106] a. Del Bimbo, E. Vicario, and D. Zingoni, "Symbolic description and visual querying of image sequences using spatio-temporal logic," *IEEE Transactions on Knowledge and Data Engineering*, vol. 7, no. 4, pp. 609–622, 1995.

[107] C. Y. Y. Nakanishi and K. Tanaka, "Querying video data by spatiotemporal relationships of moving object traces," in *Visual and Multimedia Information Management: IFIP TC 2/WG 2.6 Sixth Working Conference on Visual Database Systems, May 29-31, 2002, Brisbane, Australia*. Kluwer Academic Pub, 2002, p. 357.

[108] H. Sundaram and M. Campbell, "Event Mining in Multimedia Streams," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 623–647, Apr. 2008.

[109] J. Wu, X.-S. Hua, H.-J. Zhang, and B. Zhang, "An online-optimized incremental learning framework for video semantic classification," *Proceedings of the 12th annual ACM international conference on Multimedia - MULTIMEDIA '04*, p. 320, 2004.

[110] T. Mei, X.-S. Hua, H.-Q. Zhou, and S. Li, "Modeling and mining of users' capture intention for home videos," *Multimedia, IEEE Transactions on*, vol. 9, no. 1, pp. 66–77, 2007.

[111] D. Ke-xue, "Wu de-feng˜(1)), fuchang-jian˜(1)), li guo-hui˜(1)), li huijia˜(2))˜(1))(department of system engineering, school of info system and management, national university of defense technology, changsha 410073)˜(2))(changshajun-hong technology co., ltd., changsha 410001); video mining: A survey [j]," *Journal of Image and Graphics*, vol. 4, 2006.

[112] J. Sivic and A. Zisserman, "Video data mining using configurations of viewpoint invariant regions," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 1. IEEE, 2004, pp. I–488.

[113] a. Anjulan and N. Canagarajah, "A Unified Framework for Object Retrieval and Mining," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 1, pp. 63–76, Jan. 2009.

[114] Y. Zhang, C. Xu, J. Wang, H. Lu *et al.*, "Semantic event extraction from basketball games using multi-modal analysis," in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007, pp. 2190 – 2193.

[115] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1 – 8.

[116] Y. Ke, R. Sukthankar, and M. Hebert, "Event Detection in Crowded Videos," *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, 2007.

[117] H.-Y. Liu, T. He, and H. Zhang, "Event Detection in Sports Video Based on Multiple Feature Fusion," *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, no. Fskd, pp. 446–450, 2007.

[118] X. Li and F. M. Porikli, "A hidden markov model framework for traffic event detection using video features," in *Image Processing, 2004. ICIP'04. 2004 International Conference on*, vol. 5. IEEE, 2004, pp. 2901–2904.

[119] X. Lei, W. Qing, C. Xiumin, W. Jun, and C. Ping, "Traffic jam detection based on corner feature of background scene in video-based ITS," in *Networking, Sensing and Control, 2008. ICNSC 2008. IEEE International Conference on*. IEEE, 2008, pp. 614–619.

[120] S. V. Nath, "Crime pattern detection using data mining," in *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WIIAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on*. IEEE, 2006, pp. 41–44.

[121] M. C. Burl, "Mining patterns of activity from video data," in *SIAM Int. Conf. on Data Mining*, 2004, pp. 364–373.

[122] R. Hamid, S. Maddi, A. Bobick, and M. Essa, "Structure from statisticsunsupervised activity analysis using suffix trees," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1 – 8.

[123] P. K. Turaga, A. Veeraraghavan, and R. Chellappa, "From videos to verbs: Mining videos for activities using a cascade of dynamical systems," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1 – 8.

[124] R. Cutler and L. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 781–796, 2000.

[125] J.-Y. Pan and C. Faloutsos, "Geoplot: spatial data mining on video libraries," in *Proceedings of the eleventh international conference on Information and knowledge management*. ACM, 2002, pp. 405–412. [126] X. Zhu, X. Wu, A. K. Elmagarmid, Z. Feng, and L. Wu, "Video data mining: Semantic indexing and event detection from the association perspective," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 5, pp. 665–677, 2005.

[127] R. Yan, M.-y. Chen, and A. Hauptmann, "Mining Relationship Between Video Concepts using Probabilistic Graphical Models," *2006 IEEE International Conference on Multimedia and Expo*, pp. 301–304, Jul. 2006.

[128] 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 Transactions on Multimedia*, vol. 10, no. 2, pp. 240–251, Feb. 2008.

[129] Y.-X. Xie, X.-D. Luan, S.-Y. Lao, L.-D. Wu, P. Xiao, and Z.-G. Han, "A news video mining method based on statistical analysis and visualization," in *Image and Video Retrieval*. Springer, 2004, pp. 115–122.

[130] J. Oh, J. Lee, S. Kote, and B. Bandi, "Multimedia data mining framework for raw video sequences," in *Mining Multimedia and Complex Data*. Springer, 2003, pp. 18–35.

[131] V. Kulesh, V. A. Petrushin, and I. K. Sethi, "The perseus project: Creating personalized multimedia news portal." in *MDM/KDD*. Citeseer, 2001, pp. 31–37.

[132] M. Roach, J. S. Mason, N. W. Evans, L.-Q. Xu, and F. Stentiford, "Recent trends in video analysis: A taxonomy of video classification problems." in *IMSA*, 2002, pp. 348–353.

[133] A. Ekin, a. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization." *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 12, no. 7, pp. 796–807, Jan. 2003.

[134] a. Divakaran and a. Vetro, "Algorithms and system for segmentation and structure analysis in soccer video," *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001*, pp. 721–724, 2001.

[135] Y.-P. Tan, D. D. Saur, S. R. Kulkami, and P. J. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 10, no. 1, pp. 133–146, 2000.

[136] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," in *ACM multimedia*, vol. 95, 1995, pp. 295–304.

[137] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 1, pp. 52–64, 2005.

[138] M. Roach, J. S. Mason, and M. Pawlewski, "Motion-based classification of cartoons," in *Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on*. IEEE, 2001, pp. 146–149.

[139] W. Zhou, S. Dao, and C.-C. Jay Kuo, "On-line knowledge-and rulebased video classification system for video indexing and dissemination," *Information Systems*, vol. 27, no. 8, pp. 559–586, 2002.

[140] 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*

transactions on pattern analysis and machine intelligence, vol. 28, no. 10, pp. 1678–89, Oct. 2006.

[141] W. Zhou, A. Vellaikal, and C. C. J. Kuo, “Rule-based video classification system for basketball video indexing,” *Proceedings of the 2000 ACM workshops on Multimedia MULTIMEDIA ’00*, pp. 213–216, 2000.

[142] a. Mittal, “Addressing the problems of Bayesian network classification of video using high-dimensional features,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 2, pp. 230–244, Feb. 2004.

[143] Y. Song, G.-j. Qi, X.-s. Hua, L.-r. Dai, and R.-h. Wang, “Video Annotation by Active Learning and Semi-Supervised Ensembling,” *2006 IEEE International Conference on Multimedia and Expo*, pp. 933–936, Jul. 2006.

[144] J. Fan, A. K. Elmagarmid, X. Zhu, W. G. Aref, and L. Wu, “Classview: hierarchical video shot classification, indexing, and accessing,” *Multimedia, IEEE Transactions on*, vol. 6, no. 1, pp. 70–86, 2004.

[145] B. T. Truong and C. Dorai, “Automatic genre identification for contentbased video categorization,” in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 4. IEEE, 2000, pp. 230–233.

[146] X. Yuan, W. Lai, T. Mei, X.-s. Hua, X.-q. Wu, and S. Li, “Automatic Video Genre Categorization using Hierarchical SVM,” *2006 International Conference on Image Processing*, pp. 2905–2908, 2006.

[147] J. Yuan, J. Li, and B. Zhang, “Learning concepts from large scale imbalanced data sets using support cluster machines,” in *Proceedings of the 14th annual ACM international conference on Multimedia*. ACM, 2006, pp. 441–450.

[148] G. Lavee, E. Rivlin, and M. Rudzsky, “Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 39, no. 5, pp. 489–504, Sep. 2009.

[149] X. Yu, C. Xu, H. W. Leong, Q. Tian, Q. Tang, and K. W. Wan, “Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video,” in *Proceedings of the eleventh ACM international conference on Multimedia*. ACM, 2003, pp. 11–20.

[150] P. Chang, M. Han, and Y. Gong, “Extract highlights from baseball game video with hidden markov models,” in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1. IEEE, 2002, pp. I–609.

[151] G. Xu, Y.-f. Ma, H.-j. Zhang, and S.-q. Yang, “An HMM-based framework for video semantic analysis,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 11, pp. 1422–1433, Nov. 2005.

[152] M. Osadchy and D. Keren, “A Rejection-Based Method for Event Detection in Video,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 4, pp. 534–541, Apr. 2004.

[153] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, “Structure analysis of soccer video with domain knowledge and hidden markov models,” *Pattern Recognition Letters*, vol. 25, no. 7, pp. 767–775, 2004.

[154] H. Pan, P. Van Beek, and M. Sezan, “Detection of slow-motion replay segments in sports video for highlights generation,” in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP’01). 2001 IEEE International Conference on*, vol. 3. IEEE, 2001, pp. 1649 – 1652.

[155] G. Hong, B. Fong, and a.C.M. Fong, “An intelligent video categorization engine,” *Kybernetes*, vol. 34 , no. 6, pp. 784–802, 2005.

[156] N. Dimitrova, L. Agnihotri, and G. Wei, “Video classification based on hmm using text and faces,” in *European Conference on Signal Processing*. Citeseer, 2000.

[157] J. Tang, X.-S. Hua, M. Wang, Z. Gu, G.-J. Qi, and X. Wu, “Correlative linear neighborhood propagation for video annotation.” *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics: a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 39, no. 2, pp. 409–16, Apr. 2009.

[158] 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,” *Proceedings of the 14th annual ACM international conference on Multimedia - MULTIMEDIA ’06*, p. 421, 2006.

[159] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, “Correlative multi-label video annotation,” in *Proceedings of the 15 th international conference on Multimedia*. ACM, 2007, pp. 17–26.

[160] S. Feng, R. Manmatha, and V. Lavrenko, “Multiple bernoulli relevance models for image and video annotation,” in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–1002.

[161] M. R. Naphade and J. R. Smith, “On the detection of semantic concepts at trecvid,” in *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004, pp. 660–667.

[162] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, and J. S. Jin, “A unified framework for semantic shot classification in sports video,” *Multimedia, IEEE Transactions on*, vol. 7, no. 6, pp. 1066–1083, 2005.

[163] X. Shen, M. Boutell, J. Luo, and C. Brown, “Multilabel machine learning and its application to semantic scene classification,” in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2003, pp. 188–199.

[164] L. Hollink and M. Worring, “Building a visual ontology for video retrieval,” *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA ’05*, p. 479, 2005.

[165] Y. Wu, B. Tseng, and J. Smith, “Ontology-based multiclassification learning for video concept detection,” *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, pp. 1003–1006, 2004.

[166] J. R. Smith, M. Naphade, and A. Natsev, “Multimedia semantic indexing using model vectors,” in *Multimedia and Expo, 2003. ICME’03. Proceedings. 2003 International Conference on*, vol. 2. IEEE, 2003, pp. II–445.

[167] W. Jiang, S.-F. Chang, and A. C. Loui, “Active context-based concept fusionwith partial user labels,” in *Image Processing, 2006 IEEE International Conference on*. IEEE, 2006, pp. 2917–2920.

[168] J. Fan, H. Luo, and A. K. Elmagarmid, “Concept-oriented indexing of video databases: toward semantic sensitive retrieval and browsing.” *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 13, no. 7, pp. 974–92, Jul. 2004.

[169] R. Yan and M. Naphade, “Semi-supervised cross feature learning for semantic concept detection in videos,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 657–663.

[170] X. Yuan, X.-S. Hua, M. Wang, and X.-Q. Wu, “Manifold-ranking based video concept detection on large database and feature pool,” *Proceedings of the 14th annual ACM international conference on Multimedia - MULTIMEDIA ’06*, p. 623, 2006.

[171] M. Wang, Y. Song, X. Yuan, H.-J. Zhang, X.-S. Hua, and S. Li, “Automatic video annotation by semi-supervised learning with kernel density estimation,” *Proceedings of the 14th annual ACM international conference on Multimedia - MULTIMEDIA ’06*, p. 967, 2006.

[172] M. Wang, X.-s. Hua, J. Tang, and R. Hong, “Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation,” *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 465–476, Apr. 2009.

[173] R. Ewerth and B. Freisleben, “Semi-supervised learning for semantic video retrieval,” *Proceedings of the 6th ACM international conference on Image and video retrieval - CIVR ’07*, pp. 154–161, 2007.

[174] Y. Song, X.-S. Hua, L.-R. Dai, and M. Wang, “Semiautomatic video annotation based on active learning with multiple complementary predictors,” *Proceedings of the 7 th ACM SIGMM international workshop on Multimedia information retrieval - MIR ’05*, p. 97, 2005.

[175] 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," *Proceedings of the 8th ACM international workshop on Multimedia information retrieval - MIR '06*, no. 4, p. 289, 2006.

[176] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semanticbased surveillance video retrieval." *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 16, no. 4, pp. 1168–81, Apr. 2007.

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

[178] Y. Aytar, M. Shah, and J. Luo, "Utilizing semantic word similarity measures for video retrieval," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1 – 8.

[179] L. S. Kennedy, A. P. Natsev, and S.-F. Chang, "Automatic discovery of query-class-dependent models for multimodal search," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 882–891.

[180] R. Yan, J. Yang, and A. G. Hauptmann, "Learning query-class dependent weights in automatic video retrieval," *Proceedings of the 12th annual ACM international conference on Multimedia - MULTIMEDIA '04*, p. 548, 2004.

[181] R. Yan and A. G. Hauptmann, "Probabilistic latent query analysis for combining multiple retrieval sources," *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, p. 324, 2006.

[182] M. G. Christel and R. M. Conescu, "Mining novice user activity with trecvid interactive retrieval tasks," in *Image and Video Retrieval*. Springer, 2006, pp. 21–30.

[183] M. G. Christel, C. Huang, N. Moraveji, and N. Papernick, "Exploiting multiple modalities for interactive video retrieval," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP'04). IEEE International Conference on*, vol. 3. IEEE, 2004, pp. iii–1032.

[184] M. Worring, C. G. Snoek, O. De Rooij, G. Nguyen, and A. Smeulders, "The mediamill semantic video search engine," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1213.

[185] C. Snoek, M. Worring, D. Koelma, and a.W.M. Smeulders, "A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 280–292, Feb. 2007.

[186] R. Lienhart, "A system for effortless content annotation to unfold the semantics in videos," in *Content-based Access of Image and Video Libraries, 2000. Proceedings. IEEE Workshop on*. IEEE, 2000, pp. 45–49.

[187] Y. Wu, Y. Zhuang, and Y. Pan, "Content-based video similarity model," *Proceedings of the eighth ACM international conference on Multimedia - MULTIMEDIA '00*, pp. 465–467, 2000.

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

[189] C. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding Semantics to Detectors for Video Retrieval," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 975–986, Aug. 2007.

[190] 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 *Image and Video Retrieval*. Springer, 2006, pp. 143 – 152.

[191] A. Paul, B.-W. Chen, K. Bharanitharan, and J.-F. Wang, "Video search and indexing with reinforcement agent for interactive multimedia services," ACM Trans. Embed. Comput. Syst., vol. 12, no. 2, pp. 25:1–25:16, Feb. 2013.

[192] Z.-J. Zha, M. Wang, Y.-T. Zheng, Y. Yang, R. Hong, and T.-S. Chua, "Interactive video indexing with statistical active learning," Multimedia, IEEE Transactions on, vol. 14, no. 1, pp. 17–27, 2012