Journal of Computer Science 4 (6): 474-486, 2008
ISSN 1549-3636

A Survey of Content-Based Video Retrieval

[1]P. Geetha and [2] Vasumathi Narayanan
[1]Computer Science and Engineering, Sathyabama University, India
[2]Electronics and Communication Engineering, St.Joseph College of Engineering, India

Abstract: This study surveys current trends/methods in video retrieval. The major themes covered by the study include shot segmentation, key frame extraction, feature extraction, clustering, indexing and video retrieval-by similarity, probabilistic, transformational, refinement and relevance feedback. This work has done in an aim to assist the upcoming researchers in the field of video retrieval, to know about the techniques and methods available for video retrieval.

Key words: Shot Segmentation, key frame, feature extraction, object motion, similarity search, CBVR, SOM, edge detection, threshold, RGB, cut

INTRODUCTION

With the increasing proliferation of digital video contents, efficient techniques for analysis, indexing and retrieval of videos according to their contents have become evermore important. A common first step for most content-based video analysis techniques available is to segment a video into elementary shots, each comprising a continuous in time and space. These elementary shots are composed to form a video sequence during video sorting or editing with either cut transitions or gradual transitions of visual effects such as fades, dissolves and wipes.

Shot boundaries are typically found by, computing an image-based distance between adjacent frames of the video and noting when this distance exceeds a certain threshold. The distance between adjacent frames can be based on statistical properties of pixels [12], compression algorithms [11], or edge differences [13]. The most widely used method is based on histogram differences. If the bin-wise difference between histograms for adjacent frames exceeds a threshold, a shot boundary is assumed. Zhang *et al*. [14] used this method with two thresholds in order to detect gradual transitions.

In recent years research has focused on the use of internal features of images and videos computed in an automated or semi-automated way [29]. Automated analysis calculates statistics, which can be approximately correlated to the content features. This is useful as it provides information without costly human interaction.

The common strategy for automatic indexing had been based on using syntactic features alone. However, due to its complexity of operation, there is a paradigm shift in the research of identifying semantic features [30]. User-friendly Content-Based Retrieval (CBR) systems operating at semantic level would identify motion-features as the key besides other features like color, objects etc., because motion (either of camera motion or shot editing) adds to the meaning of the content. The focus of present motion based systems had been mainly in identifying the principal object and performing retrieval based on cues derived from such motion. With the objective of deriving semantic level indices, it becomes important to deal with the learning tools. The learning phases followed by the classification phase are two common envisioned steps in CBR systems. Rather than the user mapping the features with semantic categories, the task could be shifted to the system to perform learning (or training) with pre-classified samples and determine the patterns in an effective manner.

Shot segmentation: A shot is defined as the consecutive frames from the start to the end of recording in a camera. It shows a continuous action in an image sequence [5]. There are two different types of transitions that can occur between shots, abrupt (discontinuous) also referred as cut, or gradual (continuous) such as fades, dissolves and wipes. The cut boundaries show an abrupt change in image intensity or color, while those of fades or dissolves show gradual changes between frames. These transitions are defined as follows [13]:

- A cut is an instantaneous transition from one scene to the next and it occurs over two frames

Corresponding Author: P. Geetha Department of Computer Science and Engineering, Sathyabama University, India

- A fade is a gradual transition between a scene and a constant image (fade out) or between a constant image and a scene (fade in)
- A dissolve is a gradual transition from one scene to another, in which the first scene fades out and the second fades in
- A wipe occurs as a line moves across the screen, with the new scene appearing behind the line

There are different approaches used to detect the shot in a video and some are outlined here.

Shot boundary detection scheme based on rough-fuzzy set: Han *et al*.[1] describe a technique for video shot boundary detection using rough fuzzy set. The selected low-level features are essential to achieve high accuracy for shot boundary detection. But there are too many features available in the frame or video, such as pixel values of different color channels, statistic features, intensity and color histogram etc. By choosing the most appropriate features to represent a shot or video, the computational burden will be reduced and the efficiency will be improved. For this purpose, the feature optimal choice method based rough sets is introduced in this section. To detect the video shot boundaries, 12 candidate features, classified into 5 types, are usually extracted for common use[2-4]. The first is the RGB space model, the changes of three colors during shot transition can be measured, The 2nd is HSV space model, the component of which can be measured to the changes of hue, saturation and value between adjacent frames. In computation, we compute the mean of every component of each frame in the RGB or HSV model. The histogram features is categorized into two types: gray histogram and color histogram, which are our third and forth types of features. Finally, the statistic feature is considered as the fifth. The mean, variance and skewness of lightness component V in each frame are computed, which include cut, fade and dissolve, as well as zoom, pan and other camera motions and object motions. There are two types of false detections in videos. One results from the existence of irregular camera operations during the gradual transitions. The other is due to a lot of flash effects in a shot. The misses are mainly due to the small content changes between the frame pairs at some shot boundaries. During the experiments, we find that the false detection in the coarse detection will affect on the following feature extraction and shot boundary detection while the missed detection have less effects because the rough-fuzzy calculator weaken the mistakes in coarse detection stage, the dissimilarity function is more fit for varies video. Based on rough-fuzzy set, by which the dissimilarity function for shot boundary detection is obtained. The dissimilarity function is generated for by weighting these important features in term of their proportion in the whole feature. It shows that the proposed methods not only are both similarity and effective but also can decrease data dimensions and preserve the information of original video farthest.

Shot boundary detection in low-pass filtered histogram space: Han and Yoon [5], describes a technique for video shot detection using low pass filtered histogram space. Twin-comparison [6] was developed to find shot boundaries among cuts and fades/dissolves using two thresholds. Gunsel and Tekalp [8] proposed one threshold method using Otsu method to find the threshold automatically. However, this system was presented for detection of cut-type shot boundaries. In model-based method [7, 9], the edit effect showing gradual changes (fades, dissolves, etc.) presents edit invariant property that is used in classifying shot boundaries. In which accentuates edit constancy effects by applying low pass filtering to histogram differences between frames, while suppressing motion effects causing false alarms. Edit constancy effects are rectangular shapes of cut and triangular shapes of fades/dissolves in filtered histogram differences after applying window convolution to original histogram differences. Thus the shot detection method utilizes low-pass filter to reduce false alarms caused by image motion such as camera and objects movements. Because this method uses only color histograms as feature data, the edit constancy effects are usually distorted in real images. New features resulting in edit constancy effects similar to ideal ones will be developed in the future.

The hidden markov model technique: Boreczky and Lynn [10], describe a technique for segmenting video using hidden markov model. It uses three types of features for video segmentation-the standard histogram difference, an audio distance measure and an estimate of object motion between two adjacent frames. The histogram feature measures the distance between adjacent frames based on the distribution of luminance levels. The pixels are distributed into 64 bins based on their luminance. The bin wise difference of the histograms of adjacent frames is called the histogram feature. The audio distance is measured by first converting it to a sequence of cepstal vectors, computed every 20 ms., the likelihood measures are

computed separately over two adjacent intervals and then over their concatenation. The ratio of the two values gives the likelihood ratio for testing the hypothesis that the intervals represent the same sound type. The motion feature detects motion of objects between frames. Motion features are computed using nine motion vectors on nine blocks of the window. Magnitude of the average of the nine vectors and the average of the magnitude of these vectors helps in detecting pans and zooms. The hidden markov model has the following states cut, fade, dissolve, pan, zoom and shot. Each state of the HMM has an associated probability distribution that models the distribution of image, audio and motion features conditioned on that state. The parameters of the HMM are learned through a training phase. Once the parameters are trained, segmenting the video into its shots, camera motions and transitions is performed using the viterbi algorithm. Given the sequence of features, the viterbi algorithm produces the sequence of states most likely to have generated these features.

Shot change detection based on sliding window method: Li and Lee [15], describe a technique for shot change detection based on sliding window method. The Conventional Sliding Window (CSW) method has long been used in video segmentation for its adaptive thresholding technique. A hard cut is detected based on the ratio between the current feature value and its local neighborhood in the sliding window. Yet it has a relatively high rate of false alarm and missed cuts. Combination of the sliding window technique and color histogram differences yields good performance comparing to other methods [16, 17]. An improved sliding window method, which employs multi adaptive thresholds during the three-step processing are global pre filtering, sliding window filtering and Scene activeness investigation, of frame-by-frame discontinuity values. This method uses possibility values produced from different thresholds to measure the degree of possibility of a cuts presence. Detecting a cut based on possibility values, the method is more robust to camera/object motions. Each step produces a likelihood value, which measures the possibility of the presence of a cut. However, one of the purposes is to relax the threshold/parameter selection problem, i.e., to make the intermediate parameters to be valid for a wide range of video programs and to diminish the influence of the final threshold on the overall detection performance. Thus it improves the robustness to camera/object motions by investigating discontinuity values in sliding window more closely and relaxing threshold selection problem in some degree by using likelihood values resulted from different thresholds.

Histogram based detection: Colin *et al*.[18], present a detailed evaluation of a histogram-based shot cut detector. The algorithm was specifically applied to large and diverse digital video collection consisting of eight h of TV broadcast video. It was found that the selection of similarity thresholds for determining shot boundaries in such broadcast video was difficult and necessitates the development of adaptive thresholding in order to address the huge variation of characteristics.

The histogram creation technique used compared successive frames based upon three 64-bit histograms (one of luminance and two of chrominance). These three histograms were then concatenated to form a single N dimensional vector, where N was the total number of bins in all three histograms. The cosine measure was used for comparing histograms of adjacent frames [19]. A low cosine value indicates similarity. Setting the threshold involves a tradeoff between two apparently conflicting point's sufficiently high threshold level to insulate the detector from noise and low enough threshold to make the detector sensitive enough to recognize gradual change.

Shot segmentation by graph partitioning: Cernekova, *et al*. [20], on the detection of gradual transitions such as dissolves and wipes, which are the most difficult to be detected. Unlike the abrupt cuts, the gradual transition spreads across a number of frames. In this method an automated shot boundary detection based on the comparison of more than two consecutive frames is used. Within a temporal window we calculate the mutual information for multiple pairs of frames. This way we create a graph for the video sequence where the frames are nodes and the measures of similarity correspond to the weights of the edges. By finding and disconnecting the weak connections between nodes we separate the graph to sub-graphs ideally corresponding to the shots. The major contribution of the algorithm is the utilization of information from multiple frames within a temporal window, which ensures effective detection of gradual transitions in addition to abrupt cut detection [21, 22].

The method relies on evaluating mutual information within a certain temporal frame window. The video frames are represented as nodes in a graph, whose edge weights signify the pair wise similarities of data points. Clustering is realized by partitioning the graph into disjoint sub-graphs. The method is able to

detect efficiently abrupt cuts and all types of gradual transitions, such as dissolves, fades and wipes with very high accuracy.

Key Frame Extraction: A key-frame is a frame that represents the content of a shot or scene. This content must be the most representative as possible. In the large amount of video data, we first reduce each video to a set of representative key frames (though we enrich our representations with shot-level motion-based descriptors as well). In practice, often the first frame or center frame of a shot is chosen, which causes information loss in case of long shots containing considerable zooming and panning. This is why unsupervised approaches have been suggested that provide multiple key frames per shot [23,24]. Since for online video the structure varies strongly, we use a two-step approach that delivers multiple key frames per shot in an efficient way by following a divide and conquer strategy shot boundary detection-for which reliable standard techniques exist-is used to divide key frame extraction into shot-level sub problems that are solved separately.

Key frames selection using adaptive temporal sampling: For effective video browsing and retrieval, the selected key frames should be able to represent the content of the entire video sequence [25]. There has recently been many works related to the problem of key frame selection and several surveys on the automatic indexing of video data are presented in[25]. An initial approach was proposed in which the first frame of each shot was selected as a key frame [25]. The ordered set of key frames is sometimes referred to as a filmstrip [25]. This is not always sufficient, as there can exist salient changes within a shot due to camera or object motion. To increase the number of frames in a shot Ardizzone and Cascia [25] suggested the number of frames should be related to the length of the shot. If the shot is shorter than one sec., the middle frame was chosen and if the shot is longer, a key frame for each second was chosen. An alternative approach to find the optimal set of key frames such that the frames are maximally distinct and individually carry the most information proposed by Vermaak *et al*. [25]. Han and Yoon [5], describes a technique for key frame extraction using temporal sampling. There are two commonly employed representations of video contents shot-based and object-based. The efficient shot representation and visualization is important for indexing performance and user interface. In the shot-based video indexing system, the representation of the visual contents in a shot is generally achieved by using key frames. For some applications, such as news video indexing, a single frame may be sufficient to represent the contents of the entire frames in a shot. However, if the shot has more complex contents, more key frames are needed. The most common method to select key frames is the temporal sampling method. This method is easy to use and fast. But this method does not provide the successful representation in general because it does not consider the variation of the contents within a shot. To overcome these limitations, we propose an algorithm, called adaptive temporal sampling. We use color histogram differences, which are obtained in a shot change detection process as feature data, movement and dissolve frames. The adaptive temporal sampling method results in the key frames that are sampled at the constant interval. This algorithm select key frames considering the temporal variation of the histogram differences automatically. For more meaningful key frame selection, more information of the frame is needed, such as motion, texture.

Feature Extraction: Various high-level semantic features, concepts such as Indoor/Outdoor, people, speech etc., occur frequently in video databases. To date, techniques for video retrieval are mostly extended directly or indirectly from image retrieval techniques. Examples include first selecting key frames from shots and then extracting image features such as color and texture features from those key frames for indexing and retrieval. The success from such extension, however, is doubtful since the spatio-temporal relationship among video frames is not fully exploited. Motion features that have been used for retrieval include the motion trajectories and motion trails of objects [26], principle components of MPEG motion vectors [27] and temporal texture [28,29]. Motion trajectories and trails are used to describe the spatio-temporal relationship of moving objects across time. The relationship can be indexed as 2D or 3D strings to support spatio-temporal search. Principle components are utilized to summarize the motion information in a sequence as several major modes of motion. Temporal textures are employed to model more complex dynamic motion such as the motion of river, swimming and crowds. An important issue need to be addressed is the decomposition of camera and object motion prior to feature extraction.

Ideally, to fully explore the spatio-temporal relationship in videos, both camera and object motion need to be fully exploited in order to index the foreground and background information separately. Motion segmentation is required especially when the targets of retrieval are objects of interest. In such applications, camera motion is normally canceled by

global motion compensation and foreground objects are segmented by inter-frame subtraction [29]. However, such task is always turned up to be difficult and most importantly, poor segmentation will always lead to poor retrieval results. Although motion decomposition is a preferable step prior to the feature extraction of most videos, it may not be necessary for certain videos. If we imagine a camera as a narrative eye, the movement of eye not only tells us what to be seen but also the different ways of observing events. Typical examples include the sport events that are captured by cameras, which are mounted at the fixed locations of a stand. These camera motions are mostly regular and driven by the pace of games and the type of events that are taken place. For these videos, camera motion is always an essential cue for retrieval. Furthermore, fixed motion patterns can always be observed when camera motions are coupled with the object motion of a particular event.

Clustering: Clustering is always a solution to abbreviate and organize the content of videos, in addition, provides an efficient indexing scheme for video retrieval since similar shots are grouped under the same cluster. The proposed approaches that employ clustering structure for retrieval include [28,30,31]. For instance, Ngo *et al*.[66] proposed a two-level hierarchical clustering structure to organize the content of sport videos. The top level is clustered by color feature while the bottom level is clustered by motion feature. The top level contains various clusters including wide-angle, medium-angle and close-up shots of players from different teams. The shots inside each cluster are partitioned to form sub-clusters in the bottom level according to their motion similarity. Through empirical results, Ngo shown that the cluster-based retrieval, in addition to speed up retrieval time, will generally gives better results especially when a query is located at the boundary of two clusters.

Clustering algorithm based on K-L divergency: Cao *et al*.[32], In this, we first introduce how to get video scene boundaries by shot weave[33], a very successful technique which use unique video feature selection and efficient clustering algorithm. The goal of clustering algorithm is to correct the false scene boundaries resulted from shot weave. So the kernel idea is to find the higher probability of being false boundaries and merge those false scenes to the right position. In this, we introduce a novel approach to check the similarity between the false scene and the correct scene. We import the definition of divergency from information theory[34] to measure the similarity. Then we used a weighted K-L divergence to decompose the computation of distance between scenes into the summarization of the distance among shots. Our clustering algorithms accept audio file and video scene boundary as input. It checks the distance between each single shot scene and its neighbor scene. If the distance is less than pre-defined threshold, mark this scene and its neighbor as the same scene. Repeat this procedure until the entire single shot scenes are checked. Effective shot clustering techniques grouping related shots into scenes are important for content-based browsing and retrieval of video data. However, it is difficult to develop a good shot clustering technique if we use visual information alone. Using audio information attract relatively few attention. We present a novel audio assisted shot clustering technique based on the strict scene definition and divergency measurement. The crux of our approach is the careful analysis of feature space and distribution and the divergency concept used to measure the distance between shots.

Hierarchical clustering method: Clustering is sometimes called unsupervised learning because it classifies objects into subsets only by their actual observations [35]. No presumed classes are needed in the classification process, which is different from pattern recognition methods. There are mainly two types of clustering partition clustering arranges data in separate clusters and hierarchical clustering leads to a hierarchical classification tree.

Hierarchical methods give a more detail description about the relation among data items at different levels. They are more efficient with small data sets and less prone to various types of noise. In the case of large data set, partition methods are more suitable since instead of aiming at optimal partition at each level as hierarchical methods, they find optimal clustering directly at a certain level[36] Therefore, partition clustering methods are more suitable to obtain an abstraction of given data items. Since the abstraction will be done hierarchically in our application, either top-down or bottom-up, the entire abstraction process is actually a generalized hierarchical clustering which adapting partition at each level. This generalized clustering method is very flexible, different feature data and measure metrics, as well as clustering methods, can be applied at different level.

There are many kinds of partition clustering algorithms, among which, the mostly common used one is the iterative algorithm. The basic idea of an iterative clustering algorithm is to start with an initial partition and assign patterns to clusters so as to reduce

certain optimum functions. K-means and ISODATA are two typical iterative clustering algorithms. K-Means algorithm aims at classifying data items into a fixed number of classes starting from an initial partition. ISODATA algorithm is more general than K-means when the desired clustering number is not certain or the initial partition is poor. In ISODATA algorithms, the number of clusters can be adjusted by merging and splitting existing clusters or by removing small clusters.

To realize better and more robust clustering while the amount of data is large, we have also developed a cluster algorithm based on the unsupervised learning neural network [37], Self-Organization Map (SOM), which has been used widely in different areas.

Indexing

Syntactic indexing: Ankush Mittal[38], Some of the prominent content-based retrieval CBR systems are IBMs QBIC[39], ViBE[40], Visualseek[41] and VideoQ[42], Photobook[43] and FourEyes[44] at MIT., Chabot[45] at, MARS[46], Virage[47] and Jacob[48] These systems use syntactic features as the basis for matching and employ either Query-by-example or Query-through-dialog box to interface with the user. Thus, they operate at a lower level of abstraction and therefore, the user needs to be highly versed in the details of the CBR system to take advantage of them.

Popular automatic image indexing systems employ user-composed queries, which are provided through the dialog box. However this method is not convenient as the user needs to know the exact details of the attributes and their implementation as well as details of the search method. However, the operation of such systems is highly technical.

The only alternative to Query through dialog box was thought to be Query by example technique where the user is presented with a number of example images and he indicates the closest. The various features of the chosen image are evaluated and matched against the images in the database. The majority of work in video indexing has focused on the detection of key frames called representative frames or R-frames [49]. Indexing mainly based on minimum bounding regions [50-52]

There are a number of defects with retrieving items with Query by example:

- In contrast to a clearly defined text search, in image search, using query by example and the image can be annotated and interpreted in many ways. For example, a particular user may be interested in a waterfall, another may be interested in mountain and yet another in the sky, although all of them may be present in the same image
- It is reasonable for the user to wonder why do these two images look similar? or what specific parts of these images are contributing to the similarity? Candid [53]. Thus the user is required to know the search structure and other details for efficiently searching the database
- Since there is no matching of exactly defined fields in query by example, it requires a larger similarity threshold as it usually involves many more comparisons than query via the dialog box. The numbers of images retrieved are so many that it makes the whole task tedious and sometimes meaningless.

The survey presented before evinces the fact that most systems operate only at syntactic level and provide low-level descriptors such as color, shape and textures. Some attempted work at semantic level[53,54] confined themselves to data modeling in specific domains. Other works at semantic level exclusively tried to derive semantic properties from low-level properties. This paradigm of deriving semantic indices needs to be explored further.

Semantic indexing: A number of psychological studies and experiments emphasize the need for extracting the semantic information from images and video data. The two important researches in this direction are:

- Demonstrating that higher similarity-ratings are produced by perceptually-relevant semantic features as opposed to the features derived from color histograms on the images [55].
- The performance and the efficiency of searching is generally greatly improved by using semantic cues [56] as compared to when low-level features are employed.

One can find a lot of work, developed lately, employing semantic technique. Shannon *et al*.[53] have analyzed and looked specifically at videotaped presentations in which the camera is focused on the speakers slides projected by an overhead projector. By constraining the domain they are able to define a vocabulary of actions that people perform during a presentation.

Ferman *et al*.[57] and Naphade *et al*.[58] have recently employed probabilistic framework to construct descriptors in terms of location, objects and events. Vasconcelos *et al*.[59] have integrated shot length along

with global motion activity to characterize the video stream with properties such as violence, sex or profanity. An interesting insight that comes out from their work is that there exists a relationship between the degree of action and the structure of visual patterns that constitute a movie.

Hanjalic [60] has given a framework for adaptive extraction of highlights from a sport video based on excitement modeling. The system utilizes the expected variations in a user's excitement by observing the temporal behavior of selected audiovisual low-level features and the editing scheme of a video. Another work by Rasheed *et al.*[61] classifies movies into four broad categories comedies, action, dramas, or horror films. Inspired by cinematic principles, four computable video features average shot length, color variance, motion content and lighting key are combined in a framework to provide a mapping to these four high-level semantic classes.

Recently, many researchers have worked in semantic image classification and natural image database organization into categories like Indoor vs. Outdoor [61] etc., city vs. landscape [62,63] etc., man-made vs. natural [64], sunset vs. forest vs. mountain[65] and so on.

In summary, there is a great need to extract semantic indices for making the CBR system serviceable to the user. Though extracting all such indices might not be possible, there is a great scope for furnishing the semantic indices with a certain well-established structure.

Video retrieval: Video contains multiple types of audio and visual information, which are difficult to extract, combine or trade-off in general video information retrieval.

Similarity measure: Ngo and Pong [66], Video retrieval is still at its preliminary state, despite the fact that videos in addition to image information, consist of extra dimensional information. Three problems that have been attempted are retrieve similar videos [67], locate similar video clips in a video [68,69] retrieve similar shots [70]. In general, similarity measure can be done by matching features either locally or globally. Local matching requires aligning and matching frames (or key frames) across time. For instance, Tan *et al.*[69] employed dynamic programming to align two video sequences of different temporal length. Global matching, on the other hand, measures the similarity between two shots by computing the distance between the two representative features of shots. For retrieving similar videos, besides computing similarity among shots, the temporal order of similar shots between two videos are also taken into account [67].

More sophisticated ways of similarity measure include spatio-temporal matching [71] and nearest feature line matching [72]. The spatio-temporal matching was proposed by Chang *et al.*[31] to measure the similarity of video objects, which are represented as trajectories and trails in the spatial and temporal domains. Recently, Dagtas[71] further presented various trajectory and trail based models for motion-based video retrieval. Their proposed models emphasize both the spatial and temporal scale invariant properties for object motion retrieval. In addition, Zhao [72] described shot similarity measure by employing the concept of nearest feature line. Initially, all frames in a shot are viewed as a curve. Frames that located at the corners are extracted as key frames. Those key frames are connected by lines and form a complete graph.

Video retrieval using visual information: While analyzing the video imagery, we considered the color similarity of images and the presence of faces and text that was readable on the screen. Using the TREC video collection and the automatic known-item queries, we compared our probabilistic image retrieval model against two other vector-based image retrieval algorithms, namely the well-known QBIC image search engine and a Munsell-color histogram based image retrieval algorithm. Both of these two algorithms represent an image as a vector of features and compute the similarity between images based on the Euclidean distance between their representation vectors. The main finding from the results on individual features is probabilistic image retrieval provided the best result for any single metadata type.

It is not too surprising that the results indicate that image retrieval was the single biggest factor in video retrieval for this evaluation. Good image retrieval was the key to good performance in this evaluation, which is consistent with the intuition that video retrieval depends on finding good video images when given queries that include images or video. One somewhat surprising finding was that the speech recognition transcripts played a relatively minimal role in video retrieval for the known-item queries in our task. This may be explained by the fact that discussions among the track organizers and participants prior to the evaluation emphasized the importance of a video retrieval task.

Textual query for video retrieval: Jawahar and Chennupati [73], we present an approach that enables search based on the textual information present in the video. Regions of textual information are identified within the frames of the video. Video is then annotated

with the textual content present in the images. An advanced video retrieval solution could identify the text present in the video, recognize the text and compute the similarity between the query string [74] and pre-indexed textual information present in the video. However, success of this technique depends on two important aspects:

- Quality of the input video
- Availability of an OCR for robustly recognizing the text images

In many practical situations, we find video clips where the resolution of the broadcast is not enough even for a reasonable robust OCR to work on. Moreover for many of the languages, we do not have OCRs available for decoding the textual content in the frames. Since we do not have OCRs available to work effectively on the text in the video data, we use text images to index the videos.

This video retrieval process has two phases online and the offline phase. During the offline phase, broadcast videos in multiple languages are stored in a video database. Cuts and theme changes are identified in each video for the segmentation. These videos are further processed frame by frame for identifying possible text regions as described in the previous section. A selected set of features is extracted from these text regions and stored in a feature database.

During the online phase, a search query is entered through a graphical user interface. This query string is rendered as an image and the corresponding set of features is extracted. These features are same as those employed in the offline process. A matching algorithm then computes the degree of similarity between the features of the search query and those present in the feature database. The results are then ranked based on the similarity measure. Word form variations are handled using a partial matching method, based on Dynamic Time Warping (DTW)[75]. We have also used scrolling text as a source of information. Scrolling text, unlike the overlaid text captions, is in motion. Initially the scrolling speed of text is computed. Then a long text image of the scrolling content is generated and connected component analysis is performed to find the words. This text image is used for indexing the corresponding video clip.

We also argued that this approach has wider applicability compared to other techniques including the one based on OCRs. It will focus on improving the accuracy as well as the speed of techniques used here. Accuracy can be improved by using better techniques as well as using a large feature set, which discriminates words better from each other. Optimizing our implementation of the dynamic time warping algorithm as well as looking at related computational techniques to minimize the number of possible matches could improve speed. We are also exploring an efficient indexing scheme.

Refinemnt and relevance feedback: Several Relevance Feedback (RF) algorithms have been proposed over the last few years. The idea behind most RF-models is that the distance between images labeled as relevant and other similar images in the database should be minimal. The key factor here is that the human visual system does not follow any mathematic metric when looking for similarity in visual content and distances used in image retrieval systems are well-defined metrics in a feature space.

Pseudo-relevance feedback: Yan *et al.*[76] Nearest neighbor search is the most straightforward approach to find matching images. It contains the implicit assumption that for each feature the class posterior probabilities (distributions) are approximately constant for matching and non-matching images. However nearest neighbor search suffers from a lack of adaptability. So we have shown that some well-established classification algorithms can yield better generalization performance than nearest neighbor type algorithms. Following this direction, we used a classification-based Pseudo-Relevance Feedback (PRF) approach. The basic idea for our approach is to augment the retrieval performance by incorporating classification algorithms via PRF, with the choice of training examples based on the initial retrieval results.

Standard PRF methods, which originated from the filed of text information retrieval, view the top-ranked documents as positive examples. However, due to the limited accuracy of current video retrieval systems, even the very top-ranked results are not always the relevant, correct answers that meet the users information need, so they cannot be used as reliable positive examples for relevance feedback. However, we discovered that it is quite appropriate to make use of the lowest ranked documents in the collection, because these documents are very likely to be negative examples. While the normalization and combination of evidence is novel, this emphasizes the successful use of negative pseudo-relevance feedback to improve image retrieval performance. While the results are still far from satisfactory, PRF shows great promise for multimedia retrieval in very noisy data. One of the future directions of this approach is to study the effect

of different classification algorithms and explore better combination strategies than a simple linear combination of the individual agents.

Negative pseudo relevance feedback: Yan *et al*.[77]. In this, we present a novel automatic retrieval technique for multimedia data called negative pseudo-relevance feedback NPRF [78,79]. It attempts to learn an adaptive similarity space by automatically feeding back the training data, which are identified based on a generic similarity metric.

In the task of content-based video retrieval, a query typically consists of a text description plus audio, images or video. This query is posed against a video collection. The job of the video retrieval algorithm is to retrieve a set of relevant video shots from a given data collection [80]. In that, the positive examples are the query examples and the negative examples are sampled from the strongest negative examples. Due to the computational issues, the feedback process repeats for single iteration. Although the NPRF approach can be applied to various retrieval tasks such as text retrieval and audio retrieval, in this initial work we have mainly applied it in image retrieval feedback.

In that, our analysis is based on a statistical model of average precision. We also present several score fusion paradigms by transforming different types of similarity metrics into probabilistic outputs. Fusion of different retrieval algorithms is an effective way to address the false positive problem in NPRF. As mentioned before, combining base score and NPRF score might offer a reasonable trade-off. More interestingly, it has been found that combination of NPRF score and retrieval scores from different modalities can recover most of the performance hurt since most false positives can be filtered out by additional information. Also, these combinations can reduce the prediction variance and offer more stable results. Negative Pseudo-Relevance Feedback (NPRF)), to improve the performance of content-based video retrieval. Different from the canonical relevance feedback technique, our approach does not require users to provide judgment within a retrieval process. In this work, the task of video retrieval is framed as a concept classification problem. Specifically, the positive examples are provided by users' query and negative examples can be obtained from the worst matching examples identified based on a generic similarity metric. A margin-based learning algorithm, support vector machine SVM, is used to learn the updated similarity scores. Theoretical analysis shows that the benefit of the NPRF approach derives from the ability to adapt similarity score across queries and thus separate the means of the negative/positive score distributions. Since some extreme outliers might be misclassified as false positives, we suggest that smoothing with either the initial similarity score or the score from different modalities can safeguard against these egregious errors.

CONCLUSION

We probably covered only a small part of all existing video retrieval systems but we can still draw some conclusions from this survey. It is difficult to evaluate how successful content-based video retrieval systems are, in terms of effectiveness, efficiency and flexibility. Many articles about systems give figure about precision and recall. Most of them are good, but hard to verify. However, there are also considerations intrinsic to retrieval systems. In future, it is planned to find the concurrency in various features surveyed in this study.

REFERENCES

1. Bing Han, Xinbo Gao, Hongbing Ji, 2005. A shot boundary detection method for news video based on rough-fuzzy sets. Int. J. Inform. Technol., 11: 101-111. www.icis.ntu.edu.sg/scs-ijit/117/117_11.pdf
2. Gao, X. and X. Tang, 2002. Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing. IEEE Trans. Circuits Syst. Video Technol., 12: 765-776. Doi: 10.1109/TCSVT.2002.800510
3. Gao, X. and X. Tang, 2000. Automatic parsing of news video based on cluster analysis. In: Proceedings of 2000 Asia Pacific Conference on Multimedia Technology and Applications, Kaohsiung, Taiwai, China, Dec. 17-19, pp: 17-19. https://dspace.lib.cuhk.edu.hk/handle/2006/5923
4. Han Bing, Gao Xin-bo, Ji Hong-bing, 2003. An efficient algorithm of gradual transition for shot boundary segmentation. 3rd International Symposium on Multispectral Image Processing and Pattern recognition (MIPPR'03), Beijing, 9: 956-961. http://see.xidian.edu.cn/faculty/hbji/eng.htm
5. Seung-Hoon Han, Kuk-Jin Yoon, and In So Kweon, 2000. A new technique for shot detection and key frames selection in histogram space. In: 12th Workshop on Image Processing and Image Understanding, pp: 475-479. http://rcv.kaist.ac.kr/publication/file/domestic_conference/18_SeungHoonHan_IPIU2000.pdf

6. Zhang, H.J., A. Kankanhalli and S.W. Smoliar, 1993. Automatic partitioning of full motion video. ACM Multimedia Syst., 1: 10-28. Doi: 10. 1007/BF01210504, http://portal.acm.org/citation .cfm?id=173856.173858
7. Hampapur, A., R. Jain and T.E. Weymouth, 1995. Production model based digital video segmentation. Multimedia Tools Appl., 1: 9-46. Doi: 10.1007/BF01261224
8. Bilge, G. and A.M. Tekalp, 1998. Content based video abstraction. In: Proceedings of International Conference on Image Proceeding, Oct. 4-7, 3: 128-132, Doi: 10.1109/ICIP.1998.727150
9. Song, S.M., T.H. Kwon, W.M. Kim, H. Kim and B.D. Rhee, 1998. On detection of gradual scene changes for parsing of video data. Proc. SPIE Storage Retrieval Image Video Database, 3312: 404-413. DOI:10.1117/12.298449
10. John, S., Boreczky and D. Lynn, 1998. A hidden markov model framework for video segmentation using audio an image features. In: Proceedings of IEEE International conference on Acoustics, Speech and Signal Processing, May 12-15, 6: 3741-3744. Doi: 10.1109/ICASSP.1998.679697
11. Arman, F., A. Hsu and M.Y. Chiu, 1994. Image processing on encoded video sequences. ACM Multimedia Syst., 1: 211-219. Doi: 10.1007/BF01268945
12. Jain, R., 1988. Dynamic vision. In: 9th International Conference on Pattern Recognition, Nov. 14-17, 1: 226-235. Doi: 10.1109/ICPR.1988.28212
13. Zabih, R., J. Miller and K. Mai, 1995. A feature-based algorithm for detecting and classifying scene breaks. In: Proceedings of the 3rd ACM International Conference on Multimedia, pp: 189-200. http://doi.acm.org/10.1145/217279.215266
14. Zhang, H.J., A. Kankanhalli and S.W. Smoliar, 2001. Automatic partitioning of full-motion video. Readings Multimedia Comput. Network.ing, 321-339. http://portal.acm.org/citation.cfm?id=383905
15. Shan Li, Moon-Chuen Lee, 2005. An improved sliding window method for shot change detection. Proceeding of the 7th IASTED International Conference Signal and Image Processing, Aug. 15-17, IIonolulu, IIawaii, USA, pp: 464-468. http://appsrv.cse.cuhk.edu.hk/~sli/publication/sip05.pdf
16. Gargi, U., R. Kasturi and S.H. Strayer, 2000. Performance characterization of video-shot-change detection methods. IEEE Trans. Circuits Syst. Video Technol., 10: 1-13, 2000. Doi: 10.1109/76.825852
17. Yeo, B.L. and B. Liu, 1995. Rapid scene analysis on compressed video. IEEE Trans. Circuits Syst. Video Technol., 5: 533-544, 1995. Doi: 10.1109/76.475896
18. O'Toole, C., A. Smeaton, N. Murphy and S. Marlow, 1999. Evaluation of automatic shot boundary detection on a large video suite. In: 2nd U.K. Conference Image Retrieval: The Challenge of Image Retrieval, Feb. 25-26, Newcastle, U.K., pp: 1-12. http://doras.dcu.ie/346/
19. Cabedo Sushil, Xavier Ubiergo and K. Bhattacharjee, 1998. Shot detection tools in digital video. In: Proceedings of Non-linear Model Based Image Analysis 1998, Springer Verlag, Glasgow, pp. 121-126, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.2365
20. Cernekova, Z., N. Nikolaidis and I. Pitas, 2006. Temporal video segmentation by graph partitioning. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, May 14-19, 2: 209-212, Doi: 10.1109/ICASSP.2006.1660316
21. Cernekova, Z., C. Nikou and I. Pitas, "Shot detection in video sequences using entropy-based metrics," In: Proceedings of IEEE International Conference on Image Processing, vol: 3, pp: 421-424, 24 – 28 June 2002. Doi: 10.1109/ICIP.2002.1038995
22. Sarah V Porter, 2004. Video segmentation and indexing using motion estimation. Ph.D Thesis, University of Bristol, http://www.cs.bris.ac.uk/Publications/pub_master.jsp?id=2000202
23. Hauptmann, A.G., R. Jin and T.D. Ng, 2003. Video retrieval using speech and image information. In: SPIE Proceedings Series on Storage and Retrieval for Multimedia Databases, Jan. 20-24, 5021: 148-159. http://cat.inist.fr/?aModele=afficheN&cpsidt=15204368
24. Hafner, J., H.S. Sawhney, W. Equitz, M. Flickner and W. Niblack, 1995. Efficient color histogram indexing for quadratic form distance. IEEE Trans. Pattern Anal. Machine Intel., 17: 729-736. Doi: 10.1109/34.391417
25. Sato, T., T. Kanade, E.K Hughes and M.A. Smith, 1998. Video OCR for digital news archive. In: Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Databases, Jan. 3rd, pp: 52-60, Doi: http://doi.ieeecomputersociety.org/10.1109/CAIVD.1998.646033

26. C. Colombo, A. Del Bimbo and P. Pala, 1999. Semantics in visual information retrieval. IEEE Multimedia, 6: 38-53. Doi: 10.1109/93.790610
27. Smeulders, A.W.M., M. Worring, S. Santini, A. Gupta and R. Jain, 2000. Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Machine Intel., 22: 1349-1380. Doi: 10.1109/34.895972
28. Chong-Wah Ngo, Ting-Chuen Pong and Hong-Jiang Zhang, 2003. Motion analysis and segmentation through spatio-temporal slices processing. IEEE Trans. Image Process., 12: 341-355. Doi: 10.1109/TIP.2003.809020
29. Fablet, R., P. Bouthemy and P. Perez, 2000. Statistical motion-based video indexing and retrieval. In: Proceedings of International Conference on Content-based Multimedia Information Access, pp: 602-619. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.25.7728
30. Fan, J., A.K. Elmagarmid, X. Zhu, W.G. Aref and L. Wu, 2004. Classview: Hierarchical video shot classification, indexing and accessing. IEEE Trans. Multimedia, 6: 70-86. Doi: 10.1109/TMM.2003.819583
31. Hwanjo, Yu, Jiawei Han and Kevin Chen-Chuan Chang, 2004. PEBL: Web page classification without negative examples. IEEE Trans. Knowledge Data Eng., 16: 70-81. Doi: 10.1109/TKDE.2004. 1264823
32. Cao, Y., W. Tavanapong, K. Kim and J. Oh, 2003. Audio assisted scene segmentation for story browsing. Proc. Int. Conf. Image Video Retrieval, 2728: 446-455. Doi: 10.1007/3-540-45113-7_44
33. Tavanapong, W. and Junyu Zhou, 2004. Shot clustering techniques for story browsing. IEEE Trans. Multimedia, 6: 517-527. Doi: 10.1109/TMM.2004.830810
34. Solomon Kullback, 1997. Information Theory and Statistics. 1st Edn., Courier Dover Publications, USA., ISBN 0486696847, 978-0486696843
35. Zhong, D., H. Zhang and S.F. Chang, 1996. Clustering methods for video browsing and annotation. In: Procedding of SPIE Storage and Retrieval for still Image and Video Database IV, 2670: 239-246. http://adsabs.harvard.edu/abs/1996SPIE.2670..239Z
36. Anil K. Jain, Richard C. Dubes, "Algorithms for clustering data", Published by Prentice Hall College Div, March 1988, ISBN: 013022278X, 978-0130222787
37. Zhang, H.J. and D. Zhong, "A scheme for visual feature based image indexing". In: Proceeding of SPIE Conference on Storage and Retrieval for Image and Video Databases III, vol. 2420, pp: 36-46, February 1995. http://citeseerx.ist.psu.edu/showciting;jsessionid=EA38C101B5FA1D8EAA9F9F268750EE28?cid=1529801.
38. Ankush, M., "An overview of multimedia content-based retrieval strategies". Informatica, vol: 30, No: 3, pp: 347–356, October 2006, http://www.informatica.si/PDF/30-3/09_Mittal_An%20Overview%20of%20Multimedia%20Content.pdf.
39. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Qian Huang, Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P., "Query by image and video content: The QBIC system," IEEE Computer, vol. 28, no. 9, pp: 23-32, Sep 1995. Doi: 10.1109/2.410146.
40. Chen, J.Y., C. Taskiran, E.J. Delp and C.A. Bouman, "ViBE: A new paradigm for video database browsing and search". In Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries, pp: 96-100, 21 June 1998. 10.1109/IVL.1998.694510.
41. J. R. Smith and S.F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System," In proceedings of fourth ACM International Conference on Multimedia, Boston, MA, pp: 87-98, November 1996. http://portal.acm.org/citation.cfm?id=244130.244151.
42. Chang, S.F., W. Chen, H.J. Meng, H. Sundaram and D. Zhong, "VideoQ: an automated content based video search system using visual cues". In Proceedings of 5th ACM International conference on Multimedia, pp: 313-324, November 9-13, 1997, http://portal.acm.org/citation.cfm?id=266382.
43. Pentland, A., R.W. Picard and S. Sclaroff, 1996. Photobook: Content-based manipulation of image databases. Proc. IJCV, 18: 233-254. Doi: 10.1007/BF00123143
44. Picard, R.W. and T.P. Minka, 1995. Vision texture for annotation. Proc. ACM J. Multimedia Syst., 3: 3-14. http://portal.acm.org/citation.cfm?id=203180
45. Ogle, V.E. and M. Stonebraker, 1995. Chabot: Retrieval from a relational database of images. Proc. IEEE Comput., 28: 40-48. Doi: 10.1109/2.410150
46. Mehrotra, S., Yong Rui, M. Ortega-Binderberger and T.S. Huang, 1997. Supporting content-based queries over images in MARS. In: Proceedings of IEEE International Conference on Multimedia Computing and Systems, Jun 3-6, Ottawa, Ont., Canada, pp: 632-633. Doi: 10.1109/MMCS.1997.609791

47. Bach, J.R., C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain and C. Shu, 1996. Virage image search engine: An open framework for image image management. Proc. SPIE Conf. Storage Retrieval Still Image Video Databases, 2670: 76-87. Doi: 10.1117/12.234785
48. Ardizzone, E., La Cascia, M., "Automatic Video Database Indexing and Retrieval," Journal on Multimedia Tools and Applications, vol: 4, no. 1, pp. 29-56, January 1997, Doi 10.1023/A: 1009630331620
49. Ramin Zabih, Justin Miller, Kevin Mai, "Video Browsing Using Edges and Motion," in proceedings on IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp: 439-446, 18-20 June 1996, Doi: 10.1109/CVPR.1996.517109
50. Shearer, K., Bunke, H., Venkatesh, S., "Video indexing and similarity retrieval by largest common sub graph detection using decision trees", In proceedings of Pattern Recognition, vol: 34, No. 5, pp. 1075-1091, May 2001. Doi: 10.1016/S0031-3203(00)00048-0
51. Guttman, A., "R-trees: a dynamic index structure for spatial searching",in proceedings of ACM SIGMOD International Conference on Management of Data, vol:1, no: 2, pp:47-57, 1984, http://doi.acm.org/10.1145/602259.602266
52 White, D.A.; Jain, R., "Similarity indexing with the SS-tree", in Proceedings of the Twelfth International Conference on Data Engineering, pp: 516 - 523 , 26 Feb-March 1 1996, Doi: 10.1109/ICDE.1996.492202
53. Ju, S.X., Black, M.J., Minnerman, S. and Kimber D., " Analysis of Gesture and Action in Technical Talks for Video Indexing" In proceedings of IEEE computer society conference on Computer Vision and Pattern Recognition, pp: 595-601, 17-19 Jun 1997. http://portal.acm.org/citation.cfm?id=794189.794452.
54. Krishnapuram, R., S. Medasani, S.H. Jung, Y. Choi and R. Balasubramaniam, "Content-based image retrieval based on a fuzzy approach." IEEE Trans. Knowledge and Data engineering, vol: 16, no: 10, pp: 1185-1199, October 2004, Doi: 10.1109/TKDE.2004.53.
55. B. E. Rogowitz, T. Frese, J. Smith, C. A. Bouman, and E. Kalin, "Percpetual image similarity experiments," in IS&T/SPIE Conf. on Human Vision and Electronic Imaging III, 1998. http://serv1.ist.psu.edu:8080/viewdoc/summary?doi=10.1.1.89.5054.
56. T.V. Papathomas, T.E. Conway, I.J. Cox, J. Ghosn, M.L. Miller, T.P. Minka, P.N. Yianilos, "Psychophysical studies of the performance of an image database retrieval system" , in Proc. SPIE Conf. on Human Vision and Electronic Imaging , vol: 3299 pp. 591-602, July 1998, http://citeseerx.ist.psu.edu/showciting;jsessionid=2C93B1E50B192C14F6ABBF5335C8B100?cid=61126
57. Mufit Ferman, A. Murat Tekalp, A., "Probabilistic analysis and extraction of video content", In Proceedings of International Conference on Image Processing, Vol: 2, pp: 91-95, 1999, Doi: 10.1109/ICIP.1999.822861.
58. Naphade, M.R., Kristjansson, T., Frey, B., Huang, T.S., "Probabilistic multimedia objects (MULTIJECTS): a novel approach to video indexing and retrieval in multimedia systems", in Proceedings of International Conference on Image Processing, vol.3, pp: 536-540, October 1998. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.8930
59. Vasconcelos, N. Lippman, A., "Towards semantically meaningful feature spaces for the characterization of video content ", in Proceedings of International Conference on Image Processing, Vol: 1, pp: 25-28, 26-29 October 1997, Doi: 10.1109/ICIP.1997.647375.
60. Hanjalic, A., 2005. "Adaptive extraction of highlights from a sport video based on excitement modeling." IEEE Trans. Multimedia, 7: 1114-1122, Doi: 10.1109/TMM.2005.858397
61. Rasheed, Z. Sheikh, Y. Shah, M., "On the use of computable features for film classification", IEEE Transactions on Circuits and Systems for Video Technology, Vol: 15, no: 1, pp: 52- 64, Jan. 2005, Doi: 10.1109/TCSVT.2004.839993(410) 1.
62. Gorkani, M.M.; Picard, R.W., "Texture orientation for sorting photos at a glance", in Proceedings of the 12th IAPR International Conference on Computer Vision & Image Processing., Pattern Recognition, Vol. 1, pp: 459 – 464, 9-13 October 1994, Doi: 10.1109/ICPR.1994.576325.
63. David A. Forsyth , Jitendra Malik , Margaret M. Fleck , Hayit Greenspan , Thomas K. Leung , Serge Belongie , Chad Carson , Chris Bregler, "Finding Pictures of Objects in Large Collections of Images," Proceedings of the International Workshop on Object Representation in Computer Vision II, p.335-360, April 13-14, 1996. http://portal.acm.org/citation.cfm?id=648641.

64. Torralba, A.B. and A. Oliva, “Semantic organization of scenes using discriminant structural templates”. IEEE International Conference on Computer Vision, vol. 2, pp. 1253-1258, 1999. Doi: 10.1109/ICCV.1999.790424.
65. Aditya Vailaya, Anil Jain, "Reject Option for VQ-Based Bayesian Classification," in proceedings of 15th International Conference on Pattern Recognition , Vol. 2, pp: 48-51, 2000, Doi:10. 1109/ICPR.2000.906016
66. C.-W. Ngo, H.-J. Zhang, and T.-C. Pong, “Recent advances in content based video analysis,” International Journal of Image and Graphics, vol. 1, no. 3, pp. 445–468, 2001. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.9593.
67. Yi Wu, Y. Zhuang, and Y. Pan:, “Content-Based Video Similarity Model,” In Proc. of the 8th ACM Int. Multimedia Conf. on Multimedia, USA, pp. 465-467, 2000, http://portal.acm.org/citation.cfm?id=376380.
68. Anil Jain, Aditya Vailaya, Wei Xiong., "Query by video clip," In Proceedings of Fourteenth International Conference on Pattern Recognition, vol.1. pp: 909-911, 16-20 Aug 1998, Doi: 10.1109/ICPR.1998.711299.
69. Tan, Y.P. [Yap-Peng], Kulkarni, S.R.[Sanjeev R.], Ramadge, P.J.[Peter J.], "A Framework for Measuring Video Similarity and Its Application to Video Query By Example", in proc. IEEE International conference on image processing, vol.2, pp: 106-110, 1999, Doi: 10.1109/ICIP.1999.822864
70. Zhong, D. [Di], Chang, S.F.[Shih-Fu], "An integrated approach for content-based video object segmentation and retrieval", IEEE transactions on circuits and video technology, vol.9, no. 8, pp. 1259-1268, December 1999, Doi: 10.1109/76.809160.
71. Li Zhao, Wei Qi, Stan Z. Li, S.Q.Yang, H.J. Zhang, "Key-frame Extraction and Shot Retrieval Using Nearest Feature Line (NFL)", in Proceedings of ACM international workshop on Multimedia , pp: 217-220, November 2000, http://portal.acm.org/citation.cfm?id=357942
72. Dagntas, S., Al-Khatib, W., Ghafoor, A., Kashyap, R.L., "Models for Motion-Based Video Indexing and Retrieval", IEEE Transactions on image processing, vol.9, No. 1, pp. 88-101, January 2000, Doi: 10.1109/83.817601
73. C. V. Jawahar, BalaKrishna Chennupati, Balamanohar Paluri and Nataraj Jammalamadaka, “Video Retrieval Based on Textual Queries”, in Proceedings of the Thirteenth International Conference on Advanced Computing and Communications, Coimbatore, December 2005. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.9857.
74. Meshesha, M., C.V.Jawahar and A. Balasubramanian, “Searching in Document Images”, in Proc. of the 4th Indian. Conference on Computer Vision, Graphics and Image. Processing (ICVGIP) pp: 622–627, 2004. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.8420.
75. Rath, T.M. and R. Manmatha, “Word image matching using dynamic time warping”, In: Conference on Computer Vision and Pattern Recognition, pp: 521-527, 2003. ISBN: 0-7695-1900-8.
76. Yan, R., Jin, R., Hauptmann, A., “Multimedia search with pseudo-relevance feedback”. In Proc. Int. Conf. on Image and Video Retrieval, pp: 238-247, 2003, http://cat.inist.fr/?aModele=afficheN&cpsidt=15567430.
77. R. Yan, A. Hauptmann, and R. Jin, "Negative Pseudo-Relevance Feedback in Content-based Video Retrieval", in proceedings of 11th ACM International Conference on Multimedia, pp. 343 - 346, 2003. http://portal.acm.org/citation.cfm?id=957013.957087.
78. Yan, R., Jin, R., Hauptmann, A., "Multimedia Search with Pseudo-Relevance Feedback," AAAI Spring Symposium Series on Intelligent Multimedia Knowledge Management, Palo Alto, CA, March 24-26, 2003. http://www.informedia.cs.cmu.edu/pubs/abstract.asp?id=162.
79. Yan, R., A. Hauptmann and R. Jin, “Pseudo-Relevance Feedback for Multimedia Retrieval”, In Kluwer International series in video computing, 2003.www.visionbib.com/bibliography/applicat803.html.
80. Wu, L., C. Faloutsos, K.P. Sycara and T.R. Payne, “Multimedia queries by example and relevance feedback”, in proc. IEEE Data Engineering Bulletin 24(3), 2001, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.7.8289.