DOI:10.24996/ijs.2018.59.2A.12

ISSN: 0067-2904

Semantic Based Video Retrieval System: Survey

Matheel E. Abdulmunem[1], Eman Hato[2*]
[1]Department of Computer Science, Technology University, Baghdad, Iraq.
[2]Department of Computer Science, Mustansiriyah University, Baghdad, Iraq.

Abstract

In this review paper a number of studies and researches are surveyed, in order to assist the upcoming researchers, to know about the techniques available in the field of semantic based video retrieval. The video retrieval system is used for finding the users' desired video among a huge number of available videos on the Internet or database. This paper gives a general discussion on the overall process of the semantic video retrieval phases. In addition to its present a generic review of techniques that has been proposed to solve the semantic gap as the major scientific problem in semantic based video retrieval. The semantic gap is formed because of the difference between the low level features that are extracted from video content and user's perceptions of these features in a real world. The transformation of low level features of the video content into meaningful semantic concepts is a research topic that has received considerable attention in recent years.

Keywords: feature extraction, semantic gap, user query, video annotation, video mining.

نظام استرجاع الفيديو على أساس المعنى الدلالي: مراجعة لبحوث سابقة

مثيل عماد الدين عبد المنعم[1]، إيمان هاتو هاشم[2]*
[1]قسم علوم الحاسوب، الجامعة التكنولوجية، بغداد، العراق.
[2]قسم علوم الحاسوب، الجامعة المستنصرية، بغداد، العراق.

الخلاصة

تم في هذا البحث استعراض عدد من الدراسات والبحوث، من أجل مساعدة الباحثين القادمين، للتعرف على التقنيات والأساليب المتاحة في مجال استرجاع الفيديو على أساس المعنى الدلالي. يستخدم نظام استرجاع الفيديو للعثور على عناصر الفيديو المطلوبة من قبل المستخدمين (مشاهد، لقطات، أوصور) بين عدد كبير من أشرطة الفيديو المتاحة على شبكة الإنترنت أو قاعدة البيانات. يقدم هذا البحث فكرة عامة عن مراحل نظام استرجاع الفيديو على أساس المعنى الدلالي, بالاضافة الى مراجعة عامة للتقنيات والخورزميات التي تم اقتراحها لحل الفجوة الدلالية باعتبارها المشكلة العلمية الرئيسية لاسترجاع الفيديو. تتكون الفجوة الدلالية بسبب الاختلاف بين الميزات ذات المستوى المنخفض التي يمكن استخرجها من المعلومات المخزونة في الفديو وتفسير هذه الميزات من قبل المستخدم في العالم الواقعي. إن تحويل الميزات ذات المستوى المنخفض إلى مفاهيم ذات معنى دلالي هو موضوع بحثي مهم حظي باهتمام كبيرمن قبل الباحثين في السنوات الأخيرة.

1. Introduction

Today, through the rapid growth of information technology and multimedia strategies evolution, the amount of multimedia data accessible is increasing dramatically. One of the important forms of

*Email: emanhato@yahoo.com

multimedia data is video; it consists of different types of data like text, image, sound, and metadata. Video is heavily consumed in major applications such as surveillance, entertainment, medicine, education and sports. Searching for the elements required in these large data on the Internet can be considered an important challenge. Therefore; numerous video retrieval systems have been introduced for this intent.

There are two approaches in video retrieval: text based framework and content based frameworks [1]. In text based framework, text descriptor is used to annotate video manually. In this method, the search depends on the metadata associated with each video such as tags, title, description and keyword [2]. The drawback of this approach needs humanitarian work to manual annotation. The solution to the above drawback is content based framework. Content based framework enables the system to retrieve a video clip from a collection of videos based on the visual content which are extracted fully automatically such as color, texture, shapes, rather than on attributes irrelevant to the content [3]. This visual content bears little or no semantic content of the video.

The video carries a meaningful message that can be distinguished between one video and another, in other words the video is a sequence of events and stories, but not necessarily a sequence of color histograms or edge maps. Semantic representation is therefore the basis for building an efficient video data index. Mostly users do not care how a video looks but what the video clip about. For instance, instead of posing the query "brown object", users are most likely to ask for "mountain". Thus, the new trend of video retrieval systems aims to retrieve video clips that rely on semantic content, which is referred to as semantic based video retrieval [4, 5]. A human perception of video meaning is instantaneous unlike a computer which is far from truth. This discrepancy is indicated to as the semantic gap [6, 7]. The core idea is bridged the gap between the low level features that are extracted from video content and the interpretation of these features by humans. The transformation of low level features of the video content into semantic concepts is a research topic that has received considerable attention in recent years. To improve semantic based video retrieval techniques, it is requisite to develop a formal description of semantic video content and setup indexes that are consistent with the human perspective, and to handle, as much information as possible in a video.

The video information can be categorized into three classes: Low level features information which represent visually such as pixels, colors, texture and shape; aurally such as loudness, pitches, and frequencies, and textually such as alphabets and symbols, Syntactic information which describe the video contents, such as objects, spatial temporal position of object and spatial temporal relations between object, and Semantic information which describe what happens in a video according to the user's perception. The semantic information used to determine events of the video are: the spatial information provided through the video frame like location and objects provided in the video frame, and the temporal information provided by a series of video frames in a timely manner like actions and movements of the object displayed in a frames series [8]. Various features of different modalities (visual modality, auditory modality, and textual modality) are made use of to extract the semantic meaning of the video in order to bridge the gap between the low level features and high level semantic concepts. This paper is aim to organize the methods focused in semantic based video retrieval. The remainder of this paper is structured as follows: in the next section video terminology is defined. In section 3 the structure of semantic based video retrieval system is discussed. The general overview on contributions that have been achieved in the domain of semantic based video retrieval is presented in section 4. Lastly, the concluding remarks are given in section 5.

2. Video Terminology

Before entering into the discussion, it would be useful first to highlight the hierarchy levels of the videos [9] as illustrated in the Figure -1.

- Video: indicates a multimedia source that combines a series of images to form an animated picture, audio component that correspond to images that are displayed on the screen and textual data that is rendered by linguistic form.
- Shot: it is a successive frames captured through a single camera without significant changes in visual content. It is a brick of video streams. Shot boundary detection is a partition of the video tracks to shots to enable various processing operations on video.
- Key frame: due to the similarity among successive frames, one or more key frames are selected from a single shot based on the complication of shot content. The selected key frames represent the salient visual contents.

- Scene: is a group of semantically relevant and temporally adjacent shots that describe a high-level concept in a same location and continuous time. Physical boundaries characterize the shot, while semantic boundaries characterize the scenes.
- Video group: is a mediator entity between the physical shots and semantic scenes. The video group consumes two types of shots: temporally where shots are related in temporal series and spatially where shots are similar in visual features.

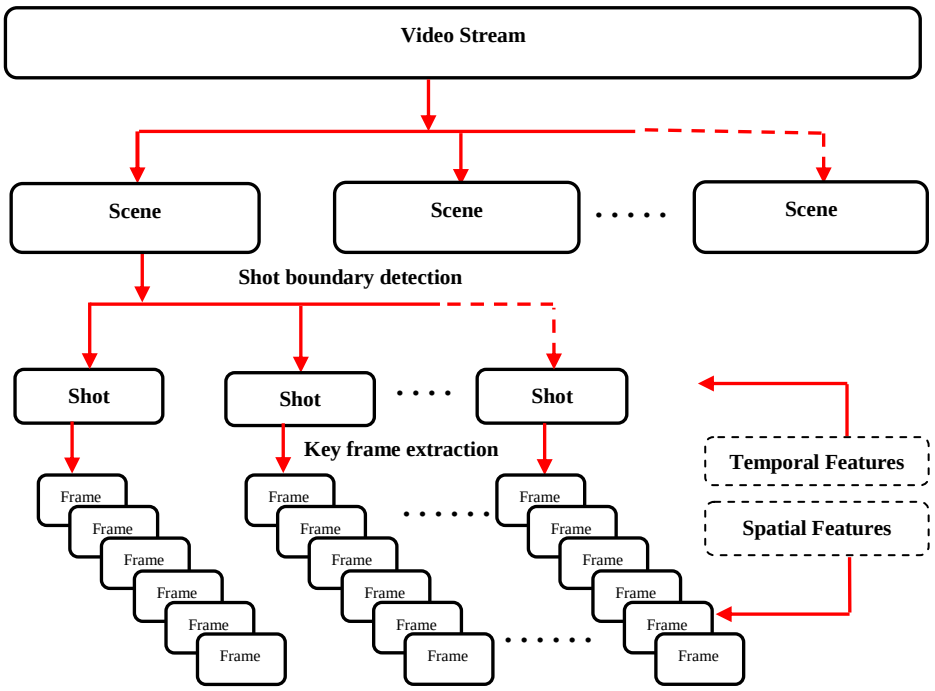

Figure 1- Video structure representation

3. Structure of Semantic Based Video Retrieval System

The general structure of the semantic based video retrieval consists of the following steps as shown in Figure -2.

- Structure analysis: includes shot boundary detection, key frames extraction and segment scenes.
- Feature extraction: includes extract features from segmented videos.
- Video mining: includes mining the extracted feature.
- Video annotation: includes the construction of the semantic index of extracted features and mining to the knowledge.
- User query: includes searching the video database for the desired video.
- Video feedback: includes optimizing search result through relevance feedback.

A general discussion on the overall process of the semantic video retrieval framework is given below:

3.1 Structure Analysis

Initially the video is split into shots through shot detection algorithms; then, key frames that represent the shot are determined.

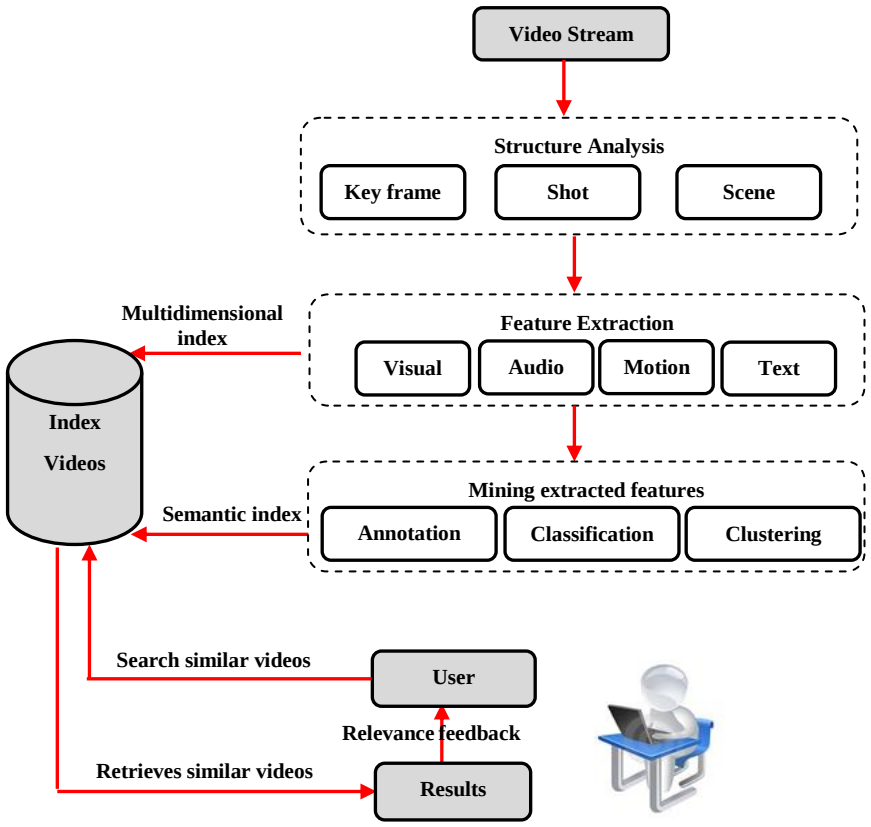

Figure 2- Structure of semantic based video retrieval system

3.1.1 Shot Boundary Detection

Shot detection is used to divide the entire video into a number of shots. A frame at a boundary position of the shot varies from its consecutive frames that belong to the next shot in visual features, and this is the essential principle that most detection algorithms depend upon. Boundaries of shot are categorized into: cut is a sudden transition between consecutive shots and gradual is soft transitions that can be of different types such as wipes, dissolve and fades [10]. The shot boundary detection approaches are usually used:

- Threshold approach: similarity between frames is compared with a predefined threshold [11].
- Statistical learning approach: detection boundary of shot is considered as classification task, where supervised learning like SVM [12] and unsupervised learning like fuzzy k-means [13] are employed.

3.1.2 Key Frame Extraction

Because of the repetition between the frames in the single shot; the frames that best reflect the contents of shot are chosen as the key frames to represent the shot. Color histogram, edges and shapes features are utilized to determine key frame. Key frame extraction is based on a sequential comparison [14], reference frame [15], clustering [16], and object-event [17].

3.1.3 Scene Segmentation

Scene has a higher level of semantic information than a shot. Scene segmentation techniques fall into three classes: key frame [18], visual information integration [19] and background [20].

3.2 Feature Extraction

Features are defined as a descriptive parameters extracted from video data. Generally features can be extracted and represented as features descriptors of the video data are: the low-level features (spatial feature), the high-level features (temporal features), the object features, the audio features and the motion features [1, 3].

3.2.1 Low Level Features

Low level features can be extracted from the key frames are global features that are extracted from the complete image and local features that describe the selected partition of an image [21].

Three classes of global features are usually used:

- Color features: typical representation of color feature involves color histogram, color moments, color correlogram, and color coherence vector [22]. Among them, color histogram is the most used, that describes the relative quantity of each color in the image while ignoring the spatial relationships between different colors.
- Texture features: texture features gives information about the arrangement of visible elements and their interconnection with the surrounding environment in an image. Texture information can be extracted using Gabor filters, wavelet transformation, orientation features and co-occurrence matrices [23, 24].
- Shape features: features of shapes are extracted by object contours or regions from the key frame after image segmentation by grouping pixels in the image based on homogeneity on color, texture, or both, or by connecting edge lines. Edge Histogram Descriptor (HDD) is an algorithm used for edges detection and describes the distribution of the edges using a histogram [25].

Local feature based methods are more robust for variations in scale, rotation, illumination and noise. The local feature extraction can be classified into two classes:

- Region detection: in order to obtain a robust detection result, the feature detectors should take into consideration factors such as photometric transformation and scale differences. Harris detector is proposed as corner detector [26].
- Region description: Region description: is used to describe the detected region in a precise and robust manner. Scale Invariant Feature Transformation (SIFT) descriptor is proposed as the feature description [27], which is measured the region around a kaypoint and describes each region using edge orientation histogram.

3.2.2 Object Features

Object features involve color, shape and texture of image regions that content the objects. These features are utilized to return video clips which are likely containing similar objects. Object representation is a method to describe an object, through which objects can be easily detected and retrieved from video stream. In general, objects can be represented by their shapes such that points based method, primitive geometric shapes and silhouette and contour and by their appearances such as probability density and templates [28]. The disadvantage of object features is that objects recognition in the video is complicated therefore; present algorithms focus on identifying a specific part of object like faces, rather than the whole object.

3.2.3 Motion Features

The motion is a key feature that basically characterizes dynamic video, representing the temporal information of video and is closer to objective and semantic concepts compared with other features such as color or texture. Motion based features fall into two classes [29]: camera based motion features such as zooming in or out, panning left or right, tilting up or down, and unknown (those are not pan, zoom, or tilt) and object based motion features that are further classified into:

- Statistics: the statistical motion features are extracted from points in video frames to form a motion distribution in the video [30]. These features cannot accurately represent action of objects and relations between objects.
- Trajectory: trajectory features are extracted through modeling motion trajectory of object in videos [31]. The accuracy of these features relies on correct segmentation and object tracking in moving videos
- Objects relationship: relationships between multiple objects are described in the temporal domain in objects relationship features [32]. The difficulty of labeling each object and its position is drawback of these features.

3.2.4 Audio Features

Audio plays an impertinent role in detecting and recognizing events in video. Audio can be speech, music or different special sounds. Audio features can be a rich source used to distinguish various speeches, different audio events, and spoken text. Generally, the audio features are divided into two groups: time domain features such as amplitudes, pitches and zero crossing rates and frequency domain features such as cepstral coefficients, spectrograms and mel frequency cepstral coefficients that are commonly used to identify speakers [33, 34].

3.2.5 Text Features

Text display in the video involves useful information for automatic annotation and indexing. Text in frames or sequence of frames will show many variations according to their properties such as motion (static, linear moment), color (polychrome, monochrome), geometry (size, alignment, inter character distance) and edge (text boundaries, strong edges) [35]. Enhancement process is required for text features due to the sensitivity to noise and low resolution for text region. Optical Character Recognition (OCR) technology is employed to extract text features and convert them into plain text.

3.2.6 Feature Fusion

The visual features, aural features and the motion features that are extracted from video data can be fused to produce more robust concept detection. Combined features improve performance at the cost of increasing complexity; to overcome the defects of fusion with the increase of accuracy detector, a form of independence is required from the features. To achieve independence there are two approaches for features fusion, the first approach is unimodal features, where the features are extracted from a single modality for example from visual features only [36].The second approach is multimodal features where features are extracted from multiple modalities, for example, the audio and the visual content [37].

3.3 Video Mining

Video mining is the process of discovering patterns and their correlation in order to extract undiscovered knowledge from the video components. The example of extracted knowledge involves structural patterns, moving objects patterns, event patterns and video semantic knowledge. Common video mining approaches include:

3.3.1 Semantic Event Mining

Video events are high level semantic information that people understand when watching a video. Video event detection seeks to make computer perception close to the human perception of events. Event understanding includes the analysis and discrimination of motion patterns, human behavior and object movement [38, 39]. The difficulty of understanding the video event is due to many reasons such as the inaccuracy of object detection and tracking, the variation in the appearance of certain events, the similarity in the appearance of different events, and the ambiguity in interpreting semantic definitions of events [40].

3.3.2 Pattern Mining

Unsupervised or semi supervised learning techniques are used to automatically detect unknown patterns. The discovered pattern can be used to detect uncommon events which are characterized by their differences from patterns discovered. Pattern mining also include discovering the special patterns that can be classified into: mining similar motion patterns [41] and mining similar objects [42].

3.3.3 Video Association Mining

Video association mining can be defined as the process of detecting unknown relationships between different events and identifying the more frequent association patterns for different objects like the occurrence of two objects in the same time [43].

3.3.4 Video Classification

Video classification task is to compile videos along with similar contents and then assign videos to a predefined class under the supervision8. Semantic based video classification is a difficult task because there is no direct link between the low level information extracted from video and human translation to this information. In video genre classification, the videos are categorized into different genres like news, movie, cartoon and sports. Video genres classification is using widely previous knowledge as well as the using of low level features due to the robustness of these features for video diversity. This technique fall into two classes:

- Rule approach is used knowledge to determine the heuristic rules and perform semantic video classification. The existing rules can be easy to insert, modify and delete when changing video classes [44]. This approach is useful for films and news videos which have obvious story structures.
- Statistical approach is used statistical machine learning that is used labeled samples with low level features to train a classifier for videos. Semantic video classification is supported by this approach through uncovers hidden rules among different video patterns, for example Bayesian network [45] and SVMs-based active learning [46] are used to classify video.

3.3.5 Video Clustering

Video clustering is the process of partitioning the video clips into different meaningful categories. It is unsupervised learning techniques used to extract knowledge from unlabeled samples of single video or a dataset [47]. Most of clustering techniques depends on the distance that calculates the similarity between the words. The Euclidean distance and Boosting method are used as distance function and distance metric evaluation for clustering respectively [48]. Partitioning and hierarchical algorithms are examples of clustering techniques [8].

3.4 Video Annotation

In semantic based retrieval, annotation is the process of assigning semantics concepts, like person, car, sky, and people walking to video shots [49]. There are two differences between video annotation and video classification: a different category or concept can be used in video classification compared to a video annotation, in spite of some concepts that can be used to both of them. The video classification applies to entire videos, while video annotation typically uses video shots as the basis unit [29]. Although the annotation is fundamental for video analysis because it helps to bridge the semantic gap; automatically producing annotations for video still a hard task. Based on the learning techniques, video annotation can be classified into three classes: supervised learning which is required a sufficiently number of labeled training samples to learn a robust detector for each concept, and the required number increases dramatically with feature dimension [50]. Active learning is an efficient approach that has been proposed to combine unlabeled sample with supervised learning techniques to handle the lack of labeled samples [51]. Semisupervised learning, it is also an effective approach that uses unlabeled samples to increase information in the available labeled examples [52].

3.5 User Query

The objective of video retrieval is to return the most relevant video given a user query. There is a wide difference in the query submission to the video search system.

3.5.1Types of Query

Types of queries can be categorized into those non-semantic based like query by objects and by example and those queries that are semantically based like queries by keywords and natural language [1, 29]:

- Query by example: the user provides an image or a video as an example to retrieve the desired video in this type of query. Low level features are extracted from a particular image or video example then similar videos are determined via measuring the features similarity. It's not always potential to get examples of the required video content.
- Query by sketch: the sketches for videos are drawn by user in order to use them to retrieve the desired video.
- Query by objects: the user provides the object's image and the system retrieve all object occurrences in the video database.
- Query by keywords: set of keywords is used to describe user's query. It has the ability to get the semantics from videos to a certain degree.
- Query by concept: also called to as query by conceptual or query by semantic, it is an extension of each keyword and example query, narrowing down results. It is depended on semantic annotations where high level concepts are linked with the video information.
- Query by natural language: it is the most natural and appropriate direction to represent the query. The hard part of this type of query is to analyze, and derive the semantics from natural language.
- Combination based query: integrate various types of queries like keywords queries and object queries. It is appropriate for the multiple model system.

3.5.2 Similarity Measure Techniques

Similarity measurements techniques are applied to the video indices in database according to the user query that given to retrieval system. Some of common similarity metrics are Euclidean distance, Squared Chord distance, Chi-Squared distance, Divergence and Correlation [53]. Depending on the query type, the approaches used to measure video similarity is determined. These approaches can be categorized into:

- Feature matching approach: the similarity between video and query is measured based on the distance between the features of the corresponding frames [54].

- Text matching approach: the similarity between the concept description text and the query text is calculated by using a vector space model after applying normalization process [55].
- Combination matching approach: It is incorporation of different matching approaches. It is resilient for the multiple models [56].

3.6 Relevance Feedback

Relevance feedback brings the user into loop retrieval in order to minimize the gap between what query represents and what the user thinks. In other words relevance feedback is optimized the retrieval results [57]. Relevance feedback reflects user's priority by ranking where the scores are given to the retrieved video based on the similarity between the query and the returned video. Videos are listed according to this result, so that the most relevant videos are displayed to the user at the top of the retrieved list. Three categories in relevance feedback:

- Explicit relevance feedback: claim the user to determine related videos that were previously recovered [58]. Although explicit feedback take advantage of users 'feedback directly, so better results can be guaranteed than others' feedback, but more interaction and user cooperation is needed.
- Implicit relevance feedback: refine the retrieval results by using click through when the user clicks on the retrieved videos [59]. Unlike the explicit feedback, the implicit feedback does not required user collaboration, making it more acceptable and practicable, but the information collected from the user is less precise than information of explicit feedback.
- Pseudo relevance feedback: positive and negative samples are selected from the prior retrieval results without the user intervention. These samples are reverted to the system for another research [60]. Although pseudo relevance feedback reduces user interaction, semantic gap causes pseudo relevance feedback is restricted in application.

4. Literature Review

The importance of the video retrieval system has led to several surveys, a good survey of video retrieval system as well as years of publication and topics, is provided in Table-1. This survey gives an overview of contributions that have been achieved in the field of semantic based video retrieval; some of previous work is provided here with a brief explanation to each of them:

Table 1- Survey papers in video retrieval

Year	paper	Topic
2008	[1]	Concept based video retrieval.
2009	[61]	Video retrieval based on spatio temporal information.
2009	[40]	Methods for understanding video events.
2011	[29]	Video indexing and retrieval based on visual content.
2012	[62]	Content based video retrieval systems.
2014	[63]	Multimodal feature extraction for semantic mining of soccer video.
2015	[64]	Reducing semantic gap in video retrieval with fusion.
2016	[28]	Human action analysis in videos for retrieval applications.

Kozintsev et al. [65] a new framework for semantic indexing and retrieving digital video is presented in this paper. A statistical framework for graph factor is proposed to detect semantic concepts using multiple media and features in order to bridge the semantic gap. Factor graphs model is used to represent the relationship between concepts while the sum product algorithm is used a tool for performing learning and inference for the global model. The results show that framework is flexible and provides a good basis for the future improvements.

Ma and Zhang [66] design a motion pattern descriptor to indicate the motion features of video in a general way. Support Vector Machines (SVMs) are employed to assign motion texture to semantic concepts. The results show that motion texture is compact and effective to represent a motion pattern as well as is improve the performance of motion based shot retrieval due to the comprehensiveness of motion pattern descriptor and the semantic classification ability.

Amir et al. [67] are designed a framework to detect events from video using trained classifiers that are used to automatically annotate video with semantic labels. The proposed framework is integrated

the visual model with speech model to detect new event in search process. The results indicate that visual features can be used to classify shots by semantic level concepts, but this requires a lot of labeled data that are very time consuming to generate. Although speech based video retrieval provides direct access to semantic information; speech retrieval fails if queries rely on visual content. The results show that combined speech and content based retrieval is best from each of the methods alone.
Yan and Naphade [68] Semi-supervised Cross Feature Learning (SCFL) algorithm is proposed in this paper. As opposite to co-training which learns each classifier by combine the selected unlabeled sample to increase the labeled set, SCFL learns separate classifiers from the selected unlabeled samples and combine them with the classifiers learned from labeled samples without noise. SCFL is more robust and extensible than co-training; which makes it more appropriate for detecting semantic concepts of video. The result indicates that SCFL improve the performance of the traditional co-training algorithms.

Bai et al. [69] are proposed a semantic analysis model based on Perception Concepts (PCs) and Finite State Machines (FSMs) to automatically describe and detect semantic patterns for sports video. Graph matching method is employed to discover high level semantic content to avoid the generating complex SQL queries. The results indicate that the designed system, yielding an average recall of 95% and an average precision of %91 for five defined event (goal scored, foul and yellow card for soccer videos, highlight attack and foul for basketball videos).

Hu et al. [70] are proposed a semantic retrieval framework for traffic video. Hierarchical clustering is applied to obtain motion trajectories using the spatial and temporal information. Spatial information is employed to cluster all trajectories into number of categories, and trajectories in each of the categories are further clustered into subcategories using the temporal information. Semantic concepts are assigned manually to the semantic content of the corresponding category to form semantic activity models. The proposed system is supported keywords, sketch and multiple object queries. The proposed algorithm is tested in a crowded traffic scene and the results demonstrate the robustness of the tracking algorithm and the effectiveness of the algorithm for learning activity models.

Shyu et al. [71] are proposed framework aim to address event detection and concept extraction without using domain knowledge by integrating data mining techniques. Distance based technique is used to implement the rough classification involve the refining, combination and selection features. While rule based algorithm is used for further classification. The system performance is evaluated by soccer videos and TRECVID (2004, 2005) broadcast news videos. The experimentation result is demonstrated the adaptively of the proposed framework for concept and event detection.

Yadav and Aygün [72] are presented a method to search for user interest clips or video in video database with semantic content. The proposed method is composed in to three steps: user browsing where the video clips is classified in to interesting and uninteresting group, query structuring where intelligent query called I-Quest is presented in order to compute the relevance of each clip to the user's query based on interesting and uninteresting sets and query processing and ranking where user browsing feedback is employed in video databases. The result shows that proposed I-Quest guides the user to access the interesting video clips when the user cannot properly formulate the query.

Hopfgartner and Jose [73] are presented a semantic based user modeling technique to capture the evolving interests of users for video news and represent these interests in dynamic user profiles. The proposed approach is employed Linked Open Data Cloud to capture and organize users interests as well as implicit relevance feedback techniques in order to return and recommend news video to users. The semantic context of the news stories in the user profile are used to fetch and present new relevant videos. Unlike standard interactive video retrieval experiments, an assessment of proposed approach is performed in an uncontrolled environment and the results shows the semantics user profiling is effective in recommending relevant results.

Memar et al. [74] are proposed method that is based on the integration of knowledge-based and corpus-based semantic word similarity measures in order to retrieve video shots for concepts whose annotations are not available for the system. Mean Average Precision (MAP) is used with TRECVID 2005 dataset to evaluate the superiority of integrated similarity method and the results show that the combined corpus based and knowledge based measures is better than each of the methods alone.

Dalton et al. [75] this study is proposed a method to model text extracted from images in the videos by Optical Character Recognition (OCR), text recognized in the speech of its audio track by Automatic Speech Recognition (ASR), as well as automatically detected semantically meaningful

visual video concepts identified in the videos for retrieval. Large external sources of text are used to construct text language models for each concept. For example, a face detector will be includes concepts that are related to faces (e.g. nose, eyes, mouth). The integrated results from different modalities achieve high precision and recall that is greater than any individual modality. The proposed work provides additional improvements of over 50% if the relevance feedback approaches is applied.

Vallet et al. [76] are designed system that is exploited external knowledge to provide visual examples related to a user search query which is employed as search inputs for low level feature retrieval models. Different external knowledge is employed in the proposed system: DBpedia (a highly structured), Flickr (a semi-structured) and Google Images (no metadata structure) which have different characteristics. To address the semantic gap, they exploit the available semantics in external knowledge above, reduce the ambiguity of the query, and focus the scope of image searches in repositories. The evaluations results show external knowledge improves the quality of the retrieved visual examples especially when the external knowledge is more structured.

Wang et al. [77] this paper presents a framework for extracting semantic information in real-world videos using audio features. The proposed system composed in to three steps. First, vocal feature over fixed length of sliding windows is extract, then, classifiers trained on audio concepts is applied in order to calculate the occurrence matrix. Finally, clip level feature is produced from the occurrence matrix. The proposed system is compared with fusion the semantic features with text feature and low level feature for the event based retrieval task and the result indicates that audio semantic concepts capture complementary information in the soundtrack.

Lin et al. [78] the objective of this study is using complex natural language queries to retrieve video. The proposed method is parsed the videos semantically using motion features and object appearance and then learned the importance of each term using structure prediction. Natural language query is parsed into a semantic graph which is then matched to the visual concepts by matching algorithm. The result show the effectiveness of proposed approach and the ability to locate a most port of the objects described in the query with high accuracy.

Jiang et al. [79] in this paper framework is proposed to addresses challenge of content based search in 100 million Internet videos. A step called concept adjustment based on a concise optimization is a key solution that aims to represent a video via a few salient and consistent concepts. Scores are linked with semantic concepts to indicate how confidently they are detected. Experimental performance indicates the scalability and efficiency of proposed algorithm. The time is needed for searching video is only 0.2 second.

Jiang et al. [60] are presented semantic based video search engine that is allowed for semantic search over Internet videos without using metadata or example videos. The paper improve zero-example search (called E-Lamp, setting in Multimedia Event Detection (MED) by the TRECVID community) that is detected the occurrence of a main event in a video and is used semantic concept as query. Different techniques can be employed such that visual and audio concept detectors or exploring interactive search schemes for improvement process. The proposed system acquires the best performance in TRECVID 2014 on collection of 200000 Internet videos according to NIST evaluation.

Chen et al. [80] are proposed a framework based on the Gaussian Mixture Model (GMM) which has the ability to retrieve semantic concepts, even from highly unbalanced datasets. The gaussian components are generated dynamically by GMM. The GMM divide the positive data instances assigned to the nearest gaussian component with positive training set to several gaussian distributed subsets. The impact of this step strengthens the newly consolidated data set. Two benchmark datasets (NUS-WIDE-LITE, MediaMill Challenge Problem) are used to evaluate the performance of proposed system in terms of the Mean Average Precision (MAP). The results show the effectiveness of the proposed GMM framework.

Agharwal et al. [81] the main objective of this study is to overcome the semantic query gap by employing the Continuous Word Space (CWS) embedding scheme to obviously compute query and detector concept similarity. The proposed technique also is utilized beforehand proposed method to create a Concept Space (CoS) video embedding pipeline, and implemented the Dictionary Space (DiS) video embedding retrieval. The experimental results indicate that proposed method is surpass beforehand methods using CoS, DiS. Although the implementation of proposed method is expensive; but it is fast computation and produce compact video representation that property to real-time interactive system.

Wu et al. [82] in this study crowd video retrieval system is proposed using hand drawn sketches as queries. The difficulty in this work is crowd motion representation and similarity measurement therefore; the motion structure coding algorithm is used for motion level crowd video indexing and sketch representation and distance metric fusion technique incorporated with Ranking SVM is utilized for measuring the relevant degree between a sketch query and the motion crowd videos. The experimental results show that the proposed method is robust and effective and outperforms of retrieval performance than alternative methods.

De Boer et al. [83] the main objective in this study is to improve video event retrieval by user feedback. The proposed method is presented the user feedback in two levels: concept level and video level. Adaptive Relevance Feedback (ARF) is presented on video level and Query Point Modification (QPM) methods with a method that changes the semantic space is presented on concept level. Results show that relevance feedback on both concept and video level improves performance compared to without using relevance feedback; relevance feedback on video level achieves higher performance compared to relevance feedback on concept level.

Zhang et al. [84] in this paper the trajectory based bag of visual words pipeline is improved to retrieve video action by combining spatial temporal information. A descriptor coding method is used to capture the spatial temporal correlations among trajectories and feature of individual trajectories. Trajectory matching stages are improved to handle with the miss alignments between dense trajectory segments. The evaluation results indicate that the proposed method improves the action video retrieval performance, especially on dynamic actions with large movements and interlocking backgrounds.

5. Conclusions

An overview of video retrieval based on the semantic is covered in this paper. The essentially task of video retrieval algorithm is return the closed similar video from a given data collection based on a user query. The performance of semantic video retrieval systems is still inadequate even there are much research efforts on the systems development. Discovering and extracting the sematic concept and knowledge of video information as well as problem of semantic gap are the main challenge of modern video retrieval system. There is no universal framework that can be applied to all kinds of video for semantic features extraction. Knowledge is employed by proposed systems in order to enhance retrieval efficiency for particular field but these systems cannot be applied to videos from other fields. Some observations are derived from the test results included in the review papers that are presented in this survey such that:

- The construction of accurate detection devices seems to be a reasonable strategy, when systems use semantic features that are automatically detected.
- Relevance feedback is an effective method to update query iteratively by gathering user's feedback during search session. As a consequence query is improved as well as increased effective of retrieval performance.
- Retrieval models may have fundamental effect on the search result, combining a reasonable strategy in order to obtain multi-modality and multi-concept learning leads to exploitation their respective strengths and upgrades the performance of retrieval system.

Although much work has been done in this scope, many issues remain open and deserve further consideration such as effective learning of high-level semantic, motion features and object tracking analysis, query-language design, hierarchical analysis of video indices, and ontology large-scale concept for videos.

References

1. Snoek, C. G. M. and Worring, M. **2008**. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, **2**(4): 215-322.
2. Sugandhiand, A. and Sharma, D. **2016**. Content based Video Retrieval using Text Annotation and Low Level Features Technique. *International Journal of Computer Applications*, **145** (14): 11-16.
3. Ansari, A. and Mohammed, M.H. **2015**. Content based Video Retrieval Systems-Methods, Techniques, Trends and Challenges. *International Journal of Computer Applications*, **112**(7): 13-23.
4. Turaga, M. P., Pugliese, R. A. and Subrahmanian, V. S. **2010**. Semantic video content analysis. In D. Schonfeld, C. Shan, D. Tao and L.Wang (eds.). *Video Search and Mining*. Berlin: Springer, 147-176.

5. Hauptmann, A. G., Christel, M. G. and Yan, R. **2008**. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, **96**(4): 602-622.

6. Chen, S., Zhao,N., and Mei-Ling, S. **2007**. Modeling semantic concepts and user preferences in content-based video retrieval. *International Journal of Semantic Computing*, **1**(03): 377- 402.

7. Memar, Sara., Affendey, L.S., Mustapha, N., and Ektefa, M. **2013**. Concept-based video retrieval model based on the combination of semantic similarity measures. In proceedings of IEEE 13th International Conference on Intelligent Systems Design and Applications, pp.: 64-68.

8. Vijayakumar, V., and Nedunchezhian, R. **2012**. A study on video data mining. *International journal of multimedia information retrieval*, **1**(3):153-172.

9. Pal, G., Rudrapaul, D., Acharjee, S., Ray, R., Chakraborty, S. and Dey, N. **2015**. Video shot boundary detection: a review. In Satapathy S.C. et al. (eds.). *Advances in Intelligent Systems and Computing*. Cham: Springer, 147-176.

10. Yuan, J., Wang, H., Xiao,L., Zheng,W., Li, J., Lin,F., and Zhang, B. **2007**. A formal study of shot boundary detection. *IEEE transactions on circuits and systems for video technology*, **17**(2):168-86.

11. Lee, H., Yu, J., Im, Y., Gil, J., and Park, D. **2011**. A unified scheme of shot boundary detection and anchor shot detection in news video story parsing. *Multimedia Tools and Applications*, **51**(3):1127-1145.

12. Ko, K., Cheon, Y. M., Kim, G., Choi, H., Shin, S., and Rhee,Y. **2006**. Video shot boundary detection algorithm. In P. Kalra and S. Peleg (eds.). *Computer Vision, Graphics and Image Processing*. Berlin: Springer, 388-396.

13. Lo, C., and Wang, S. **2001**. Video segmentation using a histogram based fuzzy c-means clustering algorithm. *Computer Standards & Interfaces*, **23**(5): 429-438.

14. Zhang, X., Liu,T., Lo, K., and Feng, J. **2003**. Dynamic selection and effective compression of key frames for video abstraction. *Pattern recognition letters*, **24**(9):1523-1532.

15. Sun, Z., Jia, K., and Chen, H. **2008**. Video key frame extraction based on spatial-temporal color distribution. In proceedings of IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp:196-199.

16. Yu, X., Wang, L., Tian, Q., and Xue, P. **2004**. Multilevel video representation with application to keyframe extraction. In proceedings of IEEE 10th International Conference on Multimedia Modelling, pp:117-123.

17. Kang, H. and Hua, X. **2005**. To learn representativeness of video frames. In Proceedings of the 13th annual ACM international conference on Multimedia, ACM, pp: 423-426.

18. Truong, B., Venkatesh, S. and Dorai, C. **2003**. Scene extraction in motion pictures. *IEEE Transactions on Circuits and Systems for Video Technology*, **13**(1): 5-15.

19. Sundaram, H. and Chang, S. **2000**. Video scene segmentation using video and audio features. In proceedings of IEEE International Conference on Multimedia and Expo, vol. 2, pp:1145-1148.

20. Chen, L., Lai, Y. and Liao, H.M. **2008**. Movie scene segmentation using background information. *Pattern Recognition*, **41**(3):1056-1065.

21. Jiang, S., Tian,Y., Huang, Q., Huang, T. and Gao, W. **2009**. Content-Based Video Semantic Analysis. In Tao, D., Xu, D., Li, X. (eds.). *Semantic Mining Technologies for Multimedia Databases*. New York: IGI Global, 211-235.

22. Liu, Y., Zhang,D., Lu,G., and Ma,W. **2007**. A survey of content-based image retrieval with high-level semantics. *Pattern recognition*, **40**(1): 262-282.

23. Ansari, M. A. and Vasishtha, H. **2016**. Content based Video Retrieval Systems Performance based on Multiple Features and Multiple Frames using SVM. *International Journal of Advanced Computer Science & Applications*, **1**(7):100-105.

24. Thepade, S. D. and Yadav, N. **2015**. Novel efficient content based video retrieval method using cosine-haar hybrid wavelet transform with energy compaction. In proceedings of IEEE International Conference on Computing Communication Control and Automation (ICCUBEA), pp: 615-619.

25. Kanagavalli, R. and Duraiswamy, K. **2012**. Shot Detection Using Genetic Edge Histogram and Object Based Video Retrieval Using Multiple Features. *Journal of Computer Science*, **8**(8): 1364-1371.

26. Lowe, D. G. **2004**. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, **60**(2): 91-110.

27. Mikolajczyk, K. and Schmid, C. **2004**. Scale & affine invariant interest point detectors. *International journal of computer vision*, **60**(1): 63-86.
28. Ramezani, M. and Yaghmaee, F. **2016**. A review on human action analysis in videos for retrieval applications. *Artificial Intelligence Review*, **46**(4): 485-514.
29. Hu, W., Xie, N., Li, L., Zeng, X. and Maybank, S. **2011**. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **41**(6): 797-819.
30. Ma, Y. and Zhang, H. **2002**. Motion texture: a new motion based video representation. In proceedings of IEEE 16th International Conference on Pattern Recognition, vol.2, pp: 548-551.
31. Dyana, A. and Das, S. **2009**. Trajectory representation using Gabor features for motion-based video retrieval. *Pattern Recognition Letters*, **30**(10):877-892.
32. Ngo, C., Pong, T. and Zhang, H. **2002**. Motion-based video representation for scene change detection. *International Journal of Computer Vision*, **50**(2):127-142.
33. Feki, I., Ammar, A. B. and Alimi, A. M. **2016**. Automatic environmental sound concepts discovery for video retrieval. *International Journal of Multimedia Information Retrieval*, **5**(2):105-15.
34. Lee, K., and Ellis, D. P. W. **2010**. Audio-based semantic concept classification for consumer video. *IEEE Transactions on Audio, Speech, and Language Processing*, **18**(6):1406-1416.
35. Bhute, A. N. and Meshram, B. B. **2014**. Text Based Approach For Indexing And Retrieval Of Image And Video: A Review. *Advances in Vision Computing: An International Journal (AVC)*, **1**(1): 27-38.
36. Mohamadzadeh, S. and Farsi, H. **2016**. Content based video retrieval based on hdwt and sparse representation. *Image Analysis & Stereology*, **35**(2): 67-80.
37. Chen, X., Hero, A.O., III and Savarese, S. **2012**. Multimodal video indexing and retrieval using directed information. *IEEE Transactions on Multimedia*, **14**(1): 1-14.
38. Wang, F., Sun, Z., Jiang, Y. and Ngo, C. **2014**. Video event detection using motion relativity and feature selection. *IEEE Transactions on multimedia*, **16**(5): 1303-1315.
39. Meng, J., Yuan, J., Yang, J., Wang, G., and Tan, Y. **2016**. Object instance search in videos via spatio-temporal trajectory discovery. *IEEE Transactions on Multimedia*, **18**(1): 116-127.
40. Lavee, G., Rivlin, E. and Rudzsky, M. **2009**. 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)*, **39**(5): 489-504.
41. Lai, C., Rafa, T. and Dwight, E. N. **2006**. Mining motion patterns using color motion map clustering. *ACM SIGKDD Explorations Newsletter*, **8**(2): 3-10.
42. Anjulan, A. and Canagarajah, N. **2007**. A novel video mining system. In proceedings of IEEE International Conference on Image Processing, vol. 1, pp: I-185.
43. Zhu, X., Wu, X. Elmagarmid, A. K., Feng, Z., and Wu, L. **2005**. Video data mining: Semantic indexing and event detection from the association perspective. *IEEE Transactions on Knowledge and Data engineering*, **17**(5): 665-677.
44. Zhou, W., Dao, S. and Kuo, C. J. **2002**. On-line knowledge-and rule-based video classification system for video indexing and dissemination. *Information Systems*, **27**(8): 559-586.
45. Ramya, S. T. and Rangarajan, P. **2011**. Knowledge based methods for video data retrieval. *International Journal of Computer Science & Information Technology*, **3**(5): 165-172.
46. Nagaraja, G. S., Rajashekara M. S. and Deepak, T.S. **2015**. Content based video retrieval using support vector machine classification. In proceedings of IEEE International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp: 821-827.
47. Saravanan, D. and Srinivasan, S. **2015**. Video Data Mining Information Retrieval Using BIRCH Clustering Technique. In L.P. Suresh et al. (eds.). *Artificial Intelligence and Evolutionary Algorithms in Engineering Systems*, India: Springer, pp: 583-594.
48. Kumar, C. R. and Sujatha, S. N. N. **2014**. STAR: Semi-supervised-clustering Technique with Application for Retrieval of video. In proceedings of IEEE International Conference on Intelligent Computing Applications, pp: 223-227.
49. Feng, J. and Zhou, W. **2014**. An efficient method for automatic video annotation and retrieval in visual sensor networks. *International Journal of Distributed Sensor Networks*, **10**(3): 1-8.

50. Snoek, C. G. M., Worring, M., Gemert, J. C. V., Geusebroek, J. and Smeulders, A. W. M. **2006**. The challenge problem for automated detection of 101 semantic concepts in multimedia. In Proceedings of the 14th ACM international conference on Multimedia, pp: 421-430.
51. Song, Y., Hua, X., Dai, L. and Wang, M. **2005**. Semi-automatic video annotation based on active learning with multiple complementary predictors. In Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval, pp: 97-104.
52. Ewerth, R. and Freisleben, B. **2007**. Semi-supervised learning for semantic video retrieval. In Proceedings of the 6th ACM international conference on Image and video retrieval, ACM, pp: 154-161.
53. Cha, S. **2007**. Comprehensive survey on distance/similarity measures between probability density functions. *International journal of mathematical models and methods in applied sciences*, **1**(4): 300-307.
54. Browne, P. and Smeaton, A. F. **2005**. Video retrieval using dialogue, keyframe similarity and video objects. In proceedings of IEEE International Conference on Image Processing, vol. 3, pp: III-1208.
55. Snoek, C. G. M., Huurnink, B. Hollink, L., Rijke, M. D., Schreiber, G. and Worring, M. **2007**. Adding semantics to detectors for video retrieval. *IEEE Transactions on multimedia*, **9**(5): 975-986.
56. Yan, R., Yang, J. and Hauptmann, A.G. **2004**. Learning query-class dependent weights in automatic video retrieval. In Proceedings of the 12th annual ACM international conference on Multimedia, pp: 548-555.
57. Aksoy, S. and Cavus, O. **2005**. A relevance feedback technique for multimodal retrieval of news videos. In proceedings of IEEE International Conference on Computer as a Tool, EUROCON, vol.1, pp: 139-14.
58. Chen, L., Chin, K. and Liao, H. **2008**. An integrated approach to video retrieval. In Proceedings of the 19 conference on Australasian database, Australian Computer Society, vol.75, pp: 49-55.
59. Ghosh, H., Poornachander, P., Mallik, A. and Chaudhury, S. **2007**. Learning ontology for personalized video retrieval. In Workshop on multimedia information retrieval on the many faces of multimedia semantics, Germany, pp: 39-46.
60. Jiang, L., Yu, S., Meng, D., Mitamura, T. and Hauptmann, A. G. **2015**. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, pp: 27-34.
61. Ren, W., Singh, S., Singh, M. and Zhu, Y. S. **2009**. State-of-the-art on spatio-temporal information-based video retrieval. *Pattern Recognition*, **42**(2): 267-282.
62. Patel, B. V. and Meshram, B. B. **2012**. Content based video retrieval systems. *International Journal of UbiComp (IJU)*, **3**(2): 13-30.
63. Oskouie, P., Alipour, S. and Eftekhari-Moghadam, A. **2014**. Multimodal feature extraction and fusion for semantic mining of soccer video: a survey. *Artificial Intelligence Review*, **42**:1-38. doi.org/10.1007/s10462-012-9332-4
64. Sudha, D. and Priyadarshini, J. **2015**. Reducing Semantic Gap in Video Retrieval with Fusion: A Survey. *Procedia Computer Science*, **50**: 496-502. doi.10.1016/j.procs.2015.04.020
65. Kozintsev, M.R.N.I., Huang, T. S. and Ramchandran, K. **2000**. A factor graph framework for semantic indexing and retrieval in video. In Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries, Hilton Head Island, South Carolina, pp.35.
66. Ma, Y. and Zhang, H. **2003**. Motion pattern-based video classification and retrieval. *EURASIP Journal on Applied Signal Processing*, **2**:199-208. doi.org/10.1155/S1110865703211021
67. Amir, A., Basu,S., Iyengar, G., Lin,C., Naphade, M., Smith, J. R., Srinivasan, S. and Tseng, B. **2004**. A multi-modal system for the retrieval of semantic video events. *Computer Vision and Image Understanding*, **96**(2): 216-236.
68. Yan, R. and Naphade, M. **2005**. Semi-supervised cross feature learning for semantic concept detection in videos. In proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, vol. 1, pp: 657-663.
69. Bai, L., Lao, S., Jones, G. J. and Smeaton, A. F. **2007**. A semantic content analysis model for sports video based on perception concepts and finite state machines. In proceedings of IEEE International Conference on Multimedia and Expo, pp: 1407-1410.

70. Hu, W., Xie, D., Fu, Z., Zeng, W. and Maybank, S. **2007**. Semantic- based surveillance video retrieval. *IEEE Transactions on image processing*, **16**(4): 1168-1181.
71. Shyu, M., Xie, Z., Chen, M. and Chen, S. **2008**. Video semantic event/concept detection using a subspace-based multimedia data mining framework. *IEEE Transactions on Multimedia*, **10**(2): 252-259.
72. Yadav, T. and Aygün, R. S. **2009**. I-Quest: an intelligent query structuring based on user browsing feedback for semantic retrieval of video data. *Multimedia Tools and Applications*, **43**(2): 145-178.
73. Hopfgartner, F. and Jose, J. M. **2010**. Semantic user modelling for personal news video retrieval. In International Conference on Multimedia Modeling, Springer Berlin Heidelberg, pp: 336-346.
74. Memar, S., Affendey, L.S., Mustapha, N., Doraisamy, S. C. and Ektefa, M. **2013**. An integrated semantic-based approach in concept based video retrieval. *Multimedia Tools and Applications*, **64**(1): 77-95.
75. Dalton, J., Allan, J. and Mirajkar, P. **2013**. Zero-shot video retrieval using content and concepts. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pp:1857-1860.
76. Vallet, D., Cantador, I. and Jose, J. M. **2013**. Exploiting semantics on external resources to gather visual examples for video retrieval. *International Journal of Multimedia Information Retrieval*, **2**(2): 117-130.
77. Wang, Y., Rawat,S. and Metze, F. **2014**. Exploring audio semantic concepts for event-based video retrieval. In proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp:1360-1364.
78. Lin, D., Fidler, S., Kong, C. and Urtasun, R. **2014**. Visual semantic search: Retrieving videos via complex textual queries. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp: 2657-2664.
79. Jiang, L., Yu,S., Meng, D., Yang,Y., Mitamura,T. and Hauptmann, A.G. **2015**. Fast and accurate content-based semantic search in 100m internet videos. In Proceedings of the 23rd ACM international conference on Multimedia, pp: 49-58.
80. Chen, C., Shyu, M. and Chen, S. **2016**. Weighted subspace modeling for semantic concept retrieval using gaussian mixture models. *Information Systems Frontiers*, **18**(5): 877-889.
81. Agharwal, A., Kovvuri, R., Nevatia, R. and Snoek, C. G. M. **2016**. Tag-based video retrieval by embedding semantic content in a continuous word space. In proceedings of IEEE Winter Conference on Applications of Computer Vision, pp: 1-8.
82. Wu, S., Yang, H., Zheng, S., Su, H., Zhou, Q. and Lu, X. **2017**. Motion sketch based crowd video retrieval. *Multimedia Tools and Applications*, **76**: 20167- 20195. doi.org/10.1007/s11042-017-4568-2
83. De Boer, M., Pingen, G., Knook,D., Schutte, K. and Kraaij, W. **2017**. Improving video event retrieval by user feedback. Multimedia Tools and Applications, 76: 22361-22381. doi.org/10.1007/s11042-017-4798-3
84. Zhang, L., Wang, Z., Yao,T., Mei, T. and Feng, D. D. **2017**. Exploiting spatial-temporal context for trajectory based action video retrieval. *Multimedia Tools and Applications*, pp:1-25. doi.org/10.1007/s11042-017-4353-2