Scientific Research Publishing

Journal of Computer and Communications, 2018, 6, 28-44
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

# Semantic-Based Video Retrieval Survey

**Shaimaa Toriah Mohamed Toriah[1*], Atef Zaki Ghalwash[2], Aliaa A. A. Youssif[2]**

[1]Department of Computer Science, Faculty of Computers and Informatics, Benha University, Benha, Egypt
[2]Computer Science Department, Faculty of Computers and Information, Helwan University, Helwan, Egypt
Email: *shaimaa_toriah@yahoo.com

**How to cite this paper:** Toriah, S.T.M., Ghalwash, A.Z. and Youssif, A.A.A. (2018) Semantic-Based Video Retrieval Survey. *Journal of Computer and Communications*, **6**, 28-44.
https://doi.org/10.4236/jcc.2018.68003

**Received:** October 8, 2017
**Accepted:** August 3, 2018
**Published:** August 6, 2018



## Abstract

There is a tremendous growth of digital data due to the stunning progress of digital devices which facilitates capturing them. Digital data include image, text, and video. Video represents a rich source of information. Thus, there is an urgent need to retrieve, organize, and automate videos. Video retrieval is a vital process in multimedia applications such as video search engines, digital museums, and video-on-demand broadcasting. In this paper, the different approaches of video retrieval are outlined and briefly categorized. Moreover, the different methods that bridge the semantic gap in video retrieval are discussed in more details.

## Keywords

Semantic Video Retrieval, Concept Detectors, Context Based Concept Fusion, Semantic Gap

## 1. Introduction

Digital data plays an essential role in our life. The digital data include videos, images, documents, sounds, etc. The video represents a rich source of information. The video can contain all the other digital data such as images, sounds, and texts. In addition, the video is characterized by its temporal consistency [1]. The rapid progress of digital devices causes inflation in the video databases. Retrieving the required information from the video database according to user's needs is called a video retrieval process. A video retrieval is a branched field from the generalized one called information retrieval. Information retrieval is considered a sub-field of computer science which organizes and retrieves the information from large database collections. Video retrieval methods are important and essential for multimedia applications such as video search engines, digital museums, video-on-demand broadcasting, etc.

Video retrieval is still an active problem due to the semantic gap, and the widespread of social media and the enormous technological development. Providing an efficient video retrieval with these huge amounts of videos on the web or even stored on the storage media is a difficult problem. The causation of the semantic gap is the difference between user requirements which are represented in queries and the low-level representation of videos on the storage media. Many methods are proposed to solve this semantic gap [2]-[7], etc., but it is not fully bridged. In this paper, a concise overview of the content-based video retrieval is mentioned. After that, the definition and the causes of a semantic gap in video retrieval will be explored. As the concept detectors [8] play a vital role in semantic video retrieval, a thorough study of the obstacles that face the construction of the generic concept detectors will be presented. Finally, the different methods model semantic concept relationships in video retrieval are categorized and explained in more details.

The main contributions of this survey are as follows:

1) We present a modern categorization of video retrieval approaches.

2) We identify the semantic gap problem.

3) We discuss the obstacles and challenges facing concept detectors construction.

4) We present the novel definitions of the different methods of semantic video retrieval.

The remainder of the paper is organized as the follows. Section 2 briefly introduces the video retrieval. Subsection 2.1 in Section 2 briefly reviews content-based retrieval. Subsection 2.2 in Section 2 discusses the semantic gap. Subsection 2.3 in Section 2 discusses the concept detectors in more details. Subsection 2.4 in Section 2 discusses the different methods of semantic video retrieval.

## 2. Video Retrieval

Video retrieval is concerned with retrieving specific video's shots according to user's needs (usually called query). Video retrieval is still an active problem due to the enormous technological development that allows easy video capturing and sharing. Maintaining an efficient video retrieval with these huge amounts of videos on the web or even stored on the storage media is a difficult problem. Video retrieval process includes video segmentation process, video low-level features extraction process, high-level concepts extraction process, video indexing, and query processing process, as shown in **Figure 1**. High-level concepts and semantics extraction process may include concept detectors and video benchmarking. In this paper, the video retrieval research is classified into content-based video retrieval, semantic gap, concept detectors, and semantic video retrieval. In **Figure 2**, the video retrieval research includes content-based video retrieval methods, the semantic gap problem, concept detectors, and finally

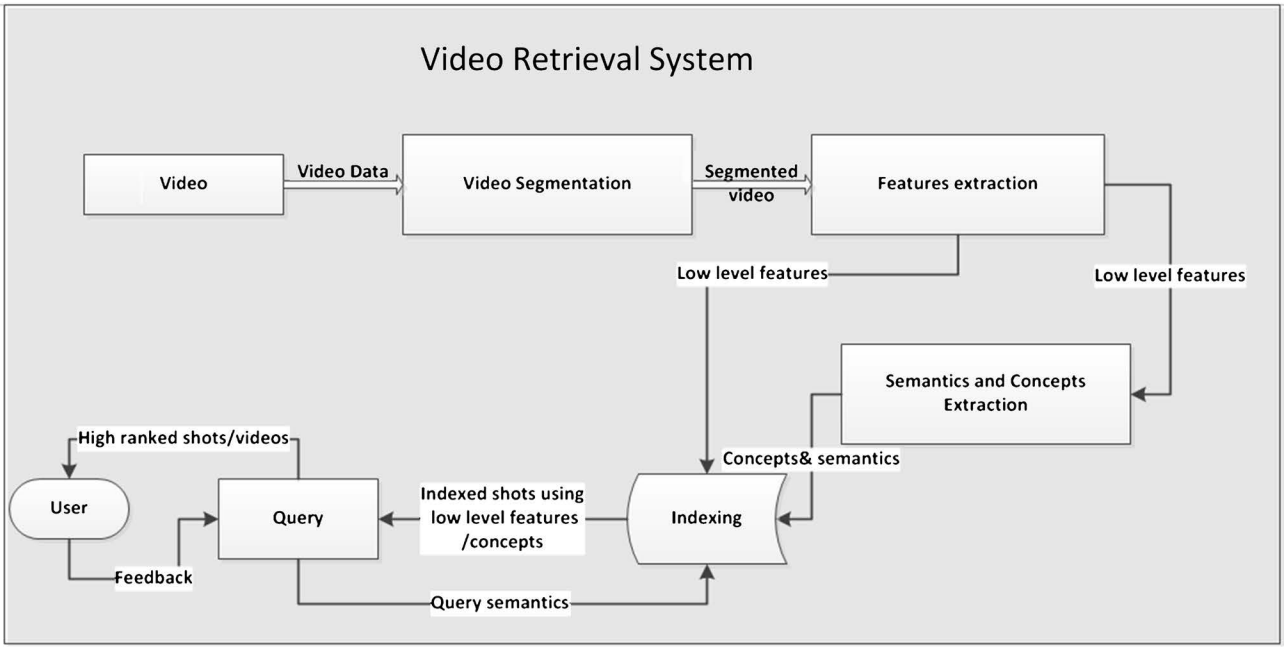


**Figure 1.** Video retrieval process.

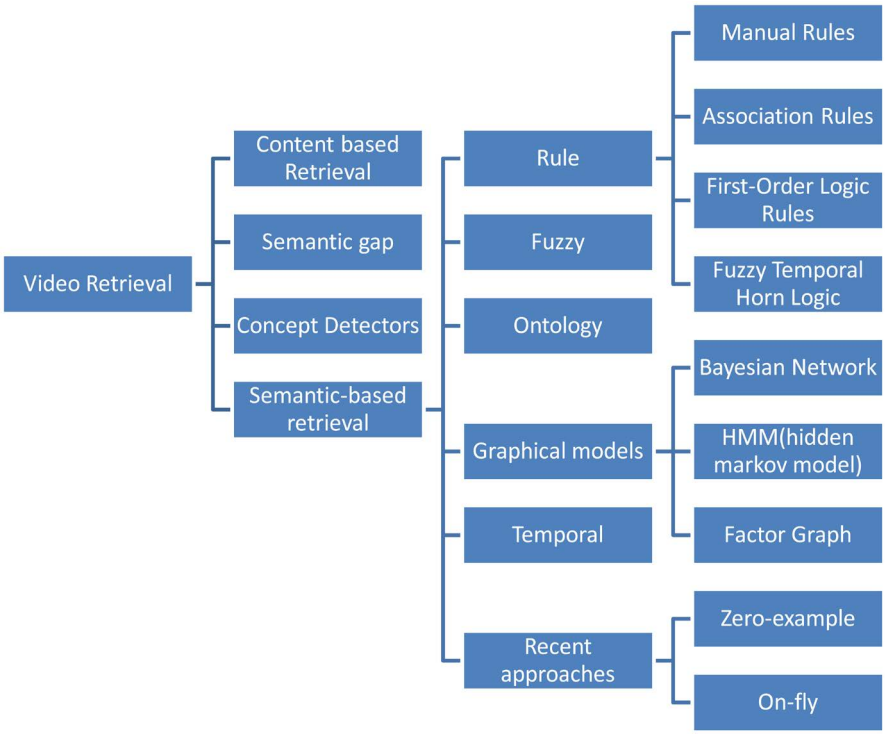


**Figure 2.** Video retrieval categories and methods.

semantic-based video retrieval methods. In addition, the content-based video retrieval methods are divided into video segmentation and video feature analysis and extraction methods. The video segmentation methods are categorized into shots boundary and keyframe detection methods, but video feature analysis and extraction methods are divided into keyframe features, motion features, and ob-

ject features extraction methods. Semantic-based video methods are categorized into the rule, fuzzy, graphical models, and temporal based retrieval methods. Finally, the recent approaches in semantic video retrieval are presented such as on-fly and zero-example video retrieval approaches.

### 2.1. Content-Based Video Retrieval

Content-based video retrieval includes video segmentation, and low-level features extraction. Video segmentation process includes shot boundary detection, key frame extraction, and scene segmentation. Feature extraction process includes extracting static features of key frames, object features, and motion features. In general, content-based video retrieval is specified for extracting low-level features from video [9], as shown in **Figure 2**. The indexed video with low-level features can meet only the query by example, where this query is supplied with a frame, image, or a sketch to get a target video.

### 2.2. Semantic Gap

Indexing videos with low-level features cannot meet most of the user s needs and requirements (query). User query can be either queried by example or text. Indexing videos using low-level features can answer queries by example only. However, there is a need to cover user's needs and views that are usually represented in user text query. Text query can contain unlimited number of high-level concepts; and video retrieval should cover the semantic gap between user views and requirements, and video low-level features. Therefore, new semantic methods in video retrieval have been developed to converge user requirements with low-level features of the video. These semantic methods are responsible for extracting the high-level features from video.

### 2.3. Concept Detectors

Due to the inability of the content-based retrieval approach to cover the semantic gap, some approaches try to detect some semantic concepts in a specific field or domain. These approaches include detecting sunsets [10], indoor and outdoor [11], etc [12]. However, these tailored methods can not support the plethora of concepts. Therefore, there is a must to emerge large scale concept detection [12]. The process of creating generic large-scale concept detectors faces some challenges such as:

1) There are infinite number of high-level concepts that are found in user perspectives, and no way to construct concept detectors for this huge number of high-level concepts.

2) Constructing a concept detector is an expensive process, consumes a huge time, and requires many steps to follow such as data preparation step (usually these data are manually annotated), low-level features extraction, building a machine learning classifier, and training the classifier.

Based on [13] [14], a limited number of reliable concept detectors should be

constructed due to its high construction cost. Alexander Hauptmann *et al.* [14] concluded that the video retrieval systems using few thousands of concept detectors are performing well, even that the individual concept detectors are low detection accuracy.

In [15], authors indicated how to select those concepts set. They employ information theoretic notion such as mutual information and pointwise mutual information to determine which concepts are helpful for retrieving the relevant shots to queries and concluded that the frequent concepts only are helpful for different queries and the rare concepts can't help and 90% of the concepts are infrequent. In addition, the concept detectors can't be built for rare concepts because statistical learning algorithms require a large number of training examples. And the frequent concepts appear in most of the shots. It is needed a limited number of concept detectors to retrieve certain shots. But this paper leaves an open question which is "what are the possible solutions for concept selection problem?".

In [16], the most existing content-based video retrieval (CBVR) systems are now amenable to support automatic low-level feature extraction, but they still have limited effectiveness from a user s perspective because of the semantic gap. The automatic video concept detection via semantic classification is one promising solution to bridge the semantic gap. A novel multi-modal boosting algorithm is proposed by incorporating feature hierarchy and boosting to reduce both the training cost and the size of training samples significantly.

In [17], authors propose constructing large-scale concept ontology for multimedia (LSCOM). In LSCOM, hundreds of concepts have been annotated and released. The LSCOM achieves a set of criteria such as utility, coverage, observability, and feasibility. LSCOM experts examined several multimedia vocabularies such as MPEG7 (moving picture experts group), TV-Anytime (Tennessee valley advertising federation), eastern stream center on resources and training 2.4 (Escort 2.4), Thesaurus of Graphical Material, etc. But most of these multimedia vocabularies didn't receive a great attention because they aren't suitable for multimedia tagging and didn't achieve the previously mentioned criteria. With the release of the LSCOM (Large Scale Concept Ontology for Multimedia), a lot of concept detectors have been developed that they can detect objects (e.g. car, people), scenes (e.g. office, outdoor) and events (e.g. walking and marching). These concept detectors are SVM classifiers trained on visual features e.g. color histograms, edge orientation histogram, scale-invariant feature transform (SIFT) descriptors, etc [2]. Relying on concept detectors, semantic video retrieval methods have been developed. As the concept detectors play a vital role in video and image retrieval process, the annual benchmarking event NIST TRECVID (national institute of standards and technology text retrieval conference & video retrieval evaluation) is held to participate in developing the search and evaluation process of concept detectors. Due to the causation of high cost of the manual annotation, the TRECVID event selects 10 - 20 concepts every year for evaluation. 10 - 20 concepts are not sufficient for video retrieval process, when-

ever thousands of concept detectors should be constructed to give an accurate video retrieval search result. In addition, some large-scale concept detectors have been developed such as Mediamill-101, Columbia374, and VIREO374 [8].

In [18], the authors explored two key problems for classifier adaptation: 1) how to transform the existing classifier(s) into an effective classifier for a new dataset with a limited number of labeled examples, and 2) how to select the best existing classifier(s) for adaptation.

Based on [19], the authors further proposed an approach for predicting the negative transfer of a concept classifier to a different domain given the observed parameters. Experimental results show that the prediction accuracy of over 75% can be achieved when transferring concept classifiers learned from LSCOM (news video domain). In [20], this paper tackled the late fusion process, which it combined many classifiers to produce a better one to detect concepts in videos. The paper applied its solutions on TRECVID 2011 Semantic Indexing task.

## 2.4. Semantic Video Retrieval

Content-based retrieval has proven their limitations in solving the semantic gap. Semantic video retrieval tries to bridge this gap using the contextual relations between concepts to deduct the existence of new concepts that haven't a detector. Semantic-based retrieval methods try to cover this gap by pooling a set of concepts and form their inter-relationships, which called context information. These relationships can be constructed by ontology, rules, etc. Some of the concepts are detected using the concept detectors that are previously mentioned. Although there are a limited number of concept detectors, there are infinite numbers of concepts in our world represented in user queries. Thus, modeling the semantic concepts relationships is urgent for discovering the new semantic concepts and it is important for refining the concept detectors scores by enhancing or refuting them. Semantic video retrieval methods are categorized and explained in the next subsections.

### 2.4.1. Graphical Models

In this section, the inter-relationships between video concepts are modeled into graphical models. Also, the relationships between the features and their concepts can be modeled graphically to enhance concept detection. The graphical models can be classified as follows.

1) Bayesian Network

This approach is concerned with modeling the relationships between concepts or high-level semantics as BN or DBN. In [21], authors identified the highlight events in the soccer video including goal event, corner kick event, penalty kick event, and card event. The proposed semantic analysis is frame-based instead of shot-based. Also, it introduced high-level, semantics-based content description analysis for reliable media access and navigation service based on the DBN. It introduced a so-called temporal intervening network to improve the accuracy of the semantic analysis as well. The most factor distinguishes this research is add-

ing the temporal intervening network to DBN to improve the semantic interpretation accuracy.

2) Hidden Markov Model

This approach models the stochastic structure of the high-level concepts as a set of HMMs. In [22], a number of algorithms are presented to classify and segment the soccer video. It was based on two defined semantics elements, play and break. Then, it described the observations of soccer game and according to the observations the features set are selected. At last, the segmentation and classification are made by HMM (hidden Markov models) followed by dynamic programming. The statistical analysis has proven that the classification accuracy is about 83.5%.

3) Factor Graph

This approach models the interaction between concepts as a factor graph. In [23], authors modeled the stochastic relationships between different concepts features as a factor graph. Factor graph has the ability to support multiple modalities and the fusion of features. The product sum algorithm was used to enhance the concept detection accuracy. It proved that the factor graph can handle the stochastic relationships between features extracted from the multi-modality. The detection performance is improved more than 22%.

4) Graph Diffusion

A graph diffusion technique is used to refine the detection scores or the binary annotations by discovering the context and relationships between the concepts [24]. This semantic graph represents concepts as nodes and edges weights reflect concept correlations. The diffusion process is used to recover the relationships between scores according to concept affinities. In addition, this approach adapts domain change of test data by extending semantic diffusion and called domain adaptive semantic diffusion. Using SD, MAP gain is 14% for the 81 concepts on the NUS-WIDE [25], and 11.8% to 15.6% for the 20 concepts on TRECVID 2005-2007. Some concepts have AP drop, as the detector quality of the contextual concepts is poor (is not good enough). DASD improves the performance for TRECVID 2006 and TRECVID 2007, but there is no improvement in TRECVID 2005 and NUS-WIDE, which is due to the fact that there is no domain change in both data. DASD could be a useful technique, if there is a shift expected to happen between training and test data. This approach relies on good performance detectors of the target concepts and the contextual concepts. But if the detector has a low performance, there may be a drop in AP value. DASD performs better than the previous works in [4] [26], and [27]. Finally, the semantic graph diffusion is an undirected weighted graph, but in some cases the contextual relationships between concepts are directional. For more accuracy, a directed graph should be used in the diffusion process to handle the directional relationships between concepts.

#### 2.4.2. Rules

1) Manual Rules

In [3], authors hybrid the manual rules with machine learning techniques to

detect specific events in the soccer game, basketball, and Australian football. This approach relies on that the occurrence of audiovisual features in play-break segments are remarkable patterns for several events.

2) Association Rules

In [7], authors try to exploit the inter-concept association relationships based on concept annotation of video shots, which discovering the hidden association between concepts. These association rules are generated using the apriori algorithm and are used to improve the detection accuracy of concept detectors using a combined ranking scheme. The combined ranking scheme integrates the detection scores of the associated concept detectors according to the association rules to infer the presence of the implied concept and to re-rank the shots. This paper explored several statistical measurements for testing whether temporal dependence among neighboring shots is statistically significant. It excluded that there is a temporal dependency between shots for different concepts where the temporal distance ranged from 1 to 20 for different concepts (such as sports, weather, maps and explosion). Thus, it smoothed the prediction of a shot with respect to a concept by a weighted combination of the inference values of its neighboring shots. Experiments on the TRECVID 2005 dataset show that the proposed framework is both efficient and effective in improving the accuracy of semantic concept detection in video.

3) First-Order Logic Rules

In [28], authors propose a framework for semantic event annotation that constructing an ontology model, referred to as pictorially enriched ontology model. This ontology includes concepts and their visual descriptors, and a method to learn a set of first-order logic rules that describing events. The proposed learning method is an adaptation of the first order inductive learner technique (FOIL). The framework improves the precision and recall, but it is tested on a limited set of concepts such as airplane flying, airplane takeoff, airplane landing, and airplane taxiing. This method needs supporting techniques to learn constants and function symbols.

4) Fuzzy Temporal Horn Logic

In [29], authors propose an approach for video annotation and retrieval based on ontologies and concept detectors. Its ontology is based on the semantic linguistic relations between concepts using wordnet. A rule-based method is used for the automatic semantic annotation for complex events. The rules are constructed using semantic web rules language (SWRL). Finally, it develops a web search engine that depends on ontologies and allows queries using a composition of Boolean and temporal relations. This paper can be improved by generating automatic-rules from the ontology. In addition, it will be more effective and efficient if the rules can handle the uncertainty nature of the semantic detectors, and using the fuzzy temporal Horn logic to overcome the limitations of (SWRL).

### 2.4.3. Ontology

In [30], it proposes ontology enriched semantic space (OSS), which is a compact

semantic space by selecting bases concepts. The concept space is constructed by selecting bases concepts. The concept space is clustered, and the concepts which are near the clusters centroids are selected and called the base concepts. Each basis is arranged to cover an approximately equal portion of subspace in OSS. In this approach, a query Q is projected to OSS space after that one or multiple clusters are selected and the clusters provide information on how to fuse multi-modality features. The statistics co-occurrence between concepts can be used to improve the performance and get more accurate results. Also, the negative relation between concepts can be used to reduce the search space.

In [31], it exhibits a complete scientific depiction plot for semantic video ontology. It is Unique from most existing video ontologies; the proposed ontology covers the three key elements of a formal ontology definition, *i.e.*, concept lexicon, concept properties, and relations among concepts. Likewise, it has understood a video ontology by utilizing a subset of LSCOM concepts as lexicon, utilizing modality weights as properties. Furthermore, it utilizes simultaneousness and hierarchy as relations. The ontology is tested over TRECVID 2005 corpus and the performance is proved over all the concepts. The proposed mathematical video ontology can be enriched with more properties and high order relations. It can be enhanced using learning algorithms.

In [5], this paper builds a context linear space which modeling the relationships between concepts. The first step removes the redundant concepts. After that, it constructs a relationship matrix which models concepts relationships by applying a spectral composition on the relationship matrix, and then the context is getting orthogonal. Then, the similarity between concepts is measured directly by getting the cosine similarities between concepts on context space. If there is a new concept not found in the context space, then this new concept is projected on the context space and the similarity between the target concept and the other concepts is measured. The highest similarities top-k concepts are selected and fused to measure the new concept in the video. This approach shows improvements in performance that is ranged from 2.8% to 38.8% on annotated data. This method can be extended to include negative and positive concept detectors in measuring the existence of the target concept. Context spaces that were learned from manual annotations and concept detection can be fused to generate a new measure for the target concept.

#### 2.4.4. Temporal

This approach tries to solve the semantic gap in video retrieval by detecting the temporal relations between video shots. However, each shot has temporal relations with the adjacent shots such as the previous and the next shots. Temporal relations make the video as a story context. Some methods have addressed the use of these temporal relationships as follows.

In [32], it presents CBCF method called temporal-spatial node balance algorithm (TSNB), which refining concepts detection scores using concept fusion task. The concept fusion task depends on concepts spatial and temporal rela-

tionships. Spatial concept relationships consider the concepts relationships in the same shot. In contrast, spatial relationships consider the relationships between consequent shots. This method is based on the physical model which considering the concepts as nodes, the relationships as forces. And the spatial and temporal relations will be balanced with the moving costs of nodes. The fusion results are defined as the steady balanced status of the whole node system. The relations among concepts can be either positive or negative. A negative relation means two concepts are mutually exclusive. When the relation is positive, it is an attractive force, when negative, it is a repulsive force. The algorithm is tested on TRECVID 2005-2010, and about 75% tested concepts are improved.

In [33], the authors proposed a new framework to encode/compress the huge amount of video data without loss of important information, where it has the ability to reduce the high-dimensional video data into a single stream of numbers (one-dimensional vector) without loss of important information. In addition, this framework has the ability to predict the behavior of neighboring shots and missing shots using the temporal prediction rules. These temporal prediction rules are generated using SPADE algorithm and have the ability to predict neighboring shots, the number of which may be 10 or more, according to the maximum window size parameter value in the SPADE algorithm. This framework is distinguished by revealing the temporal relations between the successive shots in the video. But these temporal rules are incomprehensible to non-specialized users.

#### 2.4.5. Fuzzy

In [34], this paper proposed a method to improve semantic concept detection using fuzzy ontology. The semantic concept/context information is extracted from the annotated data. A fuzzy ontology is constructed using fuzzy description logic to handle the uncertainty of contextual data. An abduction engine is used to extract more rules within concepts and context. The precision of improvement using LSCOM ontology is 11%. However, the improvement in semantic concept detection is 21% via its constructed fuzzy ontology. The recall is improved about 2% for only 5 out of 17 concepts. The improvement of the precision has declined. The proposed method should include others knowledge sources and tested on large scale of concepts. In [35], this paper proposed an approach called concept-driven multi-modality fusion. It maps a multi-modality query to a large number of semantic concepts instead of a query class, and uses the selected concepts to determine the fusion weights. The fusion method is divided into two stages: query-to-concepts mapping and context modeling. In query-to-concept mapping, a random walk process is used to determine relevancies of concepts-to-query. The second stage process these relevancies are transformed into a fusion weights, thorough a fuzzy transformation with a relation matrix. The proposed approach doesn't produce excellent performance for all queries.

### 2.4.6. Recent Approaches

1) Zero-Shot

Zero-shot approach [36]-[42] is perfectly suited for tasks such as video retrieval because of its computational efficiency. This approach applies a predefined set of concept detectors (or concept bank) to a database of videos and uses a mapping function to measure the similarity of the query with the set of concepts detectors in semantic space. After that, a ranked list of videos is returned according to the semantic similarity of the query to obtain an ordered list of videos.

Dalton *et al.* [38] propose a zero-shot retrieval framework which converts the textual description of the event into a query. It identified the concepts that are related to the text query by relying on the automatic speech recognition (ASR), optical character recognition (OCR), and high-level object and visual concepts from videos. A language model is created for each visual concept from a large text source such as web, this process called concept expansion process. The retrieved concepts are used to construct a weighted concept query to retrieve a ranked list of videos. The video data from the TRECVID 2012 Multimedia Event Detection (MED) are used for evaluation. The authors ensure that the recall-oriented measures perform poorly when the extracted text (OCR) and the recognized speech (ASR) are considered independently. 80% of the low recall-measures of the videos have no text content. It is noticeable that concept based retrieval has higher recall evidence by the higher MAP and lower MDFA04 scores. In addition, the fusion between the different modalities provides gains at high precision, 24% improvement in MAP over the single best run (Concepts). In terms of dealing with new concepts, there must be further improvements.

In [39], the authors propose a state of art system search engine for video event search without any submitted example videos to the query which is called zero-example or 0Ex. The system consists of video semantic indexing component, semantic query generation component, multimedia search, and pseudo-relevance. Video semantic indexing is off-line indexing which extracts the semantic features from input videos and indexes them. It extracts the low-level features from the video. After that, they are given to the off-shelf detectors to extract the high-level features. The high-level features include the visual/audio concepts. This component uses Wordnet and Wikipedia to convert user query to system query. Multi-modal search component retrieves a ranked list for each modality, and then the pseudo relevance refines the list. The concepts in these systems are represented as a multi-modal document which includes name, description, category, reliability, and examples of the top detected video snippet. This system is evaluated on TRECVID 2014, and it achieved the best performance according to [43], as it depends on thousands of trained concept video detectors. On the other hand, building concept detectors in this system cost 1.2 million CPU core hours. The process transferring the user query to automatic system query needs more improvement, is not understood, and can't bridge the semantic gap.

Damianos *et al.* [42] propose a zero-example event detection system (MED). This system is based on the concept expansion process which increases the concepts definitions using large external sources such as Wikipedia. The proposed system took the textual description of the event (query) and retrieved the most relevant videos. It was divided into three stages. The first stage builds concepts vector which represents the textual event query that is called Event detector. The second one is concerned with calculating the videos model vectors. Finally, the model-vectors are compared with the concepts-vector of the event query using the similarity measures, and then the most relevant videos are returned. Also, a statistical selection method is proposed to determine the suitable number of concepts k for representing the event query. It orders the event concept vector scores in descending order, constructing an exponential curve, so that then selects the first k concepts. This paper uses a large concept pool that consists of 13.488 semantic concepts, the MAP of the proposed system is slightly better than the other systems such as [39]. A real-time video retrieval can be applied to this approach because the concept detection scores are previously measured to the database of videos and the semantic similarity computing is fast. Zero-shot/Zero-example approach strives to be suitable for the queries involving concepts that are far from concept bank.

2) On-Fly-Retrieval Approach

On-the-fly methods get their training data after submitting the query (for example, from the web), rather than relying on concept detectors that have been applied on the video database. These methods make it easy to submit arbitrary queries that are not limited to a concept bank and the retrieved videos are not limited to previously saving a video database. Moreover, with these modern computing architectures such as the GPU, these methods can achieve real-time performance [44]. One of the on-fly-retrieval methods is presented in [45]. A real-time video retrieval system is presented which represents the video database and images using a fisher vector built on CNN. The on-fly-retrieval approach doesn't depend on a pre-trained data; it obtains the training data after accepting the text query from the user. For each text query, the web search returns the relevant images to query. After that, the videos in the videos database are ranked according to measuring the similarity between the fisher vectors of the web images and fisher vectors of the videos. The visual representation of the query is rebuilt according to the top selected similar videos. Finally, re-rank the videos according to the new visual representation of the query. The disadvantage of this approach is treating the videos and the web images as unordered image sets which don't extract the features of single image or frame but extracting a compact vector of features to the whole images/frames as a whole.

## 3. Conclusion and Future Work

There are many unresolved issues in semantic video retrieval such as:

- There is still a research gap for the task of selecting the appropriate features

for the semantic concept detection.

- Improving the performance of concept detectors still needs a lot of future work. This can be partially achieved by improving the relevance feedback techniques.
- Constructing a large-scale concept ontology for video retrieval.
- There is a need to construct an automatic tool which has the ability to translate the high-level concepts that were extracted from the user text query to the suitable concept detector.
- Different machine learning approaches can be fused to obtain more accurate concepts.
- Improving concept detectors performance is still a challenging problem especially for rare concepts.
- Constructing more techniques doesn't depend on prior domain knowledge.
- Constructing more techniques benefits from temporal consistency in video.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Yang, J. and Hauptmann, A.G. (2006) Exploring Temporal Consistency for Video Analysis and Retrieval. *Proceedings of the* 8*th ACM International Workshop on Multimedia Information Retrieval*, Santa Barbara, 26-27 October 2006, 33-42. https://doi.org/10.1145/1178677.1178685

[2] Aytar, Y., Shah, M. and Luo, J.B. (2008) Utilizing Semantic Word Similarity Measures for Video Retrieval. 26*th IEEE Conference on Computer Vision and Pattern Recognition*, *CVPR*, 24-26 June 2008, Anchorage.

[3] Tjondronegoro, D.W. and Chen, Y.P.P. (2010) Knowledge-Discounted Event Detection in Sports Video. *IEEE Transactions on Systems*, *Man*, *and Cybernetics—Part A*: *Systems and Humans*, **40**, 1009-1024. https://doi.org/10.1109/TSMCA.2010.2046729

[4] Weng, M.-F. and Chuang, Y.-Y. (2008) Multi-Cue Fusion for Semantic Video Indexing. *Proceedings of the* 16*th ACM International Conference on Multimedia*, Vancouver, 27-31 October 2008, 71-80.

[5] Wei, X.-Y., Jiang, Y.-G. and Ngo, C.-W. (2009) Exploring Inter-Concept Relationship with Context Space for Semantic Video Indexing. *Proceeding of the ACM International Conference on Image and Video Retrieval*, Fira, 8-10 July 2009, 15:1-15:8. https://doi.org/10.1145/1646396.1646416

[6] Jiang, Y.G., Wang, J., Chang, S.F. and Ngo, C.W. (2009) Domain Adaptive Semantic Diffusion for Large Scale Context-Based Video Annotation. *Proceedings of the IEEE International Conference on Computer Vision*, Kyoto, 29 September-2 October 2009, 1420-1427.

[7] Liu, K.H., Weng, M.F., Tseng, C.Y., Chuang, Y.Y. and Chen, M.S. (2008) Association and Temporal Rule Mining for Postfiltering of Semantic Concept Detection in Video. *IEEE Transactions on Multimedia*, **10**, 240-251. https://doi.org/10.1109/TMM.2007.911826

[8] Jiang, Y.-G., Yanagawa, A., Chang, S.-F. and Ngo, C.-W. (2008) CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection. Technical Report, Columbia University, New York.

[9] Hu, W.M., Xie, N.H., Li, L., Zeng, X.L. 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**, 797-819. https://doi.org/10.1109/TSMCC.2011.2109710

[10] Smith, J.R. and Chang, S.-F. (1997) Visually Searching the Web for Content. *IEEE MultiMedia*, **4**, 12-20. https://doi.org/10.1109/93.621578

[11] Szummer, M. and Picard, R.W. (1998) Indoor-Outdoor Image Classification. *Proceedings* 1998 *IEEE International Workshop on Content-Based Access of Image and Video Database*, Bombay, 3 January 1998, 42-51.

[12] Snoek, C.G.M. and Worring, M. (2007) Concept-Based Video Retrieval. *Foundations and Trends® in Information Retrieval*, **2**, 215-322.

[13] Hauptmann, A., Yan, R. and Lin, W.-H. (2007) How Many Highlevel Concepts Will Fill the Semantic Gap in News Video Retrieval? *Proceedings of the* 6*th ACM International Conference on Image and Video Retrieval*, Amsterdam, 09-11 July 2007, 627-634. https://doi.org/10.1145/1282280.1282369

[14] Hauptmann, A., Yan, R., Lin, W.-H., Christel, M. and Wactlar, H. (2007) Can High Level Concepts Fill the Semantic Gap in Video Retrieval? A Case Study with Broadcast News. *IEEE Transactions on Multimedia*, **9**, 958-966. https://doi.org/10.1109/TMM.2007.900150

[15] Lin, W.-H. and Hauptmann, A. (2006) Which Thousand Words Are Worth a Picture? Experiments on Video Retrieval Using a Thousand Concepts. 2006 *IEEE International Conference on Multimedia and Expo*, Toronto, 9-12 July 2006, 41-44.

[16] Fan, J., Luo, H., Gao, Y. and Jain, R. (2007) Incorporating Concept Ontology for Hierarchical Video Classification, Annotation, and Visualization. *IEEE Transactions on Multimedia*, **9**, 939-957. https://doi.org/10.1109/TMM.2007.900143

[17] Naphade, M., Smith, J.R., Tesic, J., Chang, S.-F., Hsu, W., Kennedy, L., Hauptmann, A. and Curtis, J. (2006) Largescale Concept Ontology for Multimedia. *IEEE Multimedia*, **13**, 86-91. https://doi.org/10.1109/MMUL.2006.63

[18] Yang, J., Yan, R. and Hauptmann, A.G. (2007) Cross-Domain Video Concept Detection Using Adaptive SVMS. *Proceedings of the* 15*th ACM International Conference on Multimedia*, Augsburg, 24-29 September 2007, 188-197.

[19] Yao, T., Ngo, C.-W. and Zhu, S. (2012) Predicting Domain Adaptivity: Redo or Recycle? *Proceedings of the* 20*th ACM International Conference on Multimedia*, Nara, 29 October-2 November 2012, 821-824.

[20] Strat, S.T., Benoit, A., Lambert, P., Bredin, H. and Quenot, G. (2014) Hierarchical Late Fusion for Concept Detection in Videos. In: *Fusion in Computer Vision*, Springer, Berlin, 53-77. https://doi.org/10.1007/978-3-319-05696-8_3

[21] Huang, C.-L., Shih, H.-C. and Chao, C.-Y. (2006) Semantic Analysis of Soccer Video Using Dynamic Bayesian Network. *IEEE Transactions on Multimedia*, **8**, 749-760. https://doi.org/10.1109/TMM.2006.876289

[22] Xie, L., Chang, S.F., Divakaran, A. and Sun, H. (2002) Structure Analysis of Soccer Video with Hidden Markov Models. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, 13-17 May 2002, Vol. 4, 4096-4099. https://doi.org/10.1109/ICASSP.2002.5745558

[23] Naphade, M., Kozintsev, I., Huang, T. and Ramchandran, K. (2000) A Factor Graph

Framework for Semantic Indexing and Retrieval in Video. *Proceedings Workshop on Content-Based Access of Image and Video Libraries*, Hilton Head Island, 12 June 2000, 35-39. https://doi.org/10.1109/IVL.2000.853836

[24] Jiang, Y.-G., Dai, Q., Wang, J., Ngo, C.-W., Xue, X. and Chang, S.-F. (2012) Fast Semantic Diffusion for Large-Scale Context-Based Image and Video Annotation. *IEEE Transactions on Image Processing*, **21**, 3080-3091. https://doi.org/10.1109/TIP.2012.2188038

[25] Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z. and Zheng, Y.-T. (2009) Nus-Wide: A Real-World Web Image Database from National University of Singapore. *Image and Video Retrieval*, Santorini, 8-10 July 2009.

[26] Jiang, W., Chang, S.-F. and Loui, A.C. (2007) Context-Based Concept Fusion with Boosted Conditional Random Fields. *ICASSP*, Honolulu, 15-20 April 2007, 949-952. https://doi.org/10.1109/ICASSP.2007.366066

[27] Aytar, Y., Orhan, O.B. and Shah, M. (2007) Improving Semantic Concept Detection and Retrieval Using Contextual Estimates. *IEEE International Conference on Multimedia and Expo*, 2-5 July 2007, 536-539. https://doi.org/10.1109/ICME.2007.4284705

[28] Bertini, M., Del Bimbo, A. and Serra, G. (2008) Learning Ontology Rules for Semantic Video Annotation. *ACM International Conference on Multimedia Many Faces of Multimedia Semantics MS*, 1-8.

[29] Ballan, L., Bertini, M., Del Bimbo, A. and Serra, G. (2010) Video Annotation and Retrieval Using Ontologies and Rule Learning. *IEEE MultiMedia*, **17**, 80-88. https://doi.org/10.1109/MMUL.2010.4

[30] Wei, X.-Y., Ngo, C.-W. and Jiang, Y.-G. (2008) Selection of Concept Detectors for Video Search by Ontology-Enriched Semantic Spaces. *IEEE Transactions on Multimedia*, **10**, 1085-1096. https://doi.org/10.1109/TMM.2008.2001382

[31] Zha, Z.-J., Mei, T., Wang, Z. and Hua, X.-S. (2007) Building a Comprehensive Ontology to Refine Video Concept Detection. *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, *ACM*, New York, 227-236. https://doi.org/10.1145/1290082.1290114

[32] Geng, J., Miao, Z. and Chi, H. (2013) Temporal-Spatial Refinements for Video Concept Fusion. Springer, Berlin Heidelberg, 547-559. https://doi.org/10.1007/978-3-642-37431-9_42

[33] Toriah, S.T.M., Ghalwash, A.Z. and Youssif, A.A.A. (2018) Shots Temporal Prediction Rules for High-Dimensional Data of Semantic Video Retrieval. *American Journal of Applied Sciences*, **15**, 60-69. https://doi.org/10.3844/ajassp.2018.60.69

[34] Elleuch, N., Zarka, M., Ammar, A.B. and Alimi, A.M. (2011) A Fuzzy Ontology: Based Framework for Reasoning in Visual Video Content Analysis and Indexing. *Proceedings of the* 11*th International Workshop on Multimedia Data Mining*, *ACM*, New York, Article No. 1.

[35] Wei, X.Y., Jiang, Y.G. and Ngo, C.W. (2011) Concept-Driven Multi-Modality Fusion for Video Search. *IEEE Transactions on Circuits and Systems for Video Technology*, **21**, 62-73. https://doi.org/10.1109/TCSVT.2011.2105597

[36] Chen, J., Cui, Y., Ye, G., Liu, D. and Chang, S.-F. (2014) Event-Driven Semantic Concept Discovery by Exploiting Weakly Tagged Internet Images. *Proceedings of International Conference on Multimedia Retrieval*, *ACM*, New York, 1.

[37] Cui, Y., Liu, D., Chen, J. and Chang, S.-F. (2014) Building a Large Concept Bank for Representing Events in Video.

[38] Dalton, J., Allan, J. and Mirajkar, P. (2013) Zero-Shot Video Retrieval Using Content and Concepts. *Proceedings of the* 22*nd ACM International Conference on Information & Knowledge Management*, *ACM*, New York, 1857-1860.

[39] Jiang, L., Yu, S.I., Meng, D., Mitamura, T. and Hauptmann, A.G. (2015) Bridging the Ultimate Semantic Gap: A Semantic Search Engine for Internet Videos. *Proceedings of the* 5*th ACM on International Conference on Multimedia Retrieval*, *ACM*, New York, 27-34. https://doi.org/10.1145/2671188.2749399

[40] Mazloom, M., Li, X. and Snoek, C.G.M. (2016) Tagbook: A Semantic Video Representation without Supervision for Event Detection. *IEEE Transactions on Multimedia*, **18**, 1378-1388. https://doi.org/10.1109/TMM.2016.2559947

[41] Wu, S., Bondugula, S., Luisier, F., Zhuang, X. and Natarajan, P. (2014) Zero-Shot Event Detection Using Multi-Modal Fusion of Weakly Supervised Concepts. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 2665-2672. https://doi.org/10.1109/CVPR.2014.341

[42] Galanopoulos, D., Markatopoulou, F., Mezaris, V. and Patras, I. (2017) Concept Language Models and Event-Based Concept Number Selection for Zero-Example Event Detection. *Proceedings of the ACM on International Conference on Multimedia Retrieval*, *ACM*, New York, 397-401. https://doi.org/10.1145/3078971.3079043

[43] Over, P., Fiscus, J., Sanders, G., Joy, D., Michel, M., Awad, G., Smeaton, A., Kraaij, W. and Quenot, G. (2014) Trecvid 2014—An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. *Proceedings of TRECVID*, Orlando, November 2014, 52.

[44] Chatfield, K., Simonyan, K. and Zisserman, A. (2014) Efficient on the Fly Category Retrieval Using ConvNet and GPUs. In: *Asian Conference on Computer Vision*, Springer, Berlin, 129-145.

[45] Han, X., Singh, B., Morariu, V. and Davis, L.S. (2017) VRFP: On-the-Fly Video Retrieval Using Web Images and Fast Fisher Vector Products. *IEEE Transactions on Multimedia*, **19**, 1583-1595. https://doi.org/10.1109/TMM.2017.2671414

## Nomenclature

0Ex = Zero example
AP = Average precision
ASR = Automatic speech recognition
BN = Bayesian network
CBVR = Content-based video retrieval
DBN = Dynamic Bayesian network
FOIL = First order inductive learner
LSCOM = Large scale concept ontology for multimedia
MAP = Mean average precision
MPEG7 = Moving picture experts group
NIST = National institute of standards and technology
OCR = Optical character recognition
SIFT = Scale-invariant feature transform
SVM = Support vector machine
SWRL = Semantic web rules language
TRECVID = Text retrieval conference & video retrieval evaluation
TSNB = Temporal-spatial node balance algorithm
TV-Anytime = Tennessee valley advertising federation