Semantic-Based Video Retrieval Survey

Shaimaa Toriah[1], Atef Ghalwash[2], and Aliaa Youssif[3]

[1]Teaching Assistant in the Department of Computer Science, Faculty of Computers and Information, Benha University.
[2]Prof. A.Khalwash is A Professor Emeritus,Computer Science Department, Helwan university.
[3]Prof. A.youssef is The Dean, Faculty of computers and information, Helwan university.

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. So, 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, video-on-demand broadcasting. In this paper, the different approaches of video retrieval are clearly and briefly categorized. Moreover, the different methods which try to bridge the semantic gap in video retrieval are discussed in more details.

***Index Terms*—Semantic Video Retrieval,Concept Detectors,Context Based Concept Fusion.**

I. Introduction

DIGITALdata plays an essential role in our life. The digital data includes videos, images, documents, sound and etc. One of the most important digital media is video especially the dynamic one. The video represents a rich source of information. The video can contain all the other digital data such as images, sound, and text. In addition, the video is characterized by its temporal consistency. The rapid progress of digital devices causes inflation in video database. Retrieving the required information from video database according to user needs is called a video retrieval process. A video retrieval is considered a branched field from the globalized one called information retrieval. Information retrieval is considered as a subfield of computer science that is concerned with the organization and retrieval of information from large database collections [2]. Video retrieval methods are important and essential for multimedia applications such as video search engines, digital museums, video-on-demand broadcasting, and etc.

Video retrieval is still an active problem due the semantic gap , and the wide spread of social media and the enormous technological development.To guarantee 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 semantic gap is 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, but it stills uncovered.In this paper, a concise overview on 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 play a vital role in semantic video retrieval, a thorough study of the obstacles that face the construction of generic concept detectors will be presented. Finally, The different methods which model semantic concepts relationships in video retrieval are categorized and explained in more details.

The main contributions of this survey are:

1) Clear categorization of video retrieval approaches.
2) Identifying the semantic gap problem.
3) Discussing the obstacles and challenges facing concept detectors construction.
4) Clearly defining the different methods of semantic video retrieval.
5) Presenting helpful diagrams about video retrieval system and semantic video retrieval methods.

The remainder of the paper is organized as the followings: section2 briefly introduce the video retrieval. Section 2.1 briefly reviews content-based retrieval. Section 2.2 discusses the semantic gap . In Section 2.3, it discusses concept detectors in more details. Section2.4 discusses in details the different methods of semantic video retrieval.

II. Video Retrieval

VIDEO retrieval is concerned with retrieving specific videos shots according to user needs (usually called query). Video retrieval is still an active problem due to the enormous technological development that allows easy video capturing and sharing. To guarantee 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, 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], integration and interrelation of content-based video retrieval and semantic-based video retrieval are shown. In figure[2], it illustrates the different categories and methods of video retrieval.

A. Content-Based Video Retrieval

Content based video retrieval includes video segmentation, and low level features extraction.Video segmentation process

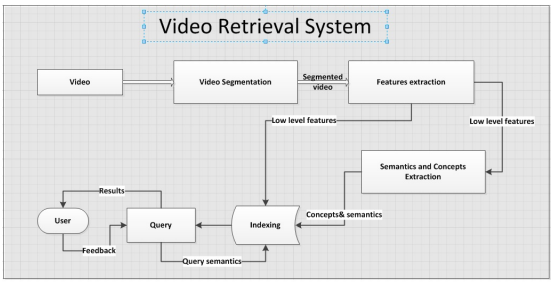

Fig. 1. Video Retrieval Process

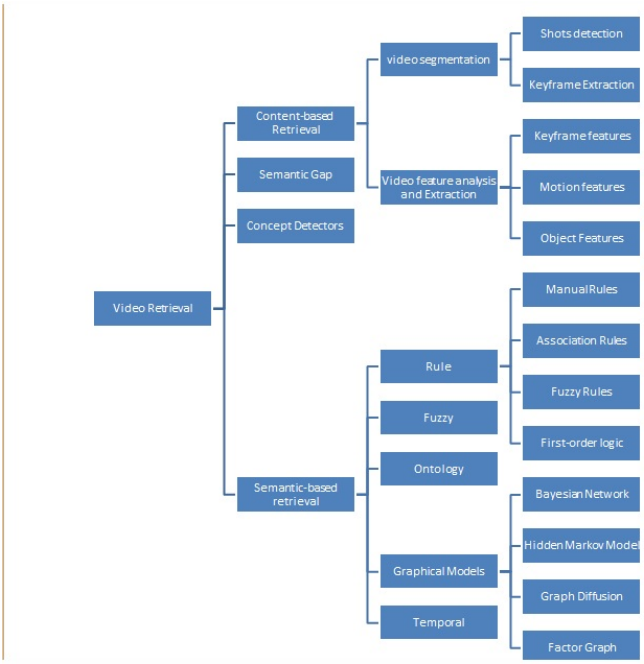

Fig. 2. Video Retrieval categories and methods

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 [11], see figure[2]. The indexed video with low level features can meet only query by example, where this query is supplied with a frame, image, or a sketch to get a target video.

B. Semantic Gap

Indexing the videos with low level features cannot meet most of the user needs and requirements (query). User query can be either query by example or a text query. 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.

C. Concept Detectors

Due to the inability of the content based retrieval approach to cover the semantic gab, some approaches try to detect some semantic concepts in a specific field or domain. These approaches include detecting sunsets [18], indoor and outdoor [8], etc [19]. However, these tailored methods can not support the plethora of concepts. therefore,there is a must to emerge large scale concept detection [19]. But, to create generic large scale concept detectors, there are some problems that represent as challenges such as:

1) There are an infinity number of high level concepts that are found in user perspectives, and no way to construct concepts 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 [10] [9], a limited number of reliable concept detectors should be constructed due its high construction cost. In [9], it concludes that the video retrieval systems that use few thousands of concept detectors are performing well , even that the individual concept detectors are low detection accuracy.

In [5],it indicates how to select those concepts set? This paper employs information theoretic notion such as mutual information and point wise 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 to , the concept detectors can't be built for rare concepts because statistical learning algorithms require 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 a certain shots. But, this paper leaves an open question which is ”what are the possible solutions for concept selection problem?”.

In [4], 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 multimodal 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], it constructs large scale concept ontology for multimedia(LSCOM). In LSCOM, hundreds of concepts have been annotated and released. The LSCOM achieves a set of criterias such as utility, coverage, observability, and feasibility. LSCOM experts examined several multimedia vocabularies

such as MPEG7, TV-Anytime, Escort 2.4, Thesaurus of Graphical Material, etc. But most of these multimedia vocabularies didn't receive a great attention because there aren't suitable for multimedia tagging and didn't achieve the previously mentioned criterias.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, SIFT descriptors, etc" [1]. 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) is hold to participate in developing the search and evaluation process of concept detectors. Due 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, whenever thousands of concept detectors should be constructed to give an accurate video retrieval search result. Also, some large scale concept detectors have been developed such as Mediamill-101, Columbia374, and VIREO374 [13].

In [25], it explores two key problems for classifier adaptation: (1) how to transform existing classifier(s) into an effective classifier for a new dataset that only has a limited number of labeled examples, and (2) how to select the best existing classifier(s) for adaptation.

Based on [26], it further proposes 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 learnt from LSCOM (news video domain) in [**?**], this paper tackle the late fusion process , which it combines many classifiers to produce a better one to detect concepts in videos. The paper applied its solutions on TRECVid 2011 Semantic Indexing task.

D. Semantic Video Retrieval

Content based retrieval has proven their limitations in solving the semantic gap. Semantic video retrieval tries to bridge this gab 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 infinity numbers of concepts in our world represented in user queries, so 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 researches are categorized and explained in the next subsections.

1) Graphical models

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

1) Bayesian Network
   In [12], it identifies the highlight events in 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 introduces high-level, semantics-based content description analysis for reliable media access and navigation service based on the DBN (dynamic Bayesian network). It also introduces a so-called temporal intervening network to improve the accuracy of the semantic analysis. What most distinguishes this research is adding the temporal intervening network to DBN to improve the semantic interpretation accuracy.
2) Hidden Markov Model
   In [24], some algorithms are presented to classify and segment the soccer video. It is based on two defined semantics elements are play and break. Then it describes 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 programing. The statistical analysis has proven that the classification accuracy is about 83.5%.
3) Factor Graph
   [16] proposes a factor graph framework which models the stochastic relationships between different concepts features. The product sum algorithm is used to enhance the concept detection accuracy. It proves that the factor graph can handle the stochastic relationships between features extracted from the multimodality.
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 [14]. This semantic graph represents concepts as nodes and edges weights reflect concepts correlations. The diffusion process is used to recover the relationships between scores according to concept affinities. This approach also adapts domain change of test data by extending semantic diffusion and called domain adaptive semantic diffusion. But the semantic graph diffusion is 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) Rules

1) Manual Rules
   in [20], it hybrids the manual rules with machine learning techniques to detect specific events in soccer game, basketball, and Australian football. This approach

relies on that the occurrence of audiovisual features in a play-break segments are remarkable patterns for several events.

2) Association rules
In [15],it tries to exploit the inter-concept association relationships based on concept annotation of video shots, which discover the hidden association between concepts. These associations rules are generated using the Apiriori algorithm and theses rules 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 concepts detectors according to the association rules to infer the presence of the implied concept and re-rank shots. This paper also explores several statistical measurements for testing whether temporal dependence among neighboring shots are statistically significant .it excludes that there are 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). So, it smoothies 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 [3],it proposes a framework for semantic event annotation that constructs 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 describes events. The proposed learning method is an adaptation of the First Order Inductive Learner technique (FOIL). The framework improves the precision and recall, however it is tested on a limited set of concepts such as airplane flying, airplane takeoff, airplane landing, and airplane taxiing. But, this method needs other techniques to learn constants and function symbols.

4) Fuzzy Temporal Horn Logic
[2],This paper proposes an approach for video annotation and retrieval based on ontologies and concept detectors. Its ontology based on the semantic linguistic relations between concepts using word net. 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. Also, 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).

In [21], this paper builds a context linear space which models 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; this new concept is projected on the context space and measure the similarity between the target concept and the other concepts. The highest similarities top-k concepts are selected and fused to measure the new concept in the video. This approach shows improvements of 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 detections can be fused to generate a new measure for the target concept.

3) Ontology

In [23], 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 bases 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.

In [27],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.

4) Temporal

In [7],it presents CBCF method called temporal spatial node balance algorithm (TNSB), which refines concepts detection scores using concept fusion task. The concept fusion task depends on concepts spatial and temporal relationships. Spatial concepts relationships consider the concepts relationships in the same shot, but spatial relationships consider the relationships between consequent shots. This method is based on the physical model which considers 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. 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 about 75% tested concepts are improved.

5) Fuzzy

In [6],this paper proposes a new 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 [22],This paper proposes an approach called concept-driven multimodality fusion. It maps a multi-modality query to the large number of semantic concepts instead of a query class, and use 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.

III. Conclusion and Future

The video is considered a rich source of information, because the video can contain others multimedia data such as audio, images, and text. Moreover, it is distinguished by its temporal consistency. The videos are spread and used in all aspects of life .Video retrieval represents a vital process in many recent applications. But, indexing videos with low level features can't cover the user high level concepts, which is called semantic gap. Enormous methods have been proposed to bridge this semantic gap. Bridging semantic gap relies on concept detectors, but building a concept detector costs too much due to the heavy cost of manual annotation and high computational cost for supervised learning. According to the previous reasons, a few concept detectors have been developed. However, 5000 concepts will be sufficient for accurate video retrieval and there are a limited number of concepts. All semantic video retrieval methods try to conclude the existence of new concepts which havent a detector through the existence of the defined concepts. The approaches that try to cover this gap are called semantic video retrieval methods. The two approaches are divided into semantic content-based-retrieval, and semantic video retrieval. The two approaches are interrelated. Finally, we can conclude that semantic video retrieval and concept based video retrieval are promising approaches to bridge the video retrieval semantic gap.Big data approaches will be a promising approach to model concepts relationships and to overcome the computational processing issue .

References

[1] Yusuf Aytar, Mubarak Shah, and Jiebo Luo. Utilizing semantic word similarity measures for video retrieval. *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008.

[2] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Video annotation and retrieval using ontologies and rule learning. *IEEE MultiMedia*, 17(4):80–88, Oct 2010.

[3] Marco Bertini, Alberto Del Bimbo, and Giuseppe Serra. Learning ontology rules for semantic video annotation. *Proc of ACM International Conference on Multimedia Many Faces of Multimedia Semantics MS*, pages 1–8, 2008.

[4] Video Classification, Jianping Fan, Hangzai Luo, Yuli Gao, and Ramesh Jain. Incorporating Concept Ontology for Hierarchical. 9(5):939–957, 2007.

[5] Ieee International Conference. WHICH THOUSAND WORDS ARE WORTH A PICTURE ? EXPERIMENTS ON VIDEO RETRIEVAL USING A THOUSAND CONCEPTS Wei-Hao Lin and Alexander Hauptmann Language Technologies Institute School of Computer Science Carnegie Mellon University. 2006.

[6] Nizar Elleuch, Mohamed Zarka, Anis Ben Ammar, and Adel M. Alimi. A fuzzy ontology: Based framework for reasoning in visual video content analysis and indexing. In *Proceedings of the Eleventh International Workshop on Multimedia Data Mining*, MDMKDD '11, pages 1:1–1:8, New York, NY, USA, 2011. ACM.

[7] Jie Geng, Zhenjiang Miao, and Hai Chi. *Temporal-Spatial Refinements for Video Concept Fusion*, pages 547–559. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[8] Fei Guo. Indoor Outdoor Image Classification. *Computer*, 2011.

[9] A Hauptmann, R Yan, W-H Lin, M Christel, and H Wactlar. Can high level concepts fill the semantic gap in video retrieval? A case study with broadcast news. *IEEE Transactions on Multimedia*, 9(5):958–966, 2007.

[10] Alexander Hauptmann and Wei-hao Lin. How many highlevel concepts will fill the semantic gap in video retrieval ? pages 627–634, 2007.

[11] W Hu, N Xie, L Li, X Zeng, and S Maybank. 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, nov 2011.

[12] Chung-Lin Huang, Huang-Chia Shih, and Chung-Yuan Chao. Semantic analysis of soccer video using dynamic Bayesian network. *IEEE Transactions on Multimedia*, 8(4):749–760, aug 2006.

[13] Yg Jiang, Yg Jiang, a Yanagawa, a Yanagawa, Sf Chang, Sf Chang, Cw Ngo, and Cw Ngo. CU-VIREO374: fusing Columbia374 and VIREO374 for large scale semantic concept .... *Columbia University ADVENT Technical Report*, 2008.

[14] Yu-gang Jiang, Qi Dai, Jun Wang, and Chong-wah Ngo. Fast Semantic Diffusion for Large-Scale Context-Based Image and Video Annotation. 21(6):3080–3091, 2012.

[15] Ken Hao Liu, Ming Fang Weng, Chi Yao Tseng, Yung Yu Chuang, and Ming Syan Chen. Association and temporal rule mining for post-filtering of semantic concept detection in video. *IEEE Transactions on Multimedia*, 10(2):240–251, 2008.

[16] M Naphade, I Kozintsev, T Huang, and K Ramchandran. A factor graph framework for semantic indexing and retrieval in video. In *2000 Proceedings Workshop on Content-based Access of Image and Video Libraries*, pages 35–39, 2000.

[17] Milind Naphade, John R. Smith, Jelena Tesic, Shih Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.

[18] J R Smith and Shih-Fu Chang. Visually searching the Web for content. *IEEE MultiMedia*, 4(3):12–20, 1997.

[19] Cees G. M. Snoek and Marcel Worring. Concept-Based Video Retrieval. *Foundations and Trends® in Information Retrieval*, 2(4):215–322, 2007.

[20] D W Tjondronegoro and Y P P Chen. Knowledge-Discounted Event Detection in Sports Video. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 40(5):1009–1024, sep 2010.

[21] Xiao-Yong Wei, Yu-Gang Jiang, and Chong-Wah Ngo. Exploring inter-concept relationship with context space for semantic video indexing. *Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 15:1—-15:8, 2009.

[22] Xiao Yong Wei, Yu Gang Jiang, and Chong Wah Ngo. Concept-driven multi-modality fusion for video search. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(1):62–73, 2011.

[23] Xiao-yong Wei, Chong-wah Ngo, and Yu-gang Jiang. Selection of Concept Detectors for Video Search by Ontology-Enriched Semantic Spaces. pages 1–12.

[24] L Xie, S F Chang, A Divakaran, and H Sun. Structure analysis of soccer video with hidden Markov models. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–4096–IV–4099, may 2002.

[25] Jun Yang and Alexander G Hauptmann. Cross-Domain Video Concept Detection Using Adaptive SVMs.

[26] Ting Yao, Chong-wah Ngo, and Shiai Zhu. Predicting Domain Adaptivity : Redo or Recycle ? pages 5–8, 2012.

[27] Zheng-Jun Zha, Tao Mei, Zengfu Wang, and Xian-Sheng Hua. Building a comprehensive ontology to refine video concept detection. In *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, MIR '07, pages 227–236, New York, NY, USA, 2007. ACM.