Automatic Semantic Video Annotation in Wide Domain Videos Based on Similarity and Commonsense Knowledgebases

Amjad Altadmri[1], Amr Ahmed[2]

School of Computer Science, University of Lincoln
Lincoln, UK
[1] atadmri@lincoln.ac.uk
[2] aahmed@lincoln.ac.uk

***Abstract*—In this paper, we introduce a novel framework for automatic Semantic Video Annotation. As this framework detects possible events occurring in video clips, it forms the annotating base of video search engine. To achieve this purpose, the system has to able to operate on uncontrolled wide-domain videos. Thus, all layers have to be based on generic features.**

This framework aims to bridge the *"semantic gap"*, which is the difference between the low-level visual features and the human's perception, by finding videos with similar visual events, then analyzing their free text annotation to find a common area then to decide the best description for this new video using commonsense knowledgebases.

Experiments were performed on wide-domain video clips from the TRECVID 2005 BBC rush standard database. Results from these experiments show promising integrity between those two layers in order to find expressing annotations for the input video. These results were evaluated based on retrieval performance.

I. Introduction

The rapidly increasing amount of video collections, available on the web or via broadcasting, motivated research towards building automatic tools for rating, indexing, searching and retrieval purposes. To achieve the most benefit from these systems they have to satisfy the human's perception and description, but these videos contain low-level visual data. This huge gap between human's perception and binary visual data is referred to as the *"semantic gap"* [1].

In this paper, we propose a novel framework that tries to help minimizing this semantic gap by detecting possible events happening in wide domain video clips. This framework finds events accompanied to objects in order to simulate the human's cognitive manner, as human beings often tend to annotate video information based on semantic events [2], not only by objects. For example, humans rank videos based on violence or sexual activities, not on detecting partially naked persons, which could also be found in some sports.

The definition of the term *"Event"* varies across the literature. While in areas like surveillance it is considered initially as one action such as *"object1 moves from the left to the right"*, these actions form *"Interest events"* in later stages combined with background information like forbidden zones [3]. However, in other areas, the term *"Event"* holds the wide meanings of *concept* (such as car, mountain, walking or sport) [4]; or sometimes it is used as a *shot class* (such as a replay or a break in a football game) [5].

The most general definition of *Event*, which is also called *"Semantic Event"* and matches the linguistic meaning, is presented in [6]. In this definition, which we adopt, *"Event"* is identified as: *"any change happens by object(s) in spatio-temporal space that holds a human meaning"*. For example, *airplane landing, person eating*.

One of the challenges is to be able to operate on wide domain uncontrolled videos. Hence, the proposed framework has to utilize domain-independent approaches. For that purpose, we utilize content-based low-level features volumes' similarity to find similar videos to an input clip, then analyse the accompanied annotations of those similar clips based on commonsense knowledgebases to understand the common area among these videos. Both these approaches based on non-domain-specific tools.

Experiments were performed on video clips from TRECVID 2005 BBC Rushes [7], which is a group of standard databases for information retrieval. The results show promising integrity between the proposed layers in order to find semantically representative annotations for the input video. These results were evaluated based on retrieval performance.

The rest of the paper is organized as follows: In section II, the key related work is discussed. Our proposed framework is presented in section III, while the experiments, results and evaluation are described in section IV. The paper is finally concluded in section V, where future work is also suggested.

II. Previous work

Considerable approaches make use of special features of specific domain videos like shot boundary types and slow movement replays in Cinematic edited videos as they add extra information in the recognition stage [5]. This make these approaches hard to be generalized as the features used are not usually found in other types of videos and their interpretation is related to the studied domain.

In contrast, as our focus is on non-domain-specific techniques, the most related key work methodologies are discussed in the following subsections:

978-1-4244-5561-4/09/$26.00 ©2009

A. State machines

State machines have been built to track the change of positions of specific objects in a video clip [6]. These approaches are considering events as a group of sequenced actions. Hence, they depend on information obtained from many previous layers, such as object detection and classification, motion analysis and motion-blob verification. As these previous layers almost depend on specific domain knowledge, they become difficult to be generalized.

One of the most important ideas in our framework is to build a system which can replace these previous layers without depending on domain knowledge.

B. Query by example

Content Based Retrieval area achieves a good improvement currently, especially on key frames, which is mainly Image Retrieval, in addition to some basic motion information [8].

Other approaches based on spatio-temporal information has achieved some reasonable results over some specific domains [9], [10], but Basharat *et al.* [11] have developed a generic technique for spatio-temporal volumes' features extraction. This method is the inspiration for the first layer in our framework, as presented in section III.

C. Learning

A number of machine learning techniques try to classify video shots under the name of event. A good SVM shots classifier, based on motion feature alone, was suggested in [4]. Bertini *et al.* [12] have built a system that learns events rules from Ontology. Others, like in [13], use association mining techniques to indicate the existence of one high-level concept from the simultaneously existence of other concepts, trying to enhance accuracy of semantic concepts detection.

In [14], decision trees were used to infer high-level concepts from low-level video visual features. Also in [3] a rule learning process, depending on both low-level and middle-level features, build a decision tree to mark rare events and concepts. In this work, they raised the class imbalance problem which faces learning process in *rare* concept/event detection. Class-imbalance means the small percentage of recording real samples of positive interest events compared to the number of negative examples. Moreover, learning usually suits the domain-specific applications more than the wider domains.

D. Ontology

The term *"Ontology"* refers to the theoretical representation model in knowledge systems [15]. Many approaches tried to use Ontology in event detection in various forms. In [16], Ontology was built by concepts' relationships' learning based on analysing co-occurrences between concepts. Other approaches have directly included visual knowledge in Multimedia domain-specific Ontology, in the form of low-level visual descriptors for concept instances, to perform semantic annotation [1].

These methods almost define rules created by human experts, which make them not practical for the definition of a large set of rules and become less efficient in wide domain. In addition to that, these rules may be subject to some inconsistency, inherited from variation of the involved human's culture, mood/personality, as well as the specific topic.

E. Commonsense Knowledgebases

Commonsense is identified as the information and facts that are expected to be known by ordinary people. Although, it maybe considered as part of Ontology, we separate them to clarify the difference between domain-specific knowledge and commonsense knowledge. Most famous commonsense knowledgebases are WordNet [17], Cyc [18] and ConceptNet [19]. Currently, ConceptNet is considered to be the biggest commonsense database built from freely entered text. This knowledgebase is very rich in relationships, the number of assertions and the types of relationships.

Commonsense knowledgebases have recently received more attention for solving problems in semantic world, by finding related concepts. In [20], a trial has started to learn the concepts' relationships in public video databases depending on ConceptNet. The novelty in our framework is mainly located in finding the connection between low-level and the linguistic terms, based on these knowledgebases, by utilizing the relationships between the various parts of the linguistic sentence, not just as concepts.

F. Multimodal

Utilizing the available multimodality in video mediums, such as audio and sometimes enclosed text, has received relatively a good attention [3], [2]. In spite of that the multimodal features analysis usually increases certainty of video annotation, as well as it can be plugged in our framework, but we preferred to analyze input video's visual features only to keep focusing on wide domain. This is also to accommodate some domains where video clips lacks audio and enclosed text, or they are not so correlated with the visual features such as wild hunts [6] and surveillance.

III. Proposed framework

From the previous mentioned methodologies, it is noticed that semantic annotation in wide domain videos has two main issues: the first is visual features processing to gain knowledge about the contents, and the second is expressing this knowledge in annotation format which needs text processing. That was the inspiration for building a framework, which is depicted in figure 1, that helps in bridging this *"semantic gap"*, which is identified in section I.

The input to our proposed framework is a new video clip that needs to be annotated (for indexing and retrieval purposes for example.) The output is an annotation that indicates to occurring events in the input video, assigned to their objects and maybe locations, as illustrated in figure 2.

As depicted in figures 1 and 2, in the first stage the input video obtains initial weighted list of free-text annotations. This is done by comparing the dominant moving objects' low-level features form the input video against corresponding features

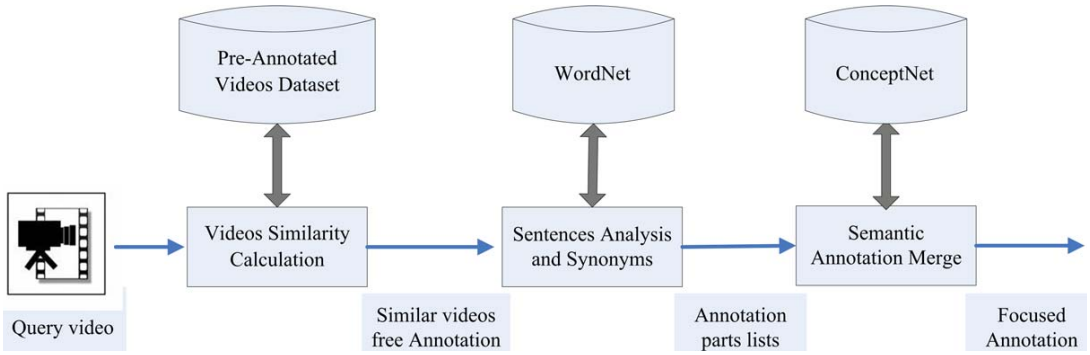

Fig. 1. Automatic Semantic Annotation Framework.

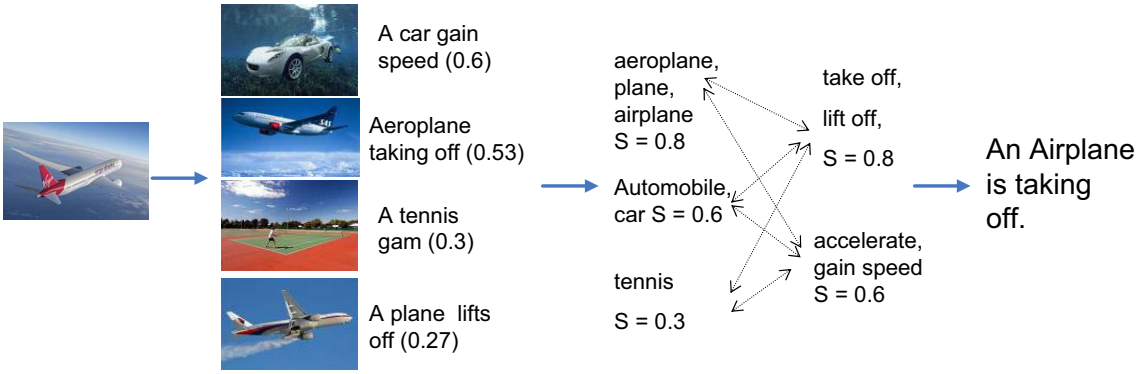

Fig. 2. An example for the framework.

from all videos in the dataset, then considering the annotations from the most matched videos.

Secondly, the relationships among annotations are analyzed to find general information about the environment of the video. As a result, some annotations gains more certainty and others are rejected which gradually reduces the available annotations list. The most candidate annotation is the highest ranked annotation in the resulted list. In the next sub sections, the proposed framework stages are described in detail:

A. *Videos Similarity Calculation*

In this layer, a generic non-domain-specific similarity calculation method is utilized to detect the most similar and related videos from a dataset to the input video. This layer is based on comparing local features obtained from spatio-temporal volumes extraction, inspired by Basharat *et al.* [11] approach. This layer is described in detail as follows:

1) Motion segmentation: First of all, interest points in all frames are calculated using SIFT features [21]. Then, temporal trajectories across all frames are built by connecting matched interest points between successive frames. A local neighbourhood rule is applied to guarantee smoothness of trajectories and to reduce false matching. At the end, broken trajectories are fixed using the same rules, then short trajectories are deleted to reduce the noise. In addition to that, interest points which do not belong to any trajectory, are also ignored.

2) Spatio-temporal volume calculation: After calculating the trajectories, a RANSAC algorithm is applied over each two successive frames to estimate different homographies. As a result, different objects in different depths are separated. An adapted connected component analysis (CCA) algorithm, to work over sparse features, is applied to separate different objects in the same depth.

After spatially separating objects within each frame, a labeling algorithm follows these objects over frames to form spatio-temporal volumes. Adjacent volumes are detected as separated, only if they manage to move apart at any time during the shot. This is done via merge and spilt algorithm described in [11].

3) Features Extraction: In previous stages volumes that represent moving objects have been calculated based on interest points matching. In this stage color, texture and motion features are also calculated for these volumes over the frames. All these features are unified in one vector and normalized to form a signature for this video.

4) Volumes Matching: As these previously mentioned stages had been pre-applied to all videos in the annotated dataset, the stored in this set is actually their features' signature. It is also applied to the new input video in the same way. The aim of this stage is to compare the extracted signature from the input video against the corresponding signature from each one of dataset videos in pairs using Earth Mover's

Distance (EMD) [22] to find the most similar videos.

The output of this layer is a weighted list of text entries that accompanied to the top similar matched videos.

B. Sentences Analysis and Synonyms

At the end of the previous layer, each text entry in the resulted list has a weight that is equal to the normalized similarity score between its source video and the input video.

The function of this layer is to find the similar meaning annotations regardless the variety of names used for the same or similar objects (e.g. car, automobile), the way of describing the event or action (e.g. speed up, accelerate, gain speed), or different spelling in various versions of the language (e.g. aeroplane in British, airplane in American English).

First of all, the sentence is divided into Object, Event and Location triplet. We use a Stanford NLP Log-linear Part-Of-Speech Tagger [23] to obtain the parts of the sentence. These tags indicate which part is the object, which is the subject in linguistic terminology, and which is the event, i.e. the verb and its related prepositions, and the location, if exists.

Three separated lists are generated from this analysis. Objects' and Locations' lists are considered as list of *nouns* when entered to WordNet [17]. This is to benefit from its very strong *Synset* relationship. The Events' list is considered as a list of verbs. This helps in preventing the similarity between some verbs and nouns like *fly*, as a verb for airplane for example, and as an insect.

The *"isA"* relationship in WordNet, which gives the synonyms, is selected because it gives equal meaning words with little amount of abstraction. Each list, separately, is extended then intersected using this relationship. The process is simply done by obtaining each item's synonyms, which match its part of speech (i.e. nouns for items in nouns' and locations' lists, and verbs for events' list). This assigns a suitable weight S_w for each synonym, calculated based on the initial word weight W_w and an un-trust decreasing constant C_d. This can be formulated as follows:

$$S_w = W_w \times C_d \tag{1}$$

The decreasing constant C_d can hold any value between 0 and 1, giving less weight for synonyms than the original word. Through experimentations, it has been found that if C_d value is small (i.e. below 0.3) the whole step become meaningless, and if it is high (i.e. more than 0.8) it starts to give the synonyms a similar strength as the original, so it increases the false alarms. Hence, the average $C_d = 0.5$ was chosen. This operation causes each list to be extended, but the matched words are grouped to increase their trust, and the resulted lists will also be normalized. At the end of this stage, the output is three sorted lists; each of which contains weighted entries for one part of the scene elements (object, event and location). Many factors give this process its strength:

- The repeated parts (or whole sentence) resulted from many similar videos, like two videos, one annotated as *"car speed up"* and the other *"car slow down"*. So the conclusion that it is a car regardless the event. Similarly, if two annotations *"boat sail on the sea"* and *"plane land on the sea"*, they refer to the fact that the environment (or the background) is the sea.
- Different words that refer to the same thing (synonym): like car, auto, and automobile. These will be grouped with higher confidence weight.
- Different spelling in different languages versions: like armored, armoured. They also will be grouped with higher confidence weight.
- Misspelling like *"mashine"* is ignored, and special annotations like *"jack running"* are assigned less weight as they do not have synonyms.

C. Semantic Annotation Merge

The aim of this layer is to check the possible conjunctions of the sentences' parts under real constraints, and to give more certainty to higher potential actions in our daily life. For this step, we use ConceptNet [19], but after some adaptation. ConceptNet consisted of a huge number of concept nodes; each concept is a semi-sentence or a phrase.

First we select the relationships' types that have usefulness in visual field. These relations are: *"capableOf"*, *"usedFor"*, *"locationAt"*, and *"isA"* . We merge both *"capableOf"* and *"usedFor"* relations into one relation, called *"event"*, by adding the score of each matched relations in both lists. Both *"event"* and *"locationAt"* types are used in annotation manipulation to connect the parts of the annotation. But it was noticed that *"isA"* includes many meanings, not only synonymic, such as abstraction and sometimes it gives a description or a property of a node. In addition to that, it's not symmetric, to be considered as synonym, neither fully asymmetric, to be considered as abstraction. To overcome these issues, WordNet *"isA"* relationship is utilized instead of it in the previous stage.

As resulted *event* relationship refers to the possibility and the score of assigning an event to an object, a full intersection operation is applied between objects' list and events' list using this relationship, and cross weighs are calculated as in equation 2. Then the same operation between objects' list and locations' list is repeated using *"locationAt"* relation.

$$T_w = N_w \times V_w \times R_s \tag{2}$$

Where: T_w is the sentence weight, N_w and V_w are the noun and verb phrases weights respectively, and R_s is the relation score.

In text mining applications, it is useful to use ConceptNet's nodes phrases directly. But in our case, the aim is to form a meaning that simulates the triplet of the visual world: objects, locations and events. In addition to that, it is important to achieve the most benefit from intersecting ConceptNet's nodes with WordNet's nodes. To achieve this, each ConceptNet node is analysed to obtain the core phrase that match its type. Finally, the rest of the node is deleted if it does not hold a full meaning or another node suits this meaning is created.

This is done like follows: First, each node's words are tagged using Stanford previously mentioned tagger [23], then

non-useful parts of sentence in visual field are deleted. These parts vary from some prepositions and stop words to some common used adjectives and adverbs, which are included in a manual written table. For example, ”fast” is visually a useful adjective because it holds a meaning related to motion, but ”better” is not. Then a split operation is applied to divide some complex nodes into parts causing new relationships to be established as depicted in figures 3 and 4.

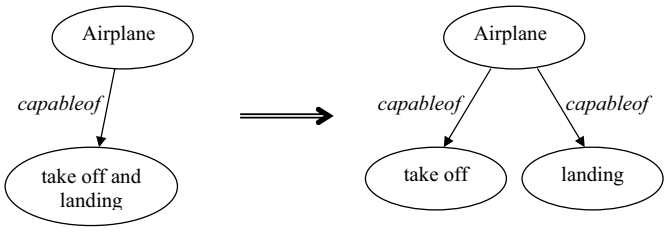

Fig. 3. Example of nodes analysis

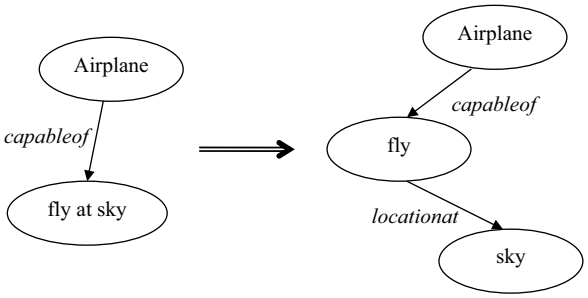

Fig. 4. Example of nodes analysis

To achieve more effective comparing between analyzed nodes and resulted candidate annotations' parts, this comparing operation is performed on the stemming level. This is done by stemming all the words of each entry, i.e. obtain the root of the word, then sorting the resulted stemmed words alphabetically. This causes the nodes that contain the same words but in different format to be comparable. For example, both ”a seasonally flower” and ”flowers of seasons” becomes ”flower season”.

IV. Experiments and results evaluation

The experiments have been performed on videos from TRECVID 2005 BBC Rushes [7], which is a group of standard databases for information retrieval. This dataset contains 335 single-shot video clips containing various types of moving vehicles like cars, tanks, airplanes and boats. These challenging uncontrolled videos contain considerable range of variations like size, appearance and shapes, viewpoint and motion of object. Also all possibilities of unknown camera quality and motion, like moving and zooming, are exists. The framework currently operates on individual video shots, but it can easily be extended by plugging a shot boundary detection layer.

Figure 5 shows the average of Recall vs. Precision, identified in equations 3 and 4, for applying this framework to annotate all database videos. Each time, one video is taken as a test and all other files in the database are considered as the pre-annotated database. Similarity has been evaluated by comparing correct similar retrieved files to all similar files, and enhancement has been evaluated by comparing correctly retrieved annotations to all possible correct annotations for the input video.

$$Recall = \frac{similar\_videos \ \cap \ retrieved\_videos}{similar\_videos} \quad (3)$$

$$Precision = \frac{similar\_videos \ \cap \ retrieved\_videos}{retrieved\_videos} \quad (4)$$

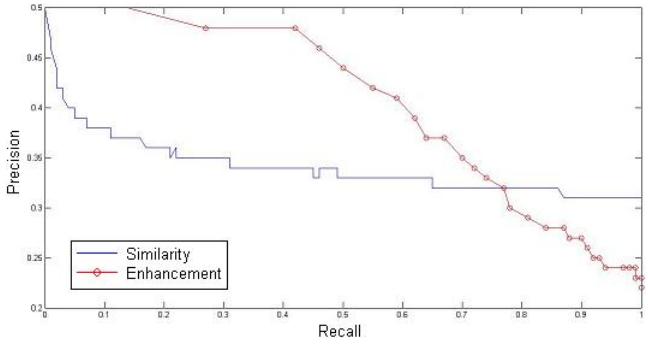

Fig. 5. Recall vs. Precision for similarity against annotation analysis enhancement

Second experiment is performed to clarify the effectiveness of existence of the input file itself in the pre-annotated dataset. It is aimed to clarify the effective of improvement of similarity layer results on annotation processing layer results. The balanced average of F-measure, identified in equation 5, against the number of top ranked files retrieved is drawn in figure 6. In this experiment, the pre-annotated dataset consists of all the 335 database videos; and each time one of these videos is considered an input file, then the average over all is drawn.

$$F = 2 \cdot (Precision \cdot Recall)/(Precision + Recall) \quad (5)$$

The results show that an acceptable percentage of good results obtained by the similarity calculation layer will lead to more accurate and precise results by annotations' analysis and enhancement. In addition to that, better similarity results will lead to find the correct annotation ranked in the top of the list.

The other notice is that bad results in the similarity layer, when precision is below 0.35 in figure 5, causes the enhancement to be almost non-beneficial. This is mainly because this layer increases the trust of similar meanings and repeated annotations but weakens the not agreed annotations. However, it is worth mentioning that the aim of the framework is not to retrieve the whole corrected list of annotations, but to find few representative annotations for the input video. This is achieved even with lower values of recall if the precision is accepted. In other words, we will take the first annotations in the list as a description for the event regardless the rest of the list.

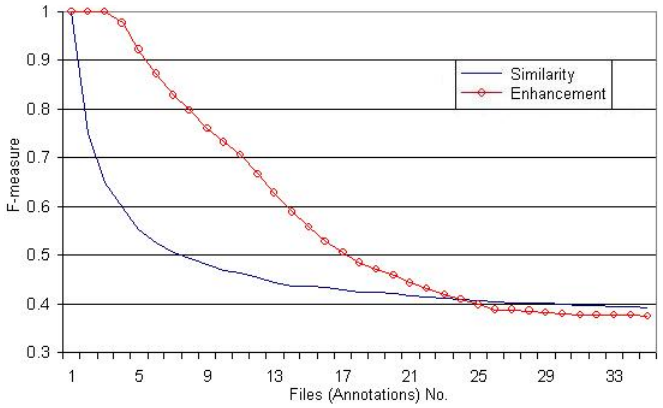

Fig. 6. F-measure for similarity against annotation analysis enhancement, compared by number of top ranked files and annotations retrieved.

V. Conclusion and Future work

In this paper, we introduced a novel framework for automatic semantic video annotation in wide domain uncontrolled videos to form the annotating part of video search engine. To achieve this, the proposed framework combines two non-domain-specific stages; low-level visual similarity and commonsense knowledgebases. As the resulted annotation is meant to satisfy human perceiving way for the visual scenes, the framework output composed of the triplet: objects, events and if detected locations.

The experiments demonstrated that this framework's stages can reach a small list of candidate annotations. However, the results, which evaluated upon retrieval performance, show that it is more likely to find the correct annotation ranked in the top. But a final model checking stage can be plugged in and applied for particular applications, to select the most related annotation.

Finally, the next step is to build an efficient volumes' comparison model, which suits large databases. This layer's mission is to retrieve the most similar videos to the input efficiently without comparing with all dataset's videos. Another step is to integrate the layers in a way that achieve interplay between visual similarity and text analysis, and achieve feeding resulted annotated video into the dataset. A final suggestion is to build a retrieval system which makes use of these knowledgebases to perform indirect semantic search queries inside the proposed indexed annotations.

Acknowledgment

The authors would like to thank Prof. Andrew Hunter for his efforts. This project is fully funded by Damascus University, Syria.

References

[1] A. D. Bagdanov, M. Bertini, A. D. Bimbo, G. Serra, and C. Torniai, "Semantic annotation and retrieval of video events using multimedia ontologies," in *International Conference on Semantic Computing*, 2007, pp. 713–720.

[2] A. Amir, S. Basu, G. Iyengar, C. Y. Lin, M. Naphade, J. R. Smith, S. Srinivasan, and B. Tseng, "A multi-modal system for the retrieval of semantic video events," *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 216–236, 2004.

[3] M. L. Shyu, Z. Xie, M. Chen, and S. C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 252–259, 2008.

[4] A. Haubold and M. Naphade, "Classification of video events using 4-dimensional time-compressed motion features," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM Press New York, NY, USA, 2007, pp. 178–185.

[5] A. Ekin and M. Tekalp, "Generic play-break event detection for summarization and hierarchical sports video analysis," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 1, 2003.

[6] N. Haering, R. J. Qian, and M. I. Sezan, "A semantic event-detection approach and its application to detecting hunts in wildlife video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 6, pp. 857–868, 2000.

[7] "Trec video retrieval track (2005)." [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[8] Y. Deng and B. Manjunath, "Content-based search of video using color, texture, and motion," in *International Conference on Image Processing*, vol. 2, 1997, pp. 534–537.

[9] S. F. Chang, W. Chen, H. Meng, and H. Sundaram, "A fully automated content-based video search engine supportingspatiotemporal queries," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 602–615, 1998.

[10] S. F. Chang, W. Chen, H. J. Meng, and H. Sundaram, "Videoq: an automated content based video search system using visual cues," in *Proceedings of the fifth ACM international conference on Multimedia*. ACM New York, NY, USA, 1997, pp. 313–324.

[11] A. Basharat, Y. Zhai, and M. Shah, "Content based video matching using spatiotemporal volumes," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 360–377, 2008.

[12] M. Bertini, A. D. Bimbo, and G. Serra, "Learning ontology rules for semantic video annotation," in *2nd ACM workshop on Multimedia semantics on International Multimedia Conference*. ACM New York, NY, USA, 2008, pp. 1–8.

[13] K. H. Liu, M. F. Weng, C. Y. Tseng, Y. Y. Chuang, and M. S. Chen, "Association and temporal rule mining for post-filtering of semantic concept detection in video," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 240–251, 2008.

[14] A. Dorado, J. Calic, and E. Izquierdo, "A rule-based video annotation system," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 622–633, 2004.

[15] B. Chandrasekaran, J. Josephson, and V. Benjamins, "What are ontologies, and why do we need them?" *IEEE Intelligent Systems and their applications*, vol. 14, no. 1, pp. 20–26, 1999.

[16] A. G. Hauptmann, M. Y. Chen, M. Christel, W. H. Lin, and J. Yang, "A hybrid approach to improving semantic extraction of news video," in *International Conference on Semantic Computing, 2007. ICSC 2007.*, 2007, pp. 79–86.

[17] C. Fellbaum, *WordNet: an electronic lexical database*. Cambridge, Mass: MIT Press, 1998.

[18] D. B. Lenat, "Cyc: A large-scale investment in knowledge infrastructure," *Communications of the ACM*, vol. 38, no. 11, pp. 33–38, 1995.

[19] H. Liu and P. Singh, "Conceptnet a practical commonsense reasoning tool-kit," *BT Technology Journal*, vol. 22, no. 4, pp. 211–226, 2004.

[20] P. Yuan, B. Zhang, and J. Li, "Semantic concept learning through massive internet video mining," in *IEEE International Conference on Data Mining Workshops*, 2008, pp. 847–853.

[21] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.

[22] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[23] S. N. Group, "The stanford nlp log-linear part of speech tagger." [Online]. Available: http://nlp.stanford.edu/software/tagger.shtml