Utilizing Semantic Word Similarity Measures for Video Retrieval

Yusuf Aytar
Computer Vision Lab,
University of Central Florida
yaytar@cs.ucf.edu

Mubarak Shah
Computer Vision Lab,
University of Central Florida
shah@cs.ucf.edu

Jiebo Luo
R&D Laboratories,
Eastman Kodak Company
jiebo.luo@kodak.com

Abstract

This is a high level computer vision paper, which employs concepts from Natural Language Understanding in solving the video retrieval problem. Our main contribution is the utilization of the semantic word similarity measures (Lin and PMI-IR similarities) for video retrieval. In our approach, we use trained concept detectors, and the visual co-occurrence relations between such concepts. We propose two methods for content-based retrieval of videos: (1) A method for retrieving a **new** concept*(a concept which is not known to the system, and no annotation is available) using semantic word similarity and visual co-occurrence. (2) A method for retrieval of videos based on their relevance to a user defined text query using the semantic word similarity and visual content of videos. For evaluation purposes, we have mainly used the automatic search and the high level feature extraction test set of TRECVID'06 benchmark, and the automatic search test set of TRECVID'07. These two data sets consist of 250 hours of multilingual news video captured from American, Arabic, German and Chinese TV channels. Although our method for retrieving a new concept is an unsupervised method, it outperforms the trained concept detectors (which are supervised) on 7 out of 20 test concepts, and overall it performs very close to the trained detectors. On the other hand, our visual content based semantic retrieval method performs 81% better than the text-based retrieval method. This shows that using visual content alone we can obtain significantly good retrieval results.*

1. Introduction

Video retrieval–searching and retrieving the videos relevant to a user defined query–is one of the most popular topics in multimedia research [10, 20, 18, 12]. Most of the current effective retrieval methods rely on the noisy text information contained in the videos. With the release of the LSCOM (Large Scale Concept Ontology for Multimedia) [9] lexicon and annotation, a large number of visual content based semantic concept detectors, which includes detectors for objects (e.g. car, people), scenes (e.g. office, outdoor) and events (e.g. walking and marching), have been developed [28, 8]. These concept detectors are essentially SVM classifiers trained on visual features e.g. color histograms, edge orientation histogram, SIFT descriptors etc. Recently, using these concept detectors, some promising video retrieval methods have been reported [22, 27, 16, 6]. In this paper, we propose a novel use of these concept detectors to further improve video retrieval.

The main contribution of this paper is utilization of the *semantic word similarity* measures for the content-based retrieval of videos. We focus on two problems: 1) concept retrieval, and 2) semantic retrieval. The aim of concept retrieval is: given a concept (e.g. "airplane" or "weather"), retrieve the most relevant videos and rank them based on their relevance to the concept. Similarly, semantic retrieval can be summarized as: given a search query (e.g. "one or more emergency vehicles in motion", "US President George Bush walking") specified in the natural language (English), return the most relevant videos and rank them based on their relevance to the query.

Although there are several approaches that exploit the context of low and mid-level features [25, 3], there are not many approaches that explore context of high-level concepts [24, 5]. This paper proposes a novel way for exploiting the context between high-level concepts. The underlying intuition behind our approach is based on the fact that certain concepts tend to occur together, therefore we can harness from this visual co-occurrence relations between concepts in order to improve retrieval. In [4] it is reported that excluding target concept's own detector, 18 out of 39 concepts can be better retrieved using other concept detectors and the visual co-occurrence relations. However, in order to obtain visual co-occurrence relations, the annotated video shots are required. The vital question here is "Can we retrieve a concept for which we don't have any annotation or training examples?" In order to accomplish this goal, we need to find some other relations to substitute the visual co-occurrences. The semantic word similarity arises as a good option for this substitution. Does semantic word similar-

ity have a strong correlation with visual co-occurrence? In other words, do we see a vehicle when we see a car? Do we see a person when we see a crowd? Do we see goalposts when we see a soccer game? These are different degrees of semantic relatedness, and intuitively it is apparent that the semantic word similarity has some correlation with the visual co-occurrence.

In this paper, we show that the semantic word similarity is a good approximation for visual co-occurrence. With the help of semantic word similarity a new concept–the concept for which we don't have any annotated video shots–can be detected sometimes better than if we had its individually trained detector (SVM classifier). The key point of our work is removing the need for annotation in order to retrieve a concept in a video. Furthermore, using the same intuition we propose a method for semantic retrieval of videos. This is based on relevance of videos to user defined queries, which is computed using the earth movers distance (EMD).

We have tested our method for retrieving new concept on Trecvid'06 [21] test set. Our retrieval method works better than the trained detectors on 7 out of 20 test concepts. Overall, our method gives results close to the trained detectors, and we show that even without any annotation we can retrieve concepts with reasonable accuracy. Also, we evaluated semantic retrieval method on Trecvid'06 and Trecvid'07 [21] sets. We have obtained 81% and more than 100% performance increase over our text based retrieval method on Trecvid'06 and Trecvid'07 sets respectively.

The paper is organized as follows: In the next section the similarity measures will be presented. In section 3, a method for retrieving new concept using similarity measures will be discussed. In section 4, we will describe our semantic video retrieval method. In section 5, we'll present experimental results on the TRECVID'06 and TRECVID'07 [21] video collections. And finally we will conclude with discussions and future work.

2. Similarity Measures

In this section, visual co-occurrence and semantic word similarity measures will be discussed. Visual co-occurrence is a relation between two concepts; it simply signifies the possibility of seeing both concepts in the same scene. In order to compute visual co-occurrence, we need concept annotations of video shots. On the other hand, the semantic word similarity is the relatedness of two words, and it is generally a common sense knowledge that we build for years. Measuring this quantity has been a challenging task for researchers, considering the subjectivity in the definition of semantic word similarity.

2.1. Visual Co-occurrence

In order to obtain visual co-occurrence we use an annotated set of video shots. Video shots are taken from Trecvid'06 development data and we use LSCOM annotation. Then the visual co-occurrence is approximated as pointwise mutual information (PMI) between two concepts as below:

$$Sim_{Visual}(c_i, c_j) = Sigmoid\left(PMI_{Visual}(c_i, c_j)\right),$$

where

$$PMI_{Visual}(c_i, c_j) = log\left(\frac{p(c_i \& c_j)}{p(c_i)p(c_j)}\right),$$

$$Sigmoid(x) = \frac{1}{1+e^{-x}},$$

c_i, c_j are the concepts and $p(c_i \& c_j)$ is the probability of both concepts occurring together, and $p(c_i), p(c_j)$ are the individual probabilities of concepts. These probabilities are computed using the annotation of training video data set. Then $Sigmoid$ function is applied for scaling the similarity measure between the interval [0-1].

2.2. Semantic Word Similarity

Semantic word similarity has been widely studied, and there are many semantic word similarity measures introduced in the literature. Due to the subjectivity in the definition of the semantic word similarity, there is no unique way to compute the performance of the proposed measures. These measures are folded into two groups in [14]: corpus-based similarity and knowledge-based similarity measures. The corpus-based measures try to identify the similarity between two concepts using the information exclusively derived from large corpora. The knowledge-based measures try to quantify the similarity using the information drawn from the semantic networks. In this paper, we examine two different semantic word similarity measures, Lin's knowledge based similarity measure [11], and PMI-IR [26] corpus based similarity measure.

2.2.1 Lin's Similarity Measure

Similar to many other knowledge based similarity measures, this measure uses WordNet [15] (which is a semantic lexicon for the English language) as the knowledge base. It mainly uses the information content (IC) of the concepts, and the least common subsumer (LCS) of the concepts in the WordNet taxonomy. LCS is the common ancestor of two concepts which has the maximum information content. In the Figure 1, LCS is described visually with an example. The key idea in this measure is to find the maximum information shared by both concepts and normalize it. Similarity is measured as the information content of LCS, which can be seen as a lower bound of the shared information between

concepts, and then normalized with the sum of information content of both concepts. The formulation is as below:

$$Sim_{Lin}(c_i, c_j) = \frac{2 \times IC(LCS(c_i, c_j))}{IC(c_i) + IC(c_j)},$$
$$IC(c) = -log(P(c)),$$

where the $Sim_{Lin}(c_i, c_j)$ is the similarity between concepts c_i, c_j, $LCS(c_i, c_j)$ is the least common subsumer of c_i, c_j, and $IC(c)$ is the information content of the concept c.

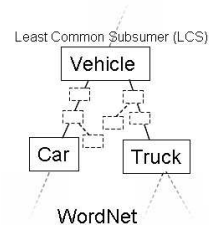

Figure 1. In this example LCS of the concepts *car* and *truck* is the *vehicle* in the given taxonomy.

2.2.2 PMI-IR Similairty

The pointwise mutual information using data collected by information retrieval (PMI-IR) was proposed as a semantic word similarity measure in [26]. The main idea behind this measure is that similar concepts tend to occur together in the documents more than the dissimilar ones. Actually this measure is very similar to the visual co-occurrence measure. The main difference is that instead of considering the visual co-occurrence here we search for the text co-occurrence. The pointwise mutual information between concepts is approximated using a web search engine, particularly in our case we use Yahoo [2] web search engine. The formulation is given as below:

$$Sim_{PMI-IR}(c_i, c_j) = Sigmoid\left(PMI_{IR}(c_i, c_j)\right),$$
$$PMI_{IR}(c_i, c_j) = log\left(\frac{p(c_i \& c_j)}{p(c_i)p(c_j)}\right),$$
$$= log\left(\frac{hits(c_i\ NEAR\ c_j) * WebSize}{hits(c_i)hits(c_j)}\right),$$

where $hits(c_i\ NEAR\ c_j)$ is the number of documents in which c_i, c_j occur in a window of ten words, $WebSize$ is the approximated by a number of documents on the web based on Yahoo Search; $hits(c_i), hits(c_j)$ are the number of retrieved documents for individual concepts. Then, the $Sigmoid$ function is applied for scaling the similarity measure between the interval [0-1].

3. Retrieving New Concept

Traditional way of retrieving a concept can be summarized in two steps. The first step is training of visual detectors for each concept. For a selected concept, using the annotated video shots, positive and negative sets of shots are extracted, and visual features like edge orientation histogram are computed from the key frame of each shot. Next, a detector (a classifier) is trained using these features. This process is repeated for all concepts. This step assumes that video shots for training have been manually annotated. In the second step, the retrieval of the desired concept is achieved by running all video shots through the desired concept detector and the detection confidences are obtained. After that the video shots are sorted using the confidences and the sorted list is returned to the user. Although this supervised training for concept detection is acceptable, manual annotation of concepts in videos is a time consuming task. Thus, supervised training is not a realistic approach for retrieving all the concepts in the real world. In this paper, we show that the retrieval of a concept can also be performed in an unsupervised manner (without having any annotated video shots of that concept) with a reasonable accuracy.

In this section, we will discuss unsupervised retrieval of a new (unknown) concept using other available concept detectors and their similarity relations with the new concept. From here on visual co-occurrence and semantic word similarity measures will be referred as similarity measures.

Assume an annotated (known) concept set $C = \{c_j\}_{j=1}^{M}$, where M is the total number of annotated concepts, and c_j is the j^{th} concept in the set; and $SD = \{s_k\}_{k=1}^{L}$ is video shot database, where L is the number of video shots, and s_k is the k^{th} video shot in the database. Then, the task of retrieving a new concept is accomplished by computing a relevance score for each shot, and then ranking the shots based on their scores. The confidence that a given shot contains a new concept is computed as a linear combination of similarity measures between the known concepts and the new concept, and the scores obtained from the known concept detectors. Then this score is normalized by the sum of the scores obtained by the known concept detectors. The formulation is as follows:

$$Score_{c_n}(s_k) = \frac{\sum_{j=1}^{M} Sim(c_j, c_n) Score_{c_j}(s_k)}{\sum_{j=1}^{M} Score_{c_j}(s_k)},$$

where $Score_{c_n}(s_k)$ and $Score_{c_j}(s_k)$ respectively are the confidences that the new concept c_n occurs in shot s_k and concept c_j occurs in shot s_k. $Sim(c_j, c_n)$ is the similarity between the new concept c_n and the annotated concept c_j.

4. The Semantic Video Retrieval

The semantic video retrieval–search and retrieval of the videos based on their relevance to a user defined text query–has attracted a noteworthy attention in the recent years. The traditional way of semantic retrieval is through the use

of the text information in the videos, which can be obtained from the closed captions, automatic speech recognition (ASR), or tagging. Several information retrieval approaches have been already proposed in the literature. On the other hand, the use of visual content in semantic retrieval is relatively new. However, see some recent approaches [22, 27, 16, 6].

In this section, we propose a new method for semantic retrieval, using the visual content of the videos through trained concept detectors. The approach stems from the intuitive idea, that is, new concepts can be detected using the context of available concept detectors and the semantic similarities between the new and known concepts. However, in this case instead of having only one new concept we may have a group of new concepts in a query. Hence, the problem becomes finding the relevance between a group of query words and a group of known concepts. The computation of this relevance is done in two steps. Initially, both the query and the video shots are expressed using appropriate representations. And then the relevance between the shot and query representations are computed using the earth movers distance (EMD) [19]. The overview of the method is visually described in Figure 2. In order to perform the comparison, we also apply a text based retrieval method which we will discuss at the end of this section.

4.1. Representation of the Query and Video Shots

Queries and video shots provide two different kinds of information, and there is no obvious way for computing the relevance between a query and a video shot. In order to compute the relevance we need similar representations. In this section, we will specify appropriate representations for both queries and video shots.

Since queries are most often expressed as sentences, there are many common words, such as 'off', 'as', 'to', which don't necessarily contribute to the meaning of the query, and create noise in the retrieval process. Therefore, initially we remove the common words from the query us-

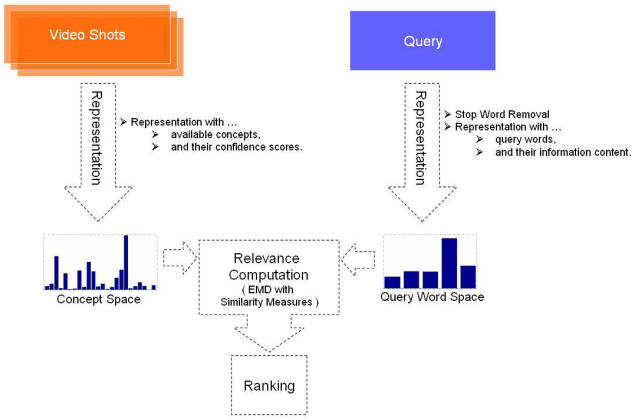

Figure 2. An overview of the Visual Content-based Semantic Video Retrieval method.

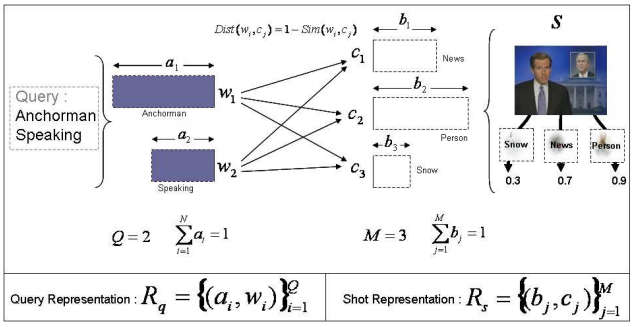

Figure 3. EMD based distance computation between the visual content and the query.

ing a common word list. Among the remaining ones, not all the words have the same significance within the query. Some words may contribute more, and some words may contribute less to the meaning of the query. For instance, in the query 'George Bush walking', it is apparent that the words 'George' and 'Bush' contribute to the query more than the word 'walking'. The contribution weight can be approximated by the specificness of the word. The information content, which specifies the amount of information that a word has, is a way to measure this specificness. Hence, we weigh the words in the query based on their information content, so that we will have a stronger representation of the underlying semantic meaning of the query.

The visual content of the video is utilized through the trained concept detectors. For a given shot, each concept detector provides a score which is the confidence that concept is present in the shot. Analogous to the query representation these scores can be seen as the weights for the corresponding concepts, and the underlying semantic meaning of the shot can be represented with concepts and their weights. Each concept is expressed with a representative word.

The query q is represented as $R_q = \{(a_i, w_i)\}_{i=1}^{Q}$, where w_i is the word, a_i is its weight, and Q is the number of the words in the query. Similarly, the video shot s is represented as $R_s = \{(b_j, c_j)\}_{j=1}^{M}$, where c_j represents the known concept, b_j is its weights. In both representations the sum of the weights is normalized to one.

4.2. Computing the Shot-Query Relevance Using Visual Content

After finding expressive representations, the next task is to compute the relevance between the shots and the query. We consider both query and the shot representations as two histograms, where concepts and query words correspond to the bins, and the weights correspond to the values of the bins (Figure 3). The computation of the distance between two histograms would be an easy task if the bins in both histogram represent the same labels. But in our case, we have two different groups of bins. Nevertheless, since we can compute the similarity between a concept and

a query word, we know distances between bin pairs. Therefore EMD (Earth Movers Distance) measure perfectly fits to this problem. In this context, the distance becomes the minimum amount of work needed to transform a query histogram into the shot histogram.

Given the query representation $R_q = \{(a_i, w_i)\}_{i=1}^{Q}$, and the shot representation $R_s = \{(b_j, c_j)\}_{j=1}^{M}$, the distance is computed solving the optimization problem given below:

$$EMD(R_q, R_s) = \underset{F=\{f_{ij}\}}{argmin} \sum_{i,j} f_{i,j} Dist(w_i, c_j),$$

with the following constraints:

$$Constraints : \sum_i f_{i,j} = b_j, \sum_j f_{i,j} = a_i,$$
$$f_{i,j} \geq 0, 1 \leq i \leq Q, 1 \leq j \leq M,$$

where $EMD(R_q, R_s)$ is the distance between query q and shot s, $f_{i,j}$ is the flow between two bins i and j, and the F is the overall flow configuration which is optimized for the minimum amount of the work. The distances between two bins is described as:

$$Dist(w_i, c_j) = 1 - Sim(w_i, c_j).$$

Finally, the score of the shot for the given query is computed as :

$$Score_q(s) = 1 - EMD(R_q, R_s).$$

This optimization problem is solved using the linear programming technique.

4.3. Retrieval Using Text Information

There are several existing text similarity measures, which have been used for the information retrieval tasks. In our text baseline, we use one of the most effective text similarity measures according to [13]. Queries are extended using synonyms of query words obtained from the WordNet. The relevance of the shot for the given query is computed as the intersection of extended query and shot words, divided by their union. Additionally, each word is weighted with its length. This weighting depends on the hypothesis that, in general, longer words are more likely to represent the subject of a text string than the shorter words.

The extended query is represented as the set $q = \{w_i\}_{i=1}^{Q}$, where w_i are the words, and Q is the number of the words. Text of the shot is represented as the set $t = \{w_t\}_{t=1}^{T}$, where w_t are the words and T is the number of the words. Then the relevance of a shot for an extended query is computed as below:

$$Score_q(t_k) = \frac{\sum_{w \in q \cap t_k} length(w)}{\sum_{w \in q \cup t_k} length(w)},$$

where $Score_q(t_k)$ is the text based relevance of shot s_k, t_k is the text information of shot s_k, and $length(w)$ is the length of the word w.

5. Experiments

For evaluation purposes, we mainly use the high level feature extraction and automatic search test data set of TRECVID'06 benchmark [21]. TRECVID'06 test data set consists of 150 hours of multilingual news videos captured from American, Arabic and Chinese TV channels and is split into 79,484 shots. On this test set, we evaluated both of our methods, a method for retrieving a new concept and a method for semantic retrieval. In addition, we also tested our semantic retrieval method on TRECVID'07 test data set which consists of 100 hours of news video entirely in German split into 18,142 shots. We used 374 concept detectors which are publicly available at [28]. These detectors are trained using another 80 hours of news video. For Lin's similarity measure we used the Wordnet::Similarity package released by [17]. And we computed PMI-IR similarity and information content using Yahoo search engine [2]. For the EMD optimization problem we used source code provided by [19], with some manipulations for our needs. As a comparison metric, we used average precision (AP) which emphasizes returning more relevant documents earlier. It is the average of precisions computed for the first k shots in the ranked list where each k is the rank (e.g. 2^{nd},5^{th},...) of a relevant document. Considering that AP is computed using the complete video shot database, even small value of AP, e.g. between 3% - 30%, leads to very nice retrieval results. We also used mean average precision (MAP) (which is the mean of average precision results for all the cases) for comparison of methods in overall.

5.1. Evaluation of Retrieving New Concept

In this evaluation, we used TRECVID'06 (high level feature extraction) test concepts. Twenty test concepts include events such as 'people-marching', scenes such as 'office' or object classes such as 'bus'. In fact, we also have access to the trained concept detectors for these concepts from 374 concept detectors set. Since test concepts should be the new concepts, during the evaluation of the new concept retrieval method we discarded the associated 20 concept detectors and used the remaining detectors and the similarity measures. Also, we used these 20 detectors as a ground truth for the comparison purpose. We performed three different evaluations of our retrieval method using visual co-occurrence, Lin's similarity and PMI-IR similarity.

Using visual co-occurrence, our method performed well, and even it outperformed the trained concept detectors on 6 out of 20 concepts. Overall, even though the method couldn't perform as well as the trained concept detectors, it performed very close. The success of visual co-occurrence was expected, but the real surprise was the result of PMI-IR similarity measure which performed better than the trained detectors on 7 out of 20 concepts, and overall it performed

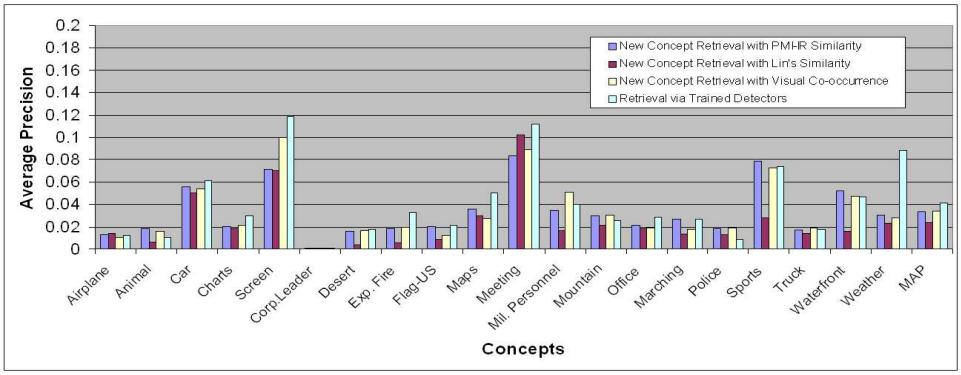

Figure 4. A comparison of different retrieval methods using Average Precision. MAP (Mean Average Precision) is shown at far right.

very close to visual co-occurence. Through these experiments, we observed that an unsupervised retrieval of new concepts can be accomplished with a reasonable accuracy. We also repeated the same experiment for Lin's similarity measure. Unfortunately, it did not perform as well as PMI-IR similarity or the visual co-occurrence. As it is shown in the Figure 7, our retrieval method using visual co-occurrence and PMI-IR similarity has almost the same precision for several recall values.

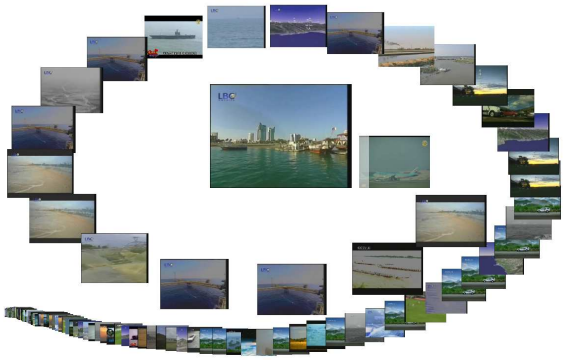

Figure 5. Top 100 retrieval results for the new concept 'fishing'.

Figure 6. Top 100 retrieval results for the new concept 'shouting'.

For 'animal' and 'police' concepts our method with PMI-IR similarity performed 80% better than the trained detectors. 'sports, mountain, waterfront' are some other concepts for which our method performed significantly better than the trained detectors. These concepts mostly have strong contextual relations. Conversely, our method couldn't retrieve that well the concepts that mostly appear in isolation, and have loose contextual relations such as 'screen, charts, maps, weather'. As a result, we observed that the concepts with strong contextual relations can be retrieved better than by using individually trained detectors. Overall, using just the context of available concept detectors and the similarities, new concepts can be retrieved with reasonably good accuracy.

We also applied our method to completely new concepts, which are different from 374 concepts, and for these concepts we don't have trained detectors. Since we don't have ground truth for these concepts, we only demonstrate top 100 shots retrieved by our method using PMI-IR similarity. Figure 6 and 5 show the retrieved shots for the "shouting" and "fishing" concepts, respectively. Indeed, we didn't know if these exact concepts are present in our shot database or not. However, this method can retrieve semantically similar results. For instance, the retrieved shots for the "fishing" concept mostly include ships, river, sea, and people which are all semantically relevant to the "fishing" concept. Since this method uses context of other concepts, it can also retrieve concepts which can not be easily recognized using visual features. For example, even though "shouting" is not a visually recognizable concept, the retrieved video shots mostly contain demonstrations, protests, parades, entertainment scenes, basketball games and fans which frequently occur together with "shouting" concept.

The accuracy of our retrieval method highly depends on the accuracy of the trained concept detectors. Even though we don't know the exact accuracy of the whole detector set, we know that the MAP of the selected 20 detectors is 4.1%.

Method Name		MAP
New Concept Retrieval	PMI-IR	3.3%
	Lin's Similarity	2.4%
	Visual Co-oc.	3.4%
Trained Concept Detector		4.1%

Table 1. MAP comparison of different concept retrieval methods.

Therefore, we believe that the MAP of the whole set would be close to 4%. And the accuracy of our method with PMI-IR similarity is 3.3%. As a conclusion, using a lexicon of 374 concepts new concept retrieval is 80% as accurate as individually trained detectors. The MAP comparisons for these methods are shown in Table 1.

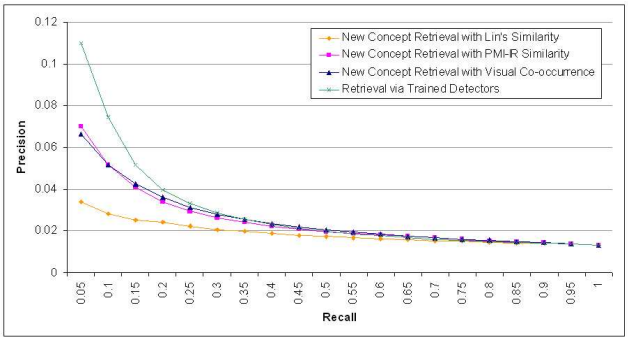

Figure 7. Average precision-recall curves for different concept retrieval methods.

5.2. The Semantic Retrieval Evaluations

These evaluations are performed on TRECVID'06 and TRECVID'07 test sets, and associated queries of automatic search challenge. These 48 queries include events such as 'walking, greeting', objects such as 'computer, book, door' and also some named entities such as 'Saddam Hussein, Dick Cheney, George Bush'. The exact expressions of the queries can be found on the official website [1] of TRECVID benchmark. In these evaluations, we compared our visual content based semantic retrieval (VC-SR) method with the text based semantic retrieval (TEXT-SR) method. For the VC-SR we tested both Lin's similarity and PMI-IR similarity. We also experimented with average fusion for combining the VC-SR and the TEXT-SR methods.

First, we tested the methods on TRECVID'06 test set. As it is shown in Figure 8, the VC-SR with PMI-IR similarity had significantly better performance than the TEXT-SR on 17 out of 24 queries. Overall, with 3.4% MAP it had 81% performance increase over the TEXT-SR. On the other hand, even though the VC-SR using Lin's similarity gives very nice results on some queries; it could neither outperform the TEXT-SR method nor the VC-SR with PMI-IR similarity. As expected, for queries which include the named entities, the VC-SR methods did not perform as well as the TEXT-SR. With a 3.7% MAP, fused results had 98% relative performance increase over the TEXT-SR, and 9% relative increase over the VC-SR with PMI-IR similarity.

The performance on TRECVID'07 test data set is similar to TRECVID'06 results. Due to the space limitations we are not able to include the results for each query, but the overall performance is shown in Table 2. As expected, the VC-SR with PMI-IR similarity gives the best performance with 3.6% MAP, and MAP of the TEXT-SR was 1.6%. In this evaluation, with 2.1% MAP, the VC-SR with Lin's similarity gives a better performance than the TEXT-SR. Since we have less named entities in TRECVID'07 queries, text based retrieval is not as effective as it is in TRECVID'06 evaluation. The VC-SR with PMI-IR gives almost the same results on TRECVID'06 and TRECVID'07. The performance of the VC-SR with Lin's similarity in TRECVID'07 is better than TRECVID'06 results, and it outperformed the TEXT-SR in TRECVID'07. Unlike TRECVID06, fusion of TEXT-SR and VC-SR with PMI-IR decreased the performance of VC-SR in TRECVID07.

Considering the fact that the MAP of trained detector set in TRECVID'06 is around 4%, having 3.4% MAP for semantic retrieval is a significant achievement. It is obvious that if we had stronger detectors the performance of the VC-SR would be much better. Moreover, in many of the previous studies [7, 22, 23] it is mentioned that increasing the number of concept detectors will increase the performance of semantic retrieval. We believe that qualitative and quantitative increase of concept detectors will leverage the quality of this approach in the future.

Semantic Retrieval Method		MAP'06	MAP'07
Vision Based Retrieval	PMI-IR	3.4%	3.6%
	Lin's Similarity	1.1%	2.1%
Text Based Retrieval		1.9%	1.6%
Average Fusion (Text + PMI-IR)		3.7%	3.5%

Table 2. MAP comparison for different semantic retrieval methods on TRECVID'06 and TRECVID'07 test data sets.

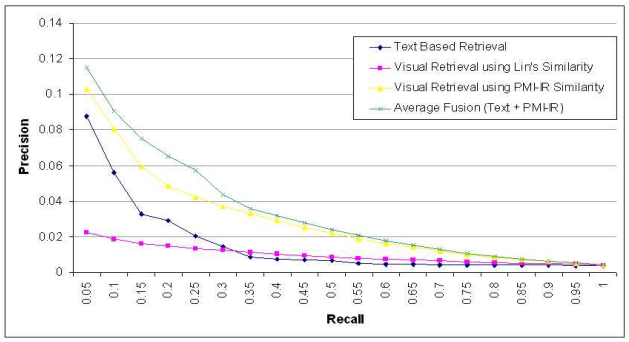

Figure 9. Average precision-recall curves for different semantic retrieval methods on TRECVID'06 queries.

6. Conclusions

This is perhaps one of the first attempts in using high level knowledge for solving video retrieval problem in computer vision. We proposed an effective way of using high level semantic relations in video retrieval problem by establishing a bridge between the low level visual concept detectors and semantic relations between such concepts.

Humans frequently use the context of known concepts in order to learn new concepts. The work in this paper is

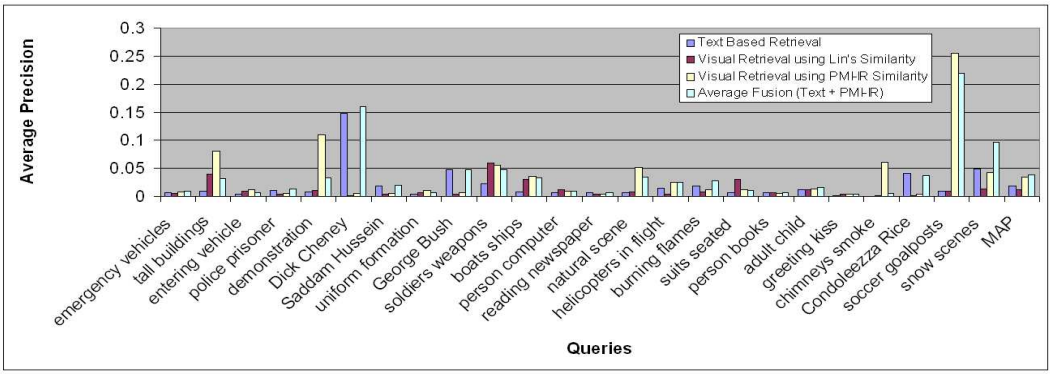

Figure 8. AP (Average Precision) results for different semantic retrieval methods for each query (represented by some selected words for each query) in TRECVID'06. MAP (Mean Average Precision) is shown at far right.

motivated by this intuitive observation. We proposed two different methods for semantic video retrieval using high level contextual relations between concepts. These relations are automatically extracted from text (language) resources available on the web or hand crafted semantic networks. For both of the proposed methods we have obtained promising results to pursue further research on this newly emerging field.

References

[1] The offical website of trecvid benchmark, http://www-nlpir.nist.gov/projects/trecvid/.

[2] Yahoo web search engine, http://www.yahoo.com.

[3] J. Amores, N. Sebe, and et al. Context-based object-class recognition and retrieval by generalized correlograms. *IEEE TPAMI*, 2007.

[4] Y. Aytar and et al. Improving semantic concept detection and retrieval using contextual estimates. *In Proc. of ICME'07*.

[5] K. Barnard and et al. Matching words and pictures. *Journal of Machine Learning Research*, 2003.

[6] A. Haubold and et al. Semantic multimedia retrieval using lexical query expansion and model-based reranking. *In Proc. of ICME'06*, 2006.

[7] A. Hauptmann and et al. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Transactions on Multimedia*, 2007.

[8] Y. Jiang and et al. Towards optimal bag-of-features for object categorization and semantic video retrieval. *ACM CIVR'07*, 2007.

[9] L. Kennedy and et al. Lscom lexicon definitions and annotations version 1.0. *DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia*, 2006. Columbia University ADVENT Technical Report #217-2006-3.

[10] M. Lew, N. Sebe, T. Huang, and et al. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing*, 2006.

[11] D. Lin. An information-theoretic definition of similarity. *In Proc. of the ICML'98*, 1998.

[12] K. J. Liu, Y. Fast video segment retrieval by sort-merge feature selection, boundary refinement, and lazy evaluation. *Computer Vision and Image Understanding*, 2003.

[13] P. Lynch and et al. An evaluation of new and old similarity ranking algorithms. *In Proc. of ITCC'04*, 2004.

[14] R. Mihalcea and et al. Corpus-based and knowledge-based measures of text semantic similarity. *In Proc. of AAAI'06*.

[15] G. Miller and et al. Wordnet: An electronic lexical database. *MIT Press*, 1998.

[16] S. Y. Neo and et al. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. *ACM CIVR'06*, 2006.

[17] T. Pedersen. Wordnet::similarity - measuring the relatedness of concepts. *In Proc. of AAAI'04*, 2004.

[18] D. Ponceleon and et al. Cuevideo: automated multimedia indexing and retrieval. *In Proc. of ACM Multimedia*, 1999.

[19] Y. Rubner and et al. A metric for distributions with applications to image databases. *In Proc. of ICCV'98*.

[20] J. Sivic, , and et al. Video google: A text retrieval approach to object matching in videos. *In Proc. of ICCV'03*.

[21] A. F. Smeaton and et al. Evaluation campaigns and trecvid. *MIR '06: 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, 2006.

[22] C. G. Snoek and et al. Adding semantics to detectors for video retrieval. *IEEE Trans. on Multimedia*, 2007.

[23] C. G. Snoek and et al. Are concept detector lexicons effective for video search? *In Proc. of ICME'07*, 2007.

[24] A. Torralba and et al. Context-based vision system for place and object recognition. *In Proc. of ICCV'03*, 2003.

[25] A. Torralba and et al. Contextual models for object detection using boosted random fields. *In NIPS'05*, 2005.

[26] P. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. *In Proc. of ECML'01*, 2001.

[27] D. Wang and et al. The importance of query-concept-mapping for automatic video retrieval. *In Proc. of ACM Multimedia*, 2007.

[28] A. Yanagawa and et al. Columbia university's baseline detectors for 374 lscom semantic visual concepts. *Columbia University ADVENT Technical Report #222-2006-8*, 2006.