SEMANTIC VIDEO RETRIEVAL USING HIGH LEVEL CONTEXT

by

YUSUF AYTAR
B.S. Ege University

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the School of Electrical Engineering and Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2008

Major Professor: Mubarak Shah

ABSTRACT

Video retrieval – searching and retrieving videos relevant to a user defined query – is one of the most popular topics in both real life applications and multimedia research. This thesis employs concepts from Natural Language Understanding in solving the video retrieval problem. Our main contribution is the utilization of the semantic word similarity measures for video retrieval through the 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, which is an unsupervised method. (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 mainly used the automatic search and the high level feature extraction test set of TRECVID'06 and TRECVID'07 benchmarks. 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 more than 100%

better than the text-based retrieval method. This shows that using visual content alone we can have significantly good retrieval results.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Video retrieval – searching and retrieving videos relevant to a user defined query – is one of the most popular topics in both real life applications and multimedia research [44, 49, 51, 50, 19, 20]. There are vast amount of video archives including broadcast news, documentary videos, meeting videos, movies etc. On the other hand video sharing on the web is growing with a tremendous speed which creates perhaps the most heterogeneous and the largest publicly available video archive [19, 20]. Finding the desired videos is becoming harder and harder everyday for the users. Research on video retrieval is aiming at the facilitation of this task.

In a video there are three main type of information which can be used for video retrieval: visual content, text information and audio information. Even though there are some studies [6] on the use of audio information, it is the least used source for retrieval of videos. Mostly audio information is converted into text using automatic speech recognition (ASR) engines and used as text information. Most of the current effective retrieval methods rely on the noisy text information attached to the videos. This text information can be ASR results, optical character recognition (OCR) texts, social tags or surrounding hypertext information. Nowadays most of the active research is conducted on the utilization of the visual content. Perhaps it is the richest source of information, however analyzing visual content is much harder than analyzing the other two.

There are two main frameworks for video retrieval: text-based and content-based. Text-based methods are originated from the information retrieval community and can be tracked back to 1970s . In these systems retrieval is achieved by using the text information attached to the video. Content-based approaches start in early 1980s with the introduction of content-based image retrieval (CBIR). In content based approaches videos are utilized through the visual features such as color, texture, shape, motion.

Users express their needs in terms of queries. In content-based retrieval there are several types of queries. These queries can be defined with text keywords, video examples, or low-level visual features. In [1] queries are split into three levels:

Level 1: Retrieval by primitive visual features such as color, texture, shape, motion or the spatial location of video elements. Examples of such queries might include “find videos with long thin dark objects in the top left-hand corner”, or most commonly “find more videos that look like this”.

Level 2: Retrieval of the concepts identified by derived features, with some degree of logical inference. Examples of such queries might include “find videos of a bus”, or “find videos of walking”.

Level 3: Retrieval by abstract attributes, involving a significant amount of high-level reasoning about the meaning and purpose of the objects or scenes depicted. Examples of such queries might include “find videos of one or more people walking up stairs”, or “find videos of a road taken from a moving vehicle through the front windshield”.

Levels 2 and 3 together are referred as *semantic video retrieval* [1], and the gap between level 1 and 2 is named as *semantic gap*. More specifically the discrepancy between limited descriptive power of low level visual features and high-level semantics is referred as *semantic gap* [2, 3].

Users in level 1 retrieval are usually required to provide example videos as queries. However it's not always possible to find examples of the desired video content. Moreover example videos may not express the user's intent appropriately. Since people use languages for the main way of communication, the most natural way of expressing themselves is through the words. In level 2 and 3 queries are mostly expressed with the words from the natural languages.

Since level 3 subsumes level 2 , we'll refer level 3 as *semantic retrieval*. And level 2 will be referred as *concept retrieval*. In this proposal we'll focus on these two levels.

With the release of the LSCOM (Large Scale Concept Ontology for Multimedia) [4] lexicon and annotation, a large number of visual content-based semantic concept detectors, which includes objects (e.g. car, people), scenes (e.g. office, outdoor) and events (e.g. walking and marching), have been developed [24, 25]. 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 [36, 38, 39, 40]. In this work, we propose a novel use of these concept detectors to further improve video retrieval.

The main contribution of this proposal is utilization of the *semantic word similarity* measures for the content-based retrieval of videos through concept detectors and the visual co-occurrence relations between concepts. 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 [45, 47], there are not many approaches that explore context of high-level concepts [46, 48]. We propose 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 [35] it is reported that excluding target concept's own detector, 18 out of 39 concepts are 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 similarity 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 proposal, 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 text queries, which is computed using the earth movers distance (EMD).

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

1.1 Related Work

There are two main research communities that explore the context information in visual understanding of images and videos: computer vision and multimedia communities.

In the computer vision community, majority of the studies are on the use of context between low-level features for object detection and localization [45, 47]. Context of low-level features is also harnessed for scene and object segmentation [26]. Although there are some approaches that use the context of high-level concepts [46, 48], it is relatively new in computer vision community.

High-level context is generally used for the post-processing step in order to refine the object detection results obtained from object recognition and segmentation framework. Usually the context information is learned from manually annotated training data sets. There are few other approaches that use semantic taxonomies for high-level semantic reasoning in object detection tasks [28, 27].

In the multimedia community, the high-level context is mostly used for improving concept detection and retrieval accuracy. Traditionally concepts are retrieved using trained concept detectors (e.g. SVM detectors) and then the high-level context is used for refining the results. Many approaches harnessed the high-level context using a second level SVM model where the inputs are the confidences obtained from the individual concept detectors [30, 29, 31]. There are also some probabilistic approaches that exploit high-level context using both directed and undirected probabilistic graphical models [32, 33, 34, 35].

Almost all the approaches descried above uses high-level context learned from the manually annotated data sets. In our study we learn the high-level context from the web and semantic networks. This unsupervised extraction of high-level context enables us to consider about the retrieval of concepts using high-level context alone.

CHAPTER 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:

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

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 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 [7]: corpus-based 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 proposal, we examine eleven different semantic word similarity measures. Seven of them are knowledge-based similarity measures and four of them are corpus-based similarity measures.

2.2.1 Knowledge-based Word Similarity Measures

The knowledge-based measures quantify the similarity between two concepts using the information drawn from the semantic networks. Most of these measure uses WordNet [16] (which is a semantic lexicon for the English language) as the semantic network. The similarity between two concepts and two words is not same. Since one word may have several senses, it can correspond to several concepts. In order to compute the semantic similarity between two words, we compute the similarity using all the senses of both words and then we pick the highest similarity score.

Some of these similarity measures uses information content (IC) which represents the amount of information belonging to a concept. It is described as:

$$IC(c) = -log(P(c)),$$

where $IC(c)$ is the information content of the concept c, and $P(c)$ is the probability of encountering an instance of the concept c in a large corpus. Another used definition is the least common subsumer (LCS) of two concepts in a taxonomy. LCS is the common ancestor of both concepts which has the maximum information content. In the Figure 2.1, LCS is described visually with an example. In the following parts, we discuss seven different knowledge-based similarity measures.

2.2.1.1 Leacock & Chodorow Similarity

This similarity measure is introduced in [9]. The similarity between two concepts is defined as:

$$Sim_{lch}(c_i, c_j) = log\left(\frac{length(c_i, c_j)}{2 \times D}\right),$$

where c_i, c_j are the concepts, $length(c_i, c_j)$ is the length of the shortest path between concepts c_i and c_j using node counting, and D is the maximum depth of the taxonomy.

2.2.1.2 Lesk Similarity

In Lesk measure [10] similarity of two concepts is defined as a function of overlap between the definitions of the concepts provided by a dictionary. It is described as:

$$Sim_{lesk}(c_i, c_j) = \frac{def(c_i) \cap def(c_j)}{def(c_i) \cup def(c_j)},$$

where $def(c)$ represents the words in definition of concept c. This measure is not limited to semantic networks, it can be computed using any electronic dictionary that provides definitions of the concepts.

2.2.1.3 Wu & Palmer Similarity

This similarity metric [12] measures the depth of two given concepts in the taxonomy, and the depth of the LCS of given concepts, and combines these figures into a similarity score:

$$Sim_{wup}(c_i, c_j) = \frac{2 \times depth(LCS(c_i, c_j))}{depth(c_i) + depth(c_j)},$$

where $depth(c)$ is the depth of the concept c in the taxonomy, and $LCS(c_i, c_j)$ is the LCS of the concepts c_i and c_j.

2.2.1.4 Resnik Similarity

Resnik similarity measure [11] is defined as the information content of the LCS of two concepts:

$$Sim_{res}(c_i, c_j) = IC(LCS(c_i, c_j)).$$

2.2.1.5 Lin's Similarity

The key idea in this measure is to find the maximum information shared by both concepts and normalize it. Lin's similarity [8] is measured as the information content of LCS, which can be seen as a lower bound of the shared information between two concepts, and then normalized with the sum of information contents of both concepts. The formulation is as below:

$$Sim_{lin}(c_i, c_j) = \frac{2 \times IC(LCS(c_i, c_j))}{IC(c_i) + IC(c_j)}.$$

2.2.1.6 Jiang & Conrath Similarity

This measures is introduced in [13]. This measure also uses IC and LCS. It is defined as below:

$$Sim_{jnc}(c_i, c_j) = \frac{1}{IC(c_i) + IC(c_j) - 2 \times IC(LCS(c_i, c_j))}.$$

2.2.1.7 Hirst & St-Onge Similairty

This measure is a path based measure, and classifies relations in WordNet as having direction. For example, is-a relations are upwards, while has-part relations are horizontal. It establishes the similarity between two concepts by trying to find a path between them that is neither too long nor that changes direction too often. This similarity measure is represented with Sim_{hso}. Detailed description of this method can be found in [14].

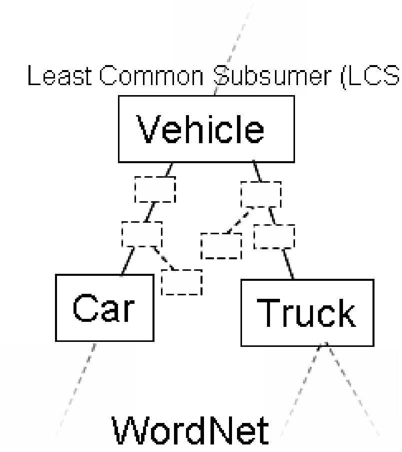


Figure 2.1: In this example LCS of the concepts *car* and *truck* is the *vehicle* in the given taxonomy.

2.2.2 Corpus-based Word Similarity Measures

Corpus-based measures try to identify the similarity between two concepts using the information exclusively derived from large corpora. In this section we focus on PMI-IR similarity measure computed from four different sources.

2.2.2.1 PMI-IR Similairty

The pointwise mutual information using data collected by information retrieval (PMI-IR) was proposed as a semantic word similarity measure in [15]. The main idea behind this measure is that similar concepts tend to occur together in the documents more than 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 two concepts is approximated using a web search engine. The formulation is given as below:

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

where $hits(c_i, c_j)$ is the number of documents that contain c_i, c_j concepts together, $WebSize$ is the approximated number of all documents indexed in the search engine; $hits(c_i), hits(c_j)$ 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].

We use four different sources for computation of Sim_{PMI-IR}. Initially for $Sim_{PMI-IR-WebAND}$ we use Yahoo [18] web search engine, and $hits(c_i, c_j)$ is computed as the number of documents that include both c_i and c_j concepts. In the second measure $Sim_{PMI-IR-WebNEAR}$, we again use the Yahoo web search engine. But in this case with the help of NEAR operator, $hits(c_i, c_j)$ is computed as the number of documents in which c_i, c_j occur in a window of ten words. Third similarity measure $Sim_{PMI-IR-WebImage}$ is obtained from Yahoo image search engine [21], and $hits(c_i, c_j)$ is computed as the number of returned images when we search for c_i and c_j concepts together. The last similarity measure $Sim_{PMI-IR-Flickr}$ is extracted from Flickr image search engine [22], and

$hits(c_i, c_j)$ is computed as the number of returned images when we search for c_i and c_j concepts together.

CHAPTER 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 then 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 proposal, we show that the retrieval of a concept can also be done in an unsupervised manner (without having any annotated video shots of that concept) with 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.

CHAPTER 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 [36, 38, 39, 40].

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) [41]. The overview of the method is visually

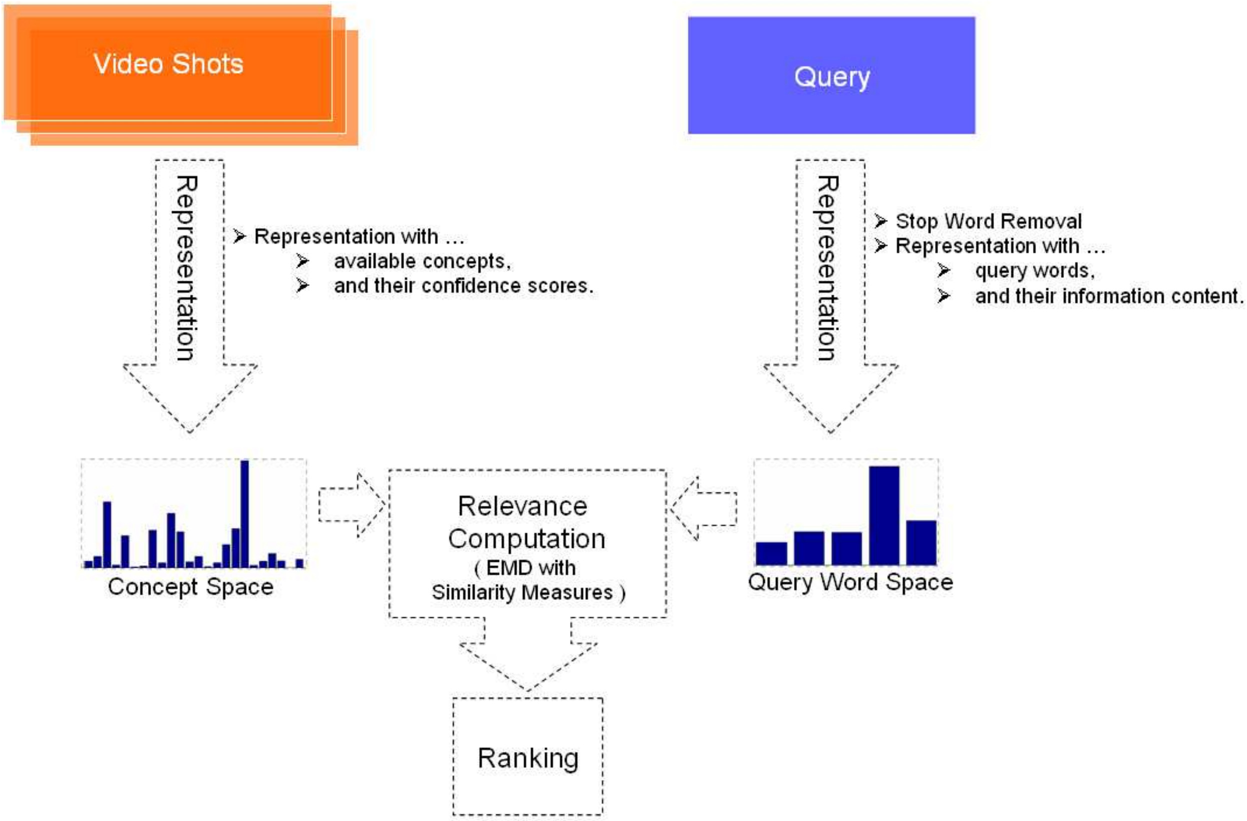


Figure 4.1: An overview of the Visual Content-based Semantic Video Retrieval method.

described in Figure 4.1. 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 using 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 4.2). 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:

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

with the following constraints:

$$Constraints : \sum_{i} f_{i,j} = b_j, \sum_{j} f_{i,j} = a_i,$$

$$f_{i,j} \geq 0, 1 \leq i \leq Q, 1 \leq j \leq M,$$

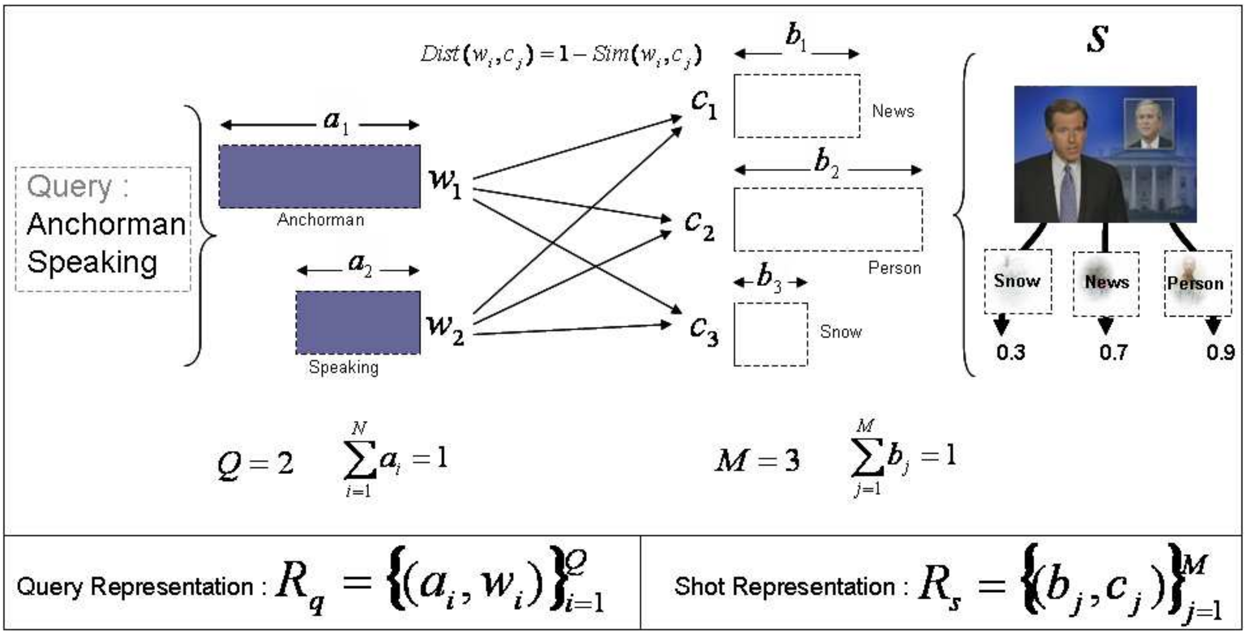

$Dist(w_i, c_j) = 1 - Sim(w_i, c_j)$
Query : Anchorman Speaking
a_1
Anchorman
w_1
a_2
Speaking
w_2
c_1
c_2
c_3
b_1
News
b_2
Person
b_3
Snow
S
Snow
News
Person
0.3
0.7
0.9
$Q = 2$ $\sum_{i=1}^{N} a_i = 1$
$M = 3$ $\sum_{j=1}^{M} b_j = 1$
Query Representation : $R_q = \{(a_i, w_i)\}_{i=1}^{Q}$
Shot Representation : $R_s = \{(b_j, c_j)\}_{j=1}^{M}$

Figure 4.2: EMD based distance computation between the visual content and the query.

where $EMD(R_q, R_s)$ is the distance between query q and shot s, $f_{i,j}$ is the flow between bin pairs, and the F is the overall flow configuration which is optimized for the minimum amount of the work. The distances between bin pairs are 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 [42]. 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.

CHAPTER 5
EXPERIMENTS

For evaluation purposes, we use the high level feature extraction and automatic search test data set of TRECVID'06 and TRECVID'07 benchmarks [5]. 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. TRECVID'07 test data set consists of 100 hours of news video entirely in German and is split into 18,142 shots.

Our development set is the common development set for TRECVID'05 and TRECVID'06. It contains 80 hours of video from American, Arabic and Chinese TV channels and is split into 61,901 shots. We also have annotations of LSCOM concepts for each shot in the development set. We use two set of 374 concept detectors which are released by Columbia University [24] and City University of Hong-Kong [25]. These detectors are trained using development set. These detector sets will be referred as *Columbia* and *Vireo* detector sets.

For the knowledge-based semantic words similarity measures we use the Wordnet::Similarity package released by [17]. And we compute PMI-IR similarity from Yahoo web search engine [18], Yahoo image search engine [21], and Flicker image search engine [22]. Information content is computed using Yahoo web search engine [18].

For the EMD optimization problem we use source code provided by [41], with some manipulations for our needs. As a comparison metric, we use average precision (AP) which emphasizes returning more relevant documents earlier. It is the average of precisions computed after truncating the list after each of the relevant documents. Considering that AP is computed from the complete video shot database, even small values between 3% - 30% lead very nice retrieval results. We also use mean average precision (MAP) (which is the mean of average precision results for all the cases) for comparison of methods in overall.

Using all these resources, we evaluate both of our methods, a method for retrieving new concept and a method for visual content-based semantic retrieval. For retrieving a new concept we use two subset of detectors. First subset is extracted from 374 concept detectors by checking if associated concepts exist in WordNet or not. This set contains 201 concepts. The second set is the complete set of concepts which includes 374 concept detectors. For semantic retrieval we only use complete detector set. Before starting the evaluations of these methods, we'll discuss the development and test data sets and we'll give some motivation for the feasibility of video retrieval using only high-level context.

5.1 Data Analysis

We have three data sets: development set, TRECVID'06 and TRECVID'07 test sets. First we'll discuss the strength of the context and frequency of concepts using the development set. Then we'll give some insight information about TRECVID'06 and TRECVID'07 test sets.

There are two concept sets: 201 concepts and 374 concepts. Former will be referred as the small concept set and the latter will be referred as the complete concept set. Although some of these concepts have very high frequencies, most of them have relatively small frequencies in the

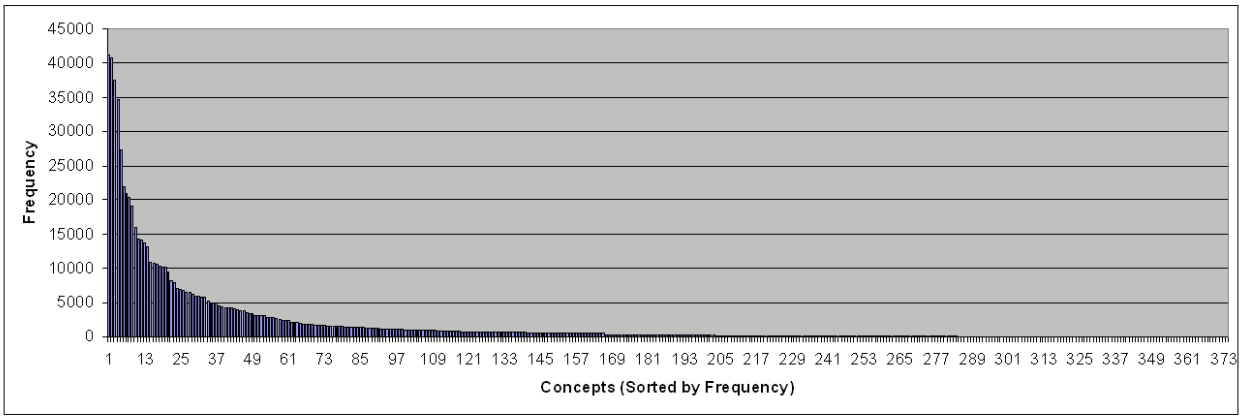


Figure 5.1: Frequency of the concepts in descending order using the complete concept set and the development data.

development set. In the small concept set, 5.9% of the concepts has more than 10,000 occurrences, 23.3% of the concepts has more than 1000 occurrences and 31.3% of the concepts has less than 100 occurrences. Distribution of concept frequencies for the small set is shown in figure 5.3. In the complete concept set, 5.3% of the concepts has more than 10,000 occurrences, 26% of the concepts has more than 1000 occurrences and 31% of the concepts has less than 100 occurrences. Distribution of concept frequencies for the complete set shown in figure 5.1. And there are 20 concepts which have less than 20 occurrences in the complete set.

Since the development set is extracted from broadcast news, it is very challenging in nature. Sometimes even two people may disagree about the existence of a concept. Detection of the concepts in this data set is not as easy as detection of concepts in Caltech [52], Pascal [53] or LabelMe [54] data sets. As a consequence of both the complexness of the detection task and low frequencies of concepts, trained detectors have relatively low precisions when compared to detectors trained

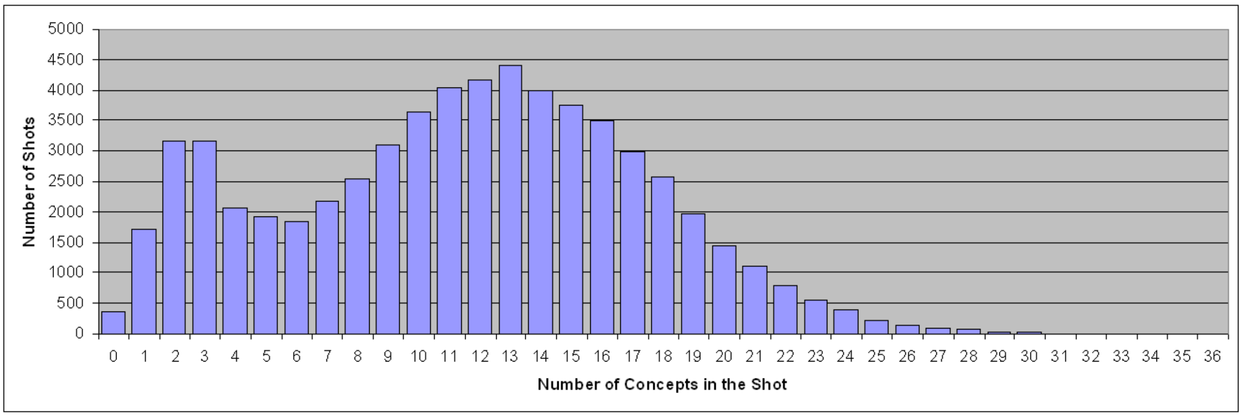


Figure 5.2: The distribution of the number of concepts in the shots using the complete concept set and the development data.

and tested on other data sets. Nevertheless none of the Caltech, Pascal, LabelMe data sets has wide coverage (<40) of concepts. Since in our problem we retrieve the videos using only high level context, the wide coverage of concepts makes the TRECVID data set a good candidate for testing our methods.

Since we only use high-level context, we need strong contextual relations between concepts in order to have accurate retrieval results. When we think of a real life scene, we can easily find many concepts occurring in that scene. For instance, most of the scenes are either indoor or outdoor scenes, therefore we'll at least see one concept in the scene. We can find numerous concepts just in one shot. We analyzed our development set for this purpose to see the strength of the context. Using small concept set, we find out that only 5.8% of the shots contain one concept. And most of the shots contain 7 or 8 concepts. Using complete concept set, 2.7% of the concepts contain one concept. And most of the shots contain 11,12,13 or 14 concepts. Distributions of concept number

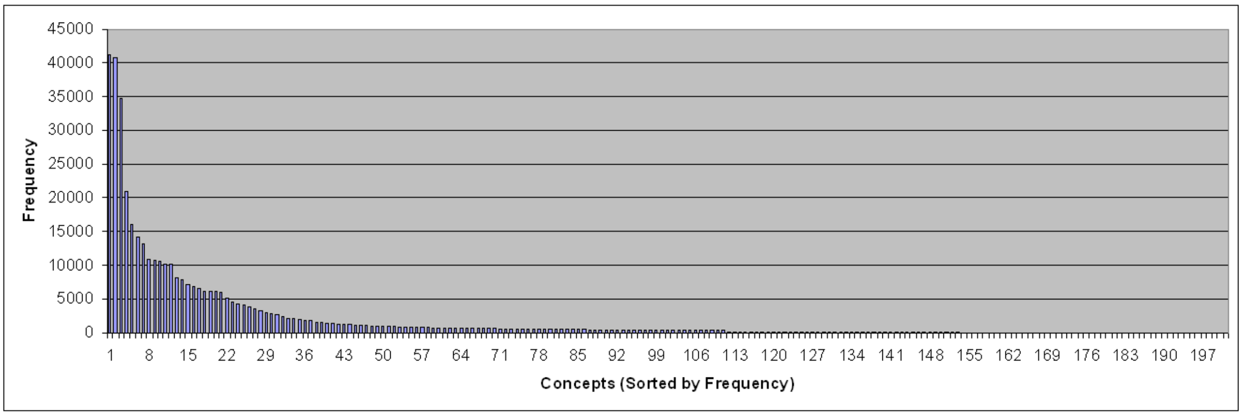


Figure 5.3: Frequency of the concepts in descending order using the small concept set and the development data.

in the shots can be seen in the figure 5.2 and 5.4. Since 97.3% of the shots in the complete set have more than one concept, in each of these shots we can predict a score for the existence of concepts using context. The certainty of detection depends on the strength of the context. As a result of these analysis, it is clear that when we increase the number of concepts the number shots that contain one concept is decreasing, and the high-level context is becoming stronger.

For testing purposes we use TRECVID'06 and TRECVID'07 test sets. Even though they are both broadcast news, the TRECVID'07 test set is more challenging than TRECVID'06. Additionally the queries in TRECVID'07 has less named entities and harder than TRECVID'06 queries. This fact decreases the performance of text-based retrieval in TRECVID'07. For instance, one of the queries in TRECVID'07 is "Find shots of a door being opened". Obviously no one will say "I am opening the door" while performing this action. On the other hand, one example of

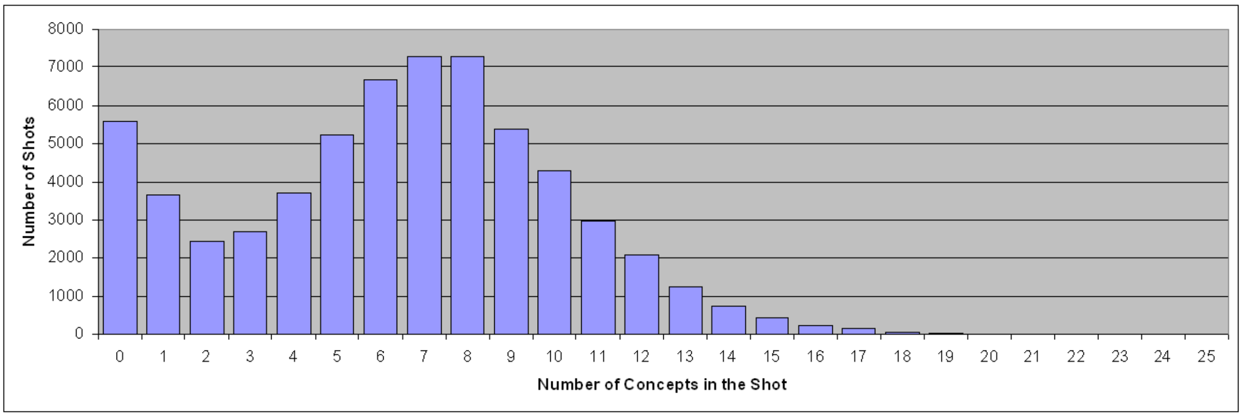


Figure 5.4: The distribution of the number of concepts in the shots using the small concept set and the development data.

TRECVID'06 queries is “Find shots of US Vice President Dick Cheney”. And most probably before or during related shots anchorman will use one of the query words.

5.2 Evaluation of Retrieving New Concept

In this evaluation, we used TRECVID'06 and TRECVID'07 (high level feature extraction) test concepts. Both of the test sets contain twenty test concepts including events such as 'people-marching', scenes such as 'office' or object classes such as 'bus'. We examine retrieving new concept method with knowledge-based and corpus-based similarity measures. Since almost all knowledge-based similarity measures uses WordNet, in the first experiment we used small concept set (concepts which has WordNet entries) for testing the performance of all similarity measures. We performed this test on 13 test concepts because only these concepts are in the small concept set.

In the second experiment, we evaluated our method using complete concept set with corpus-based similarity measures.

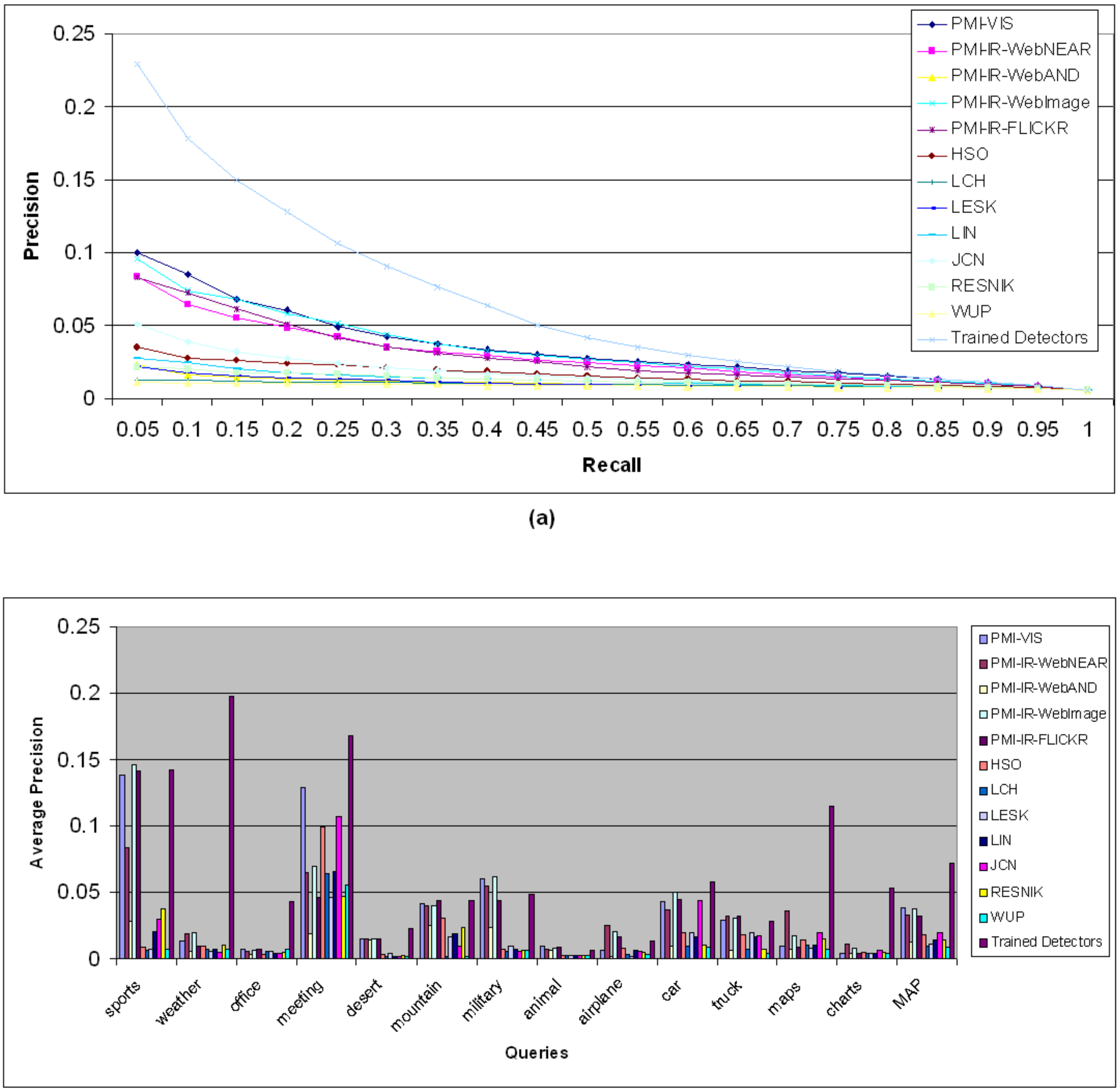
In fact, we also have access to the trained concept detectors for test concepts from the set of complete concept detectors. Since test concepts should be new concepts, during the evaluation of the retrieving new concept method we discarded the associated test concept detectors and used the remaining detectors and the similarity measures. Also, we used these test detectors for the comparison purpose.

5.2.1 Experiments with the Small Concept Set

In this experiment, we performed twelve different evaluations on our retrieval method using visual co-occurrence, seven knowledge-based similarities and four corpus-based similarities. These evaluations are performed using Columbia detector set.

In general visual co-occurrence performed very well and it outperformed the trained detectors on many concepts. Corpus-based similarity measures had almost the same performance with visual co-occurrence. Although HSO and JCN had some nice results, knowledge-based similarity measures didn't perform very well.

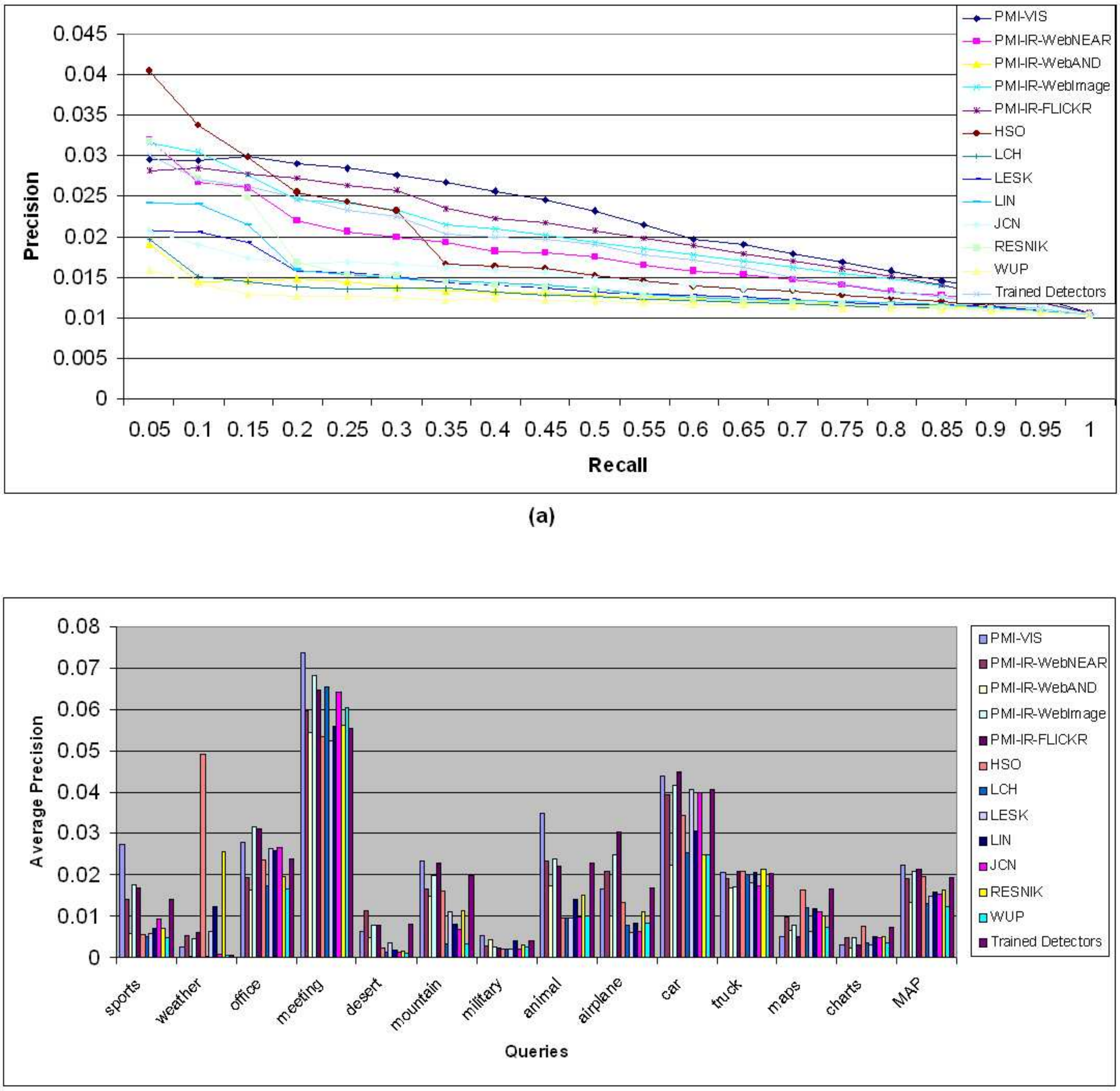
In TRECVID'07, using visual co-occurrence our method outperformed the trained concept detectors on 9 out of 13 concepts. Overall, its performance is better than trained concept detectors. Except PMI-IR-WebAND, corpus-based similarity measures had almost the same performance with visual co-occurrence. Knowledge-based similarity measures didn't perform well. Only HSO



(a)

(b)

Figure 5.5: (a) demonstrates average precision-recall curves for different concept retrieval methods on TRECVID'06 test data. (b) shows AP results for each concept using all concept retrieval methods on TRECVID'06 test data. MAP (Mean Average Precision) is shown at far right. Both of the results are obtained from small concept set and Columbia detectors.



(a)

(b)

Figure 5.6: (a) demonstrates average precision-recall curves for different concept retrieval methods on TRECVID'07 test data. (b) shows AP results for each concept using all concept retrieval methods on TRECVID'07 test data. MAP (Mean Average Precision) is shown at far right. Both of the results are obtained from small concept set and Columbia detectors.

Method Name			MAP’06	MAP’07
Retrieving New Concept	PMI-VIS		3.9%	2.2%
	Corpus-based Similarity Measures	PMI-IR-WebNEAR	3.3%	1.9%
		PMI-IR-WebAND	1.2%	1.4 %
		PMI-IR-WebImage	3.8%	2.1%
		PMI-IR-FLICKR	3.2%	2.1 %
	Knowledge-based Similarity Measures	HSO	1.8%	2.0 %
		LCH	1%	1.3 %
		LESK	1.2%	1.5 %
		LIN	1.4%	1.6 %
		JCN	2%	1.5%
		RESNIK	1.4%	1.6%
		WUP	0.9%	1.2 %
Trained Detectors			7.2%	1.9%

Table 5.1: Overall MAP(Mean Average Precision) comparison of trained concept detectors and retrieving new concept methods using visual co-occurrence, knowledge-based similarity measures and corpus-based similarity measures on TRECVID’06 and TRECVID’07 data with small concept set and Columbia detectors.

performed better than trained concept detectors, but it was not better than corpus-based measures. The results on TRECVID'07 is shown in figure 5.6.

In TRECVID'06, using visual co-occurrence our method outperformed the trained concept detectors on 3 out of 13 concepts. Overall, its performance is close to trained concept detectors on many concepts. Corpus-based similarity measures performed similar to visual co-occurrence. Knowledge-based similarity measures didn't perform well. The best of knowledge-based measures was JCN which has a very poor performance when compared to corpus based measures.The results on TRECVID'06 is shown in figure 5.5.

The main purpose of this experiment is comparison of knowledge-based and corpus-based similarity measures for retrieving new concepts. From the results, its clear that corpus-based measures perform much better than knowledge-based measures. The overall results are shown in table 5.1. The quality of corpus-based measures will be elaborated in the next experiments.

5.2.2 Experiments with the Complete Concept Set

In this experiment, we performed five different evaluations on our retrieval method using visual co-occurrence and four corpus-based similarities. These evaluations are performed using complete set of both Columbia and Vireo detectors.

In TRECVID'07, retrieving new concept methods performed very well. Using Columbia detectors, new concepts are retrieved better than trained concept detectors in overall. Using Vireo detectors, the overall performance of the methods was very close to the trained concept detec-

Method Name		Columbia Detectors		Vireo Detectors	
		MAP’06	MAP’07	MAP’06	MAP’07
Retrieving New Concept	PMI-VIS	3.8%	2.8%	5.6%	3.4%
	PMI-IR-WebNEAR	4.3%	2.4%	5.5%	3.4%
	PMI-IR-WebAND	1.7%	1.8 %	2.2%	2%
	PMI-IR-WebImage	3.8%	2.7%	5.7%	3.2%
	PMI-IR-FLICKR	3.1%	3%	4.7%	2.9%
Trained Detectors		6.6%	2.1%	10.1%	4.4%

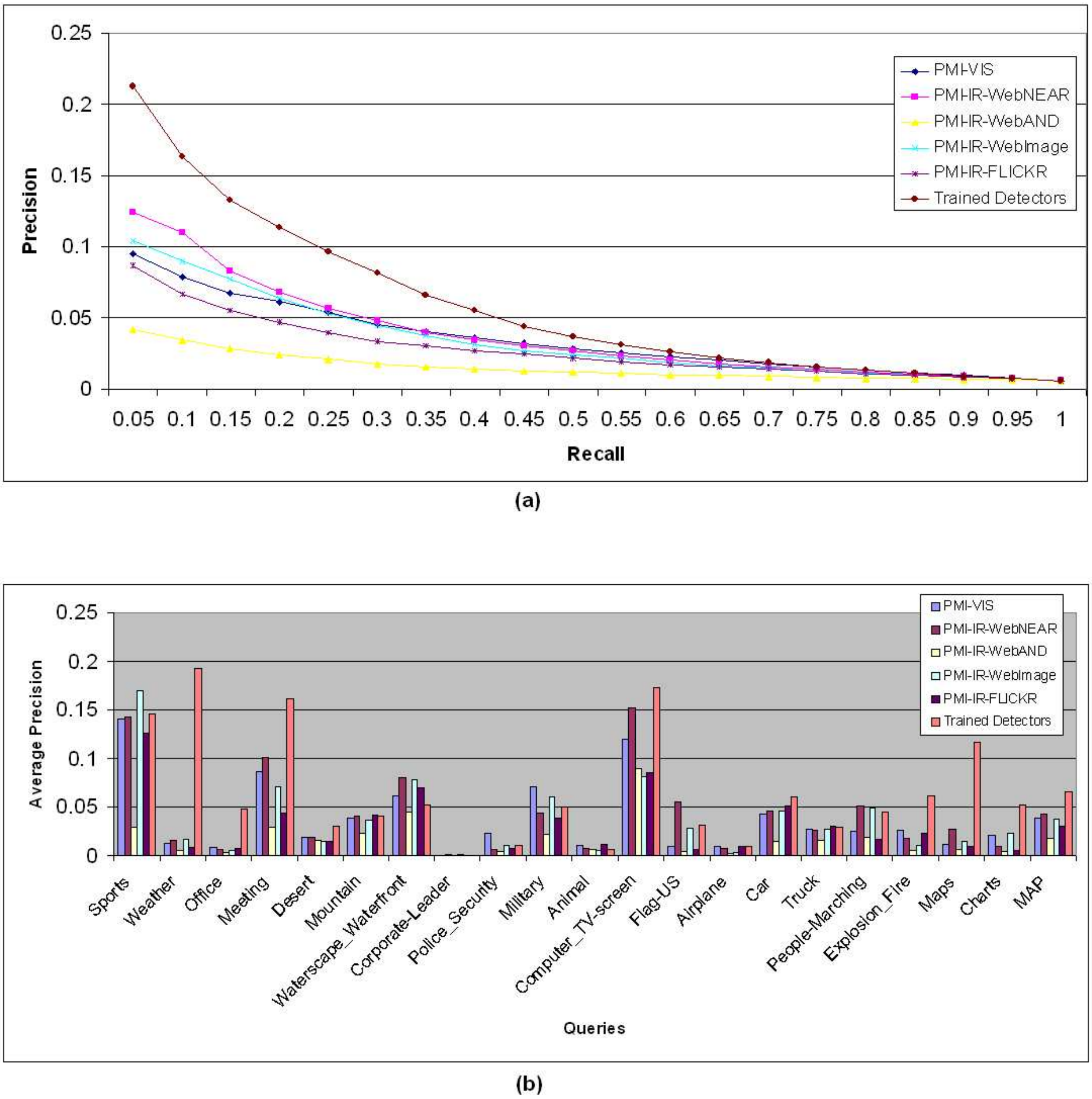
Table 5.2: Overall MAP(Mean Average Precision) comparison of trained concept detectors and retrieving new concept methods using visual co-occurrence and corpus-based similarity measures on TRECVID’06 and TRECVID’07 data with complete concept set.

tors. The performance of the methods in TRECVID'07 using Columbia and Vireo detectors are presented in figure 5.7 and 5.9, respectively.

In TRECVID'06, performance of retrieving new concept methods was reasonably well. In overall, using both Columbia and Vireo detectors the best MAP of retrieving new concept methods was approximately 60% of the MAP of trained detectors. The performance of the methods in TRECVID'06 using Columbia and Vireo detectors are presented in figure 5.8 and 5.10, respectively.

In general our method with visual co-occurrence performed well. It outperformed the trained detectors on many concepts and evaluations. Except the PMI-IR-WebAND, all the corpus-based similarity measures had similar results with visual co-occurrence. Particularly PMI-IR-WebNEAR and PMI-IR-WebImage performed very well. For some cases they even outperformed the visual co-occurrence. PMI-IR-Flicker also had good results but not as good as PMI-IR-WebNEAR and PMI-IR-WebImage in overall. As it is shown in the figures 5.7, 5.8, 5.9 and 5.10, our retrieval method using visual co-occurrence and corpus-based similarities have almost the same precision for several recall values. Overall MAP comparison is presented in table 5.2. Through these experiments, we observed that retrieving new concepts using high-level context is possible. And we can perform this task in an unsupervised way with the help of corpus based similarity measures.

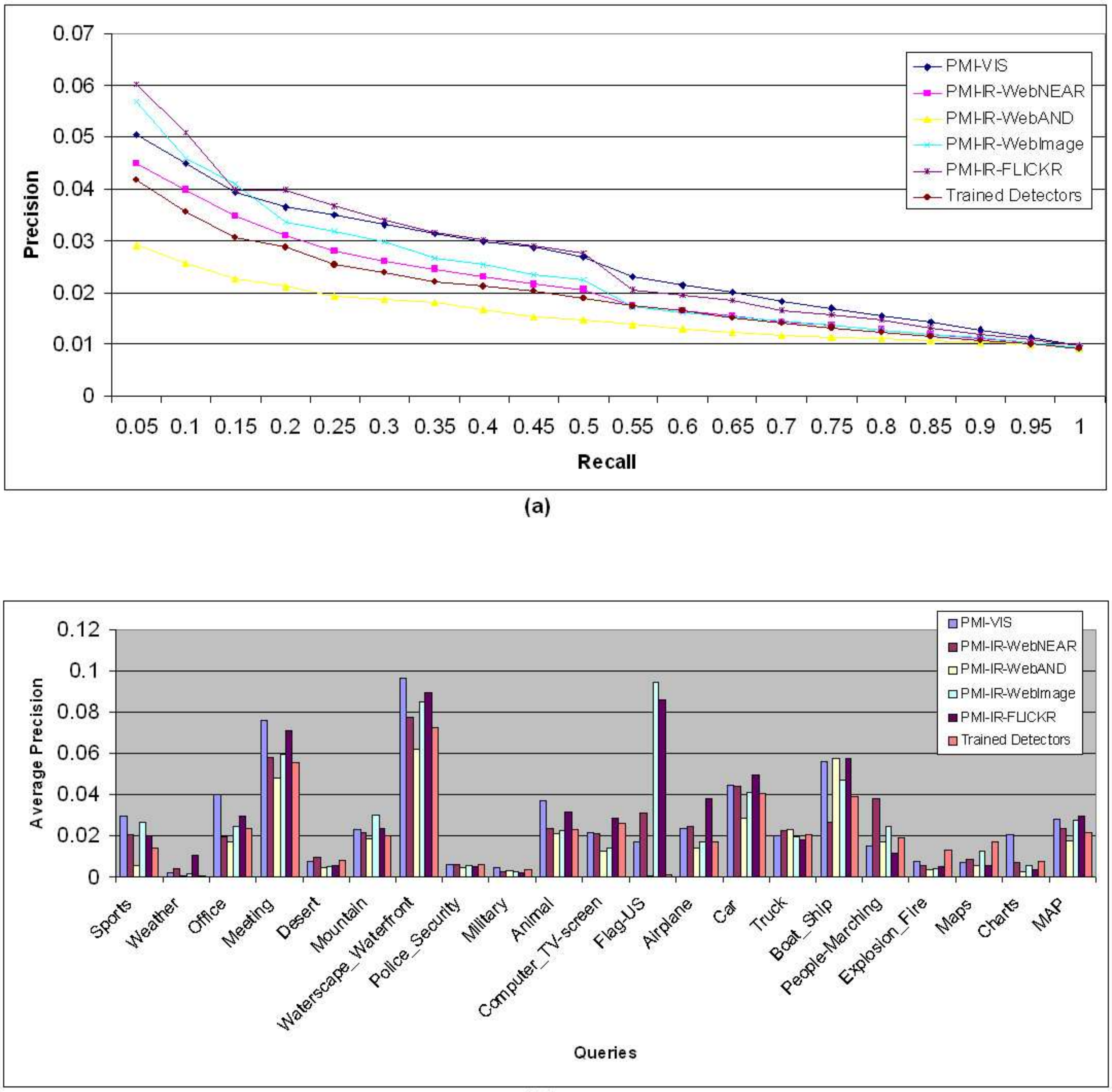
In order to make more certain judgements we analyzed the methods for each concept. For 'sports, waterfront, corporate leader, police, military, animal, airplane, people marching,' concepts our method has significantly better performance than the trained detectors. These concepts mostly have strong contextual relations. Conversely, the concepts that mostly appear in isolation, and have



(a)

(b)

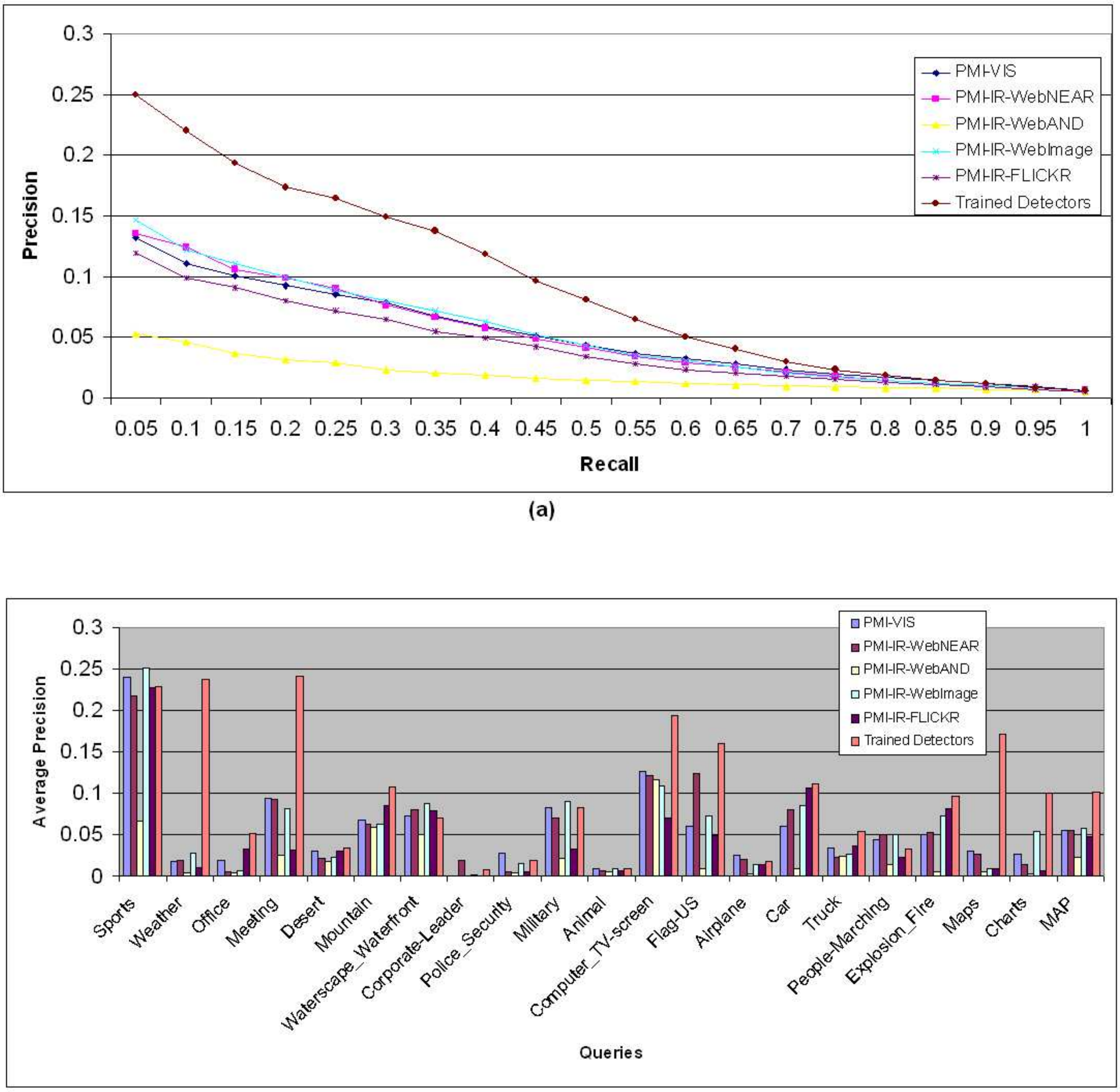
Figure 5.7: (a) demonstrates average precision-recall curves for different concept retrieval methods on TRECVID'06 test data. (b) shows AP results for each concept using several concept retrieval methods on TRECVID'06 test data. MAP (Mean Average Precision) is shown at far right. Both of the results are obtained from complete concept set and Columbia detectors.



(a)

(b)

Figure 5.8: (a) demonstrates average precision-recall curves for different concept retrieval methods on TRECVID'07 test data. (b) shows AP results for each concept using several concept retrieval methods on TRECVID'07 test data. MAP (Mean Average Precision) is shown at far right. Both of the results are obtained from complete concept set and Columbia detectors.



(a)

(b)

Figure 5.9: (a) demonstrates average precision-recall curves for different concept retrieval methods on TRECVID'06 test data. (b) shows AP results for each concept using several concept retrieval methods on TRECVID'06 test data. MAP (Mean Average Precision) is shown at far right. Both of the results are obtained from complete concept set and Vireo detectors.

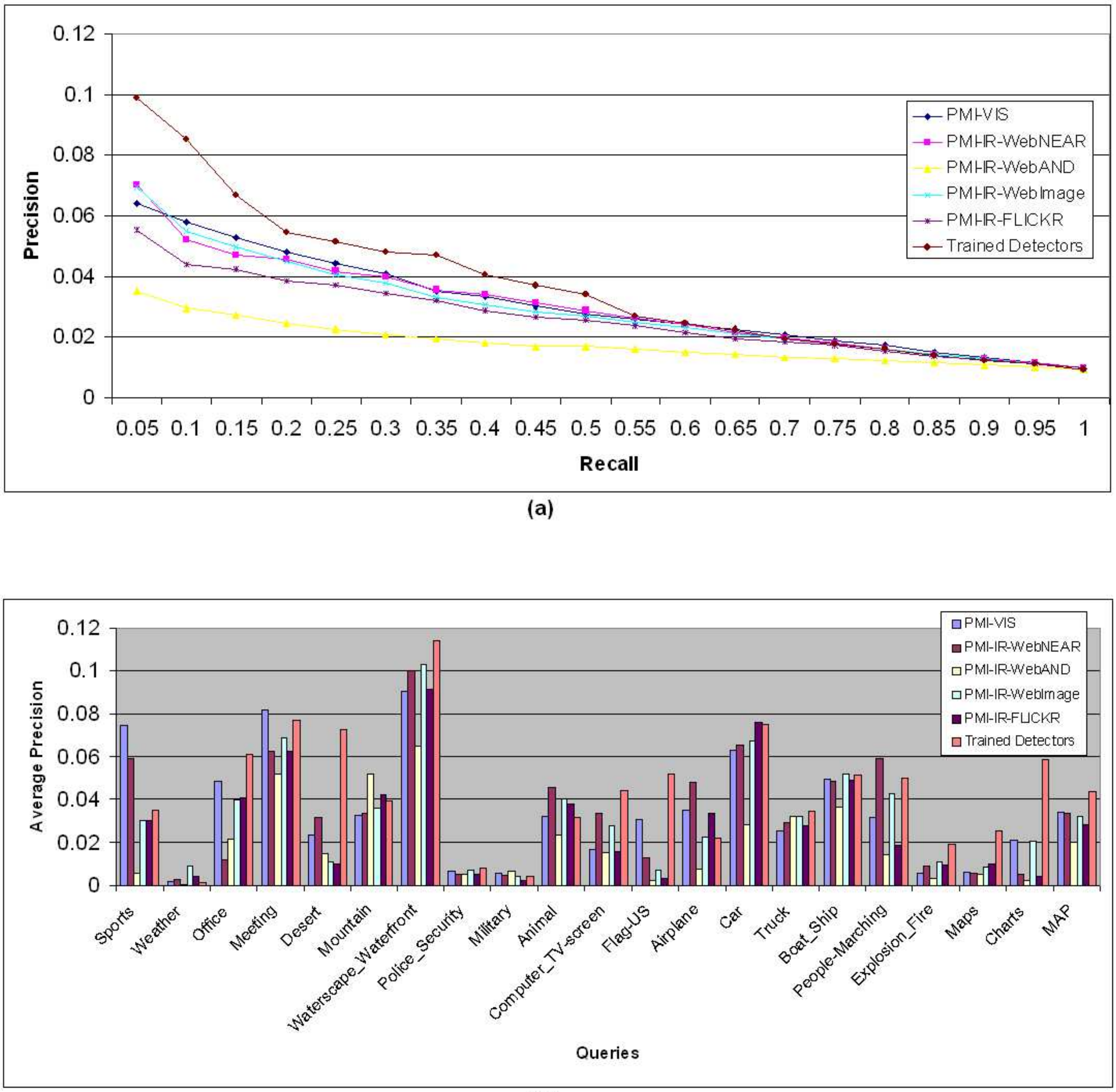


Figure 5.10: (a) demonstrates average precision-recall curves for different concept retrieval methods on TRECVID'07 test data. (b) shows AP results for each concept using several concept retrieval methods on TRECVID'07 test data. MAP (Mean Average Precision) is shown at far right. Both of the results are obtained from complete concept set and Vireo detectors.

loose contextual relations such as 'screen, charts, maps, weather' couldn't be retrieved as well as others using our method. The main reason for the low MAP of the retrieving new concept methods is contextually weak concepts. When we exclude these contextually weak concepts our methods have almost the same performance with trained detectors.

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.

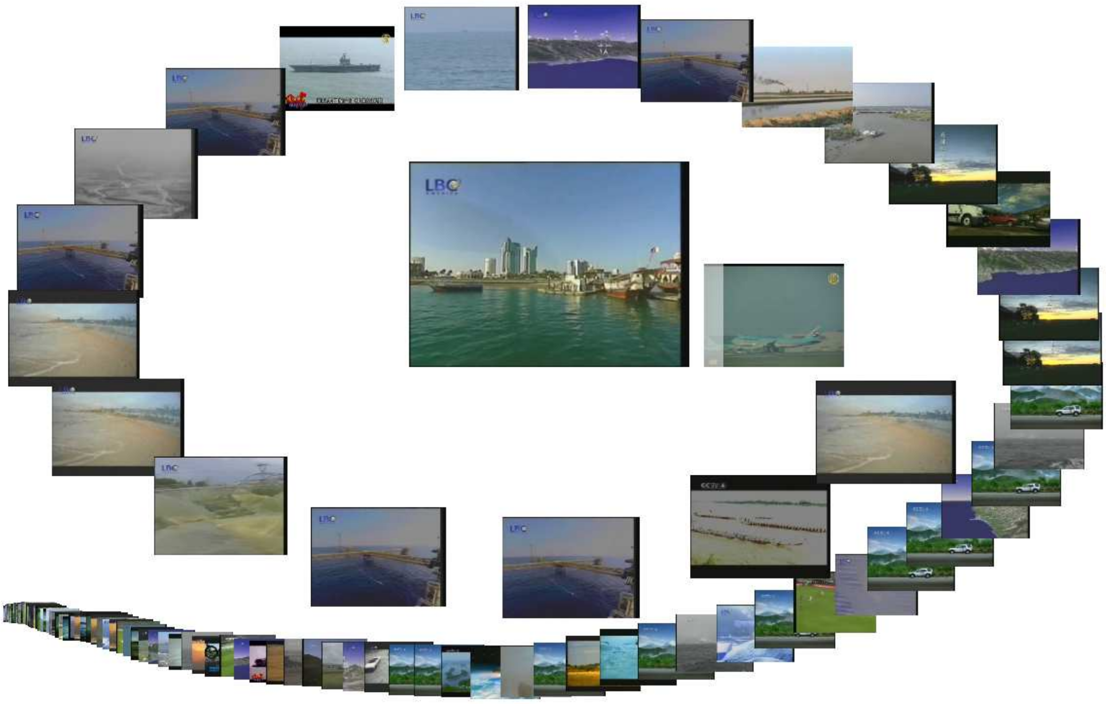


Figure 5.11: Top 100 retrieval results for the new concept 'fishing'.

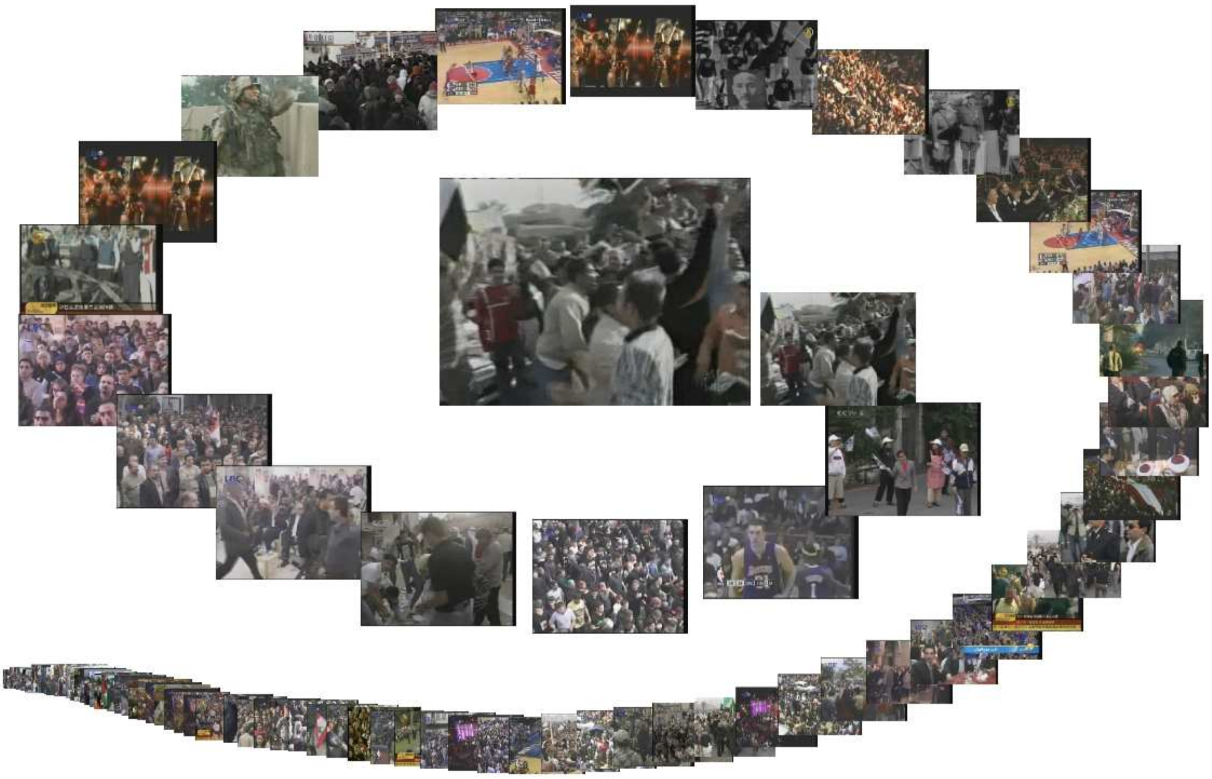


Figure 5.12: Top 100 retrieval results for the new concept 'shouting'.

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-WebNEAR similarity. Figure 5.12 and 5.11 show the retrieved shots for the "shouting" and "fishing" concepts, respectively. Indeed, we don'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

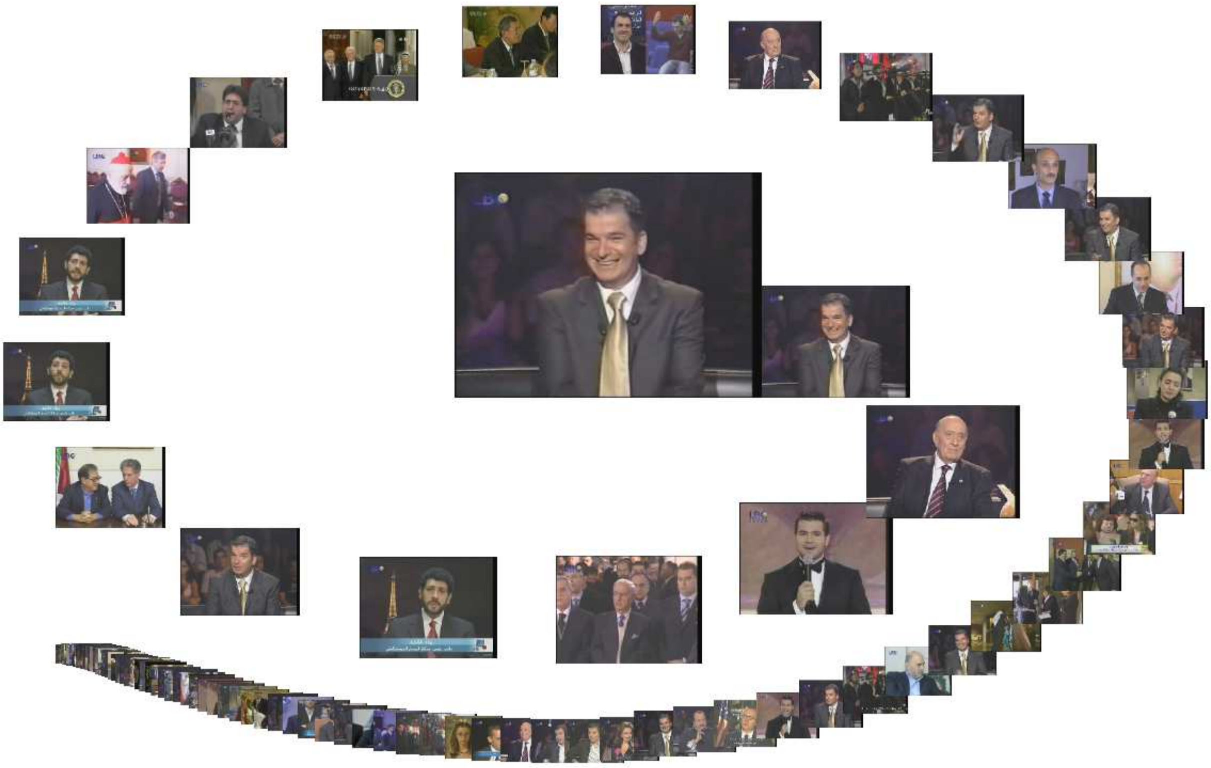


Figure 5.13: Top 100 retrieval results for the new concept 'politics'.

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. More qualitative results are shown in figures 5.13 and 5.14.

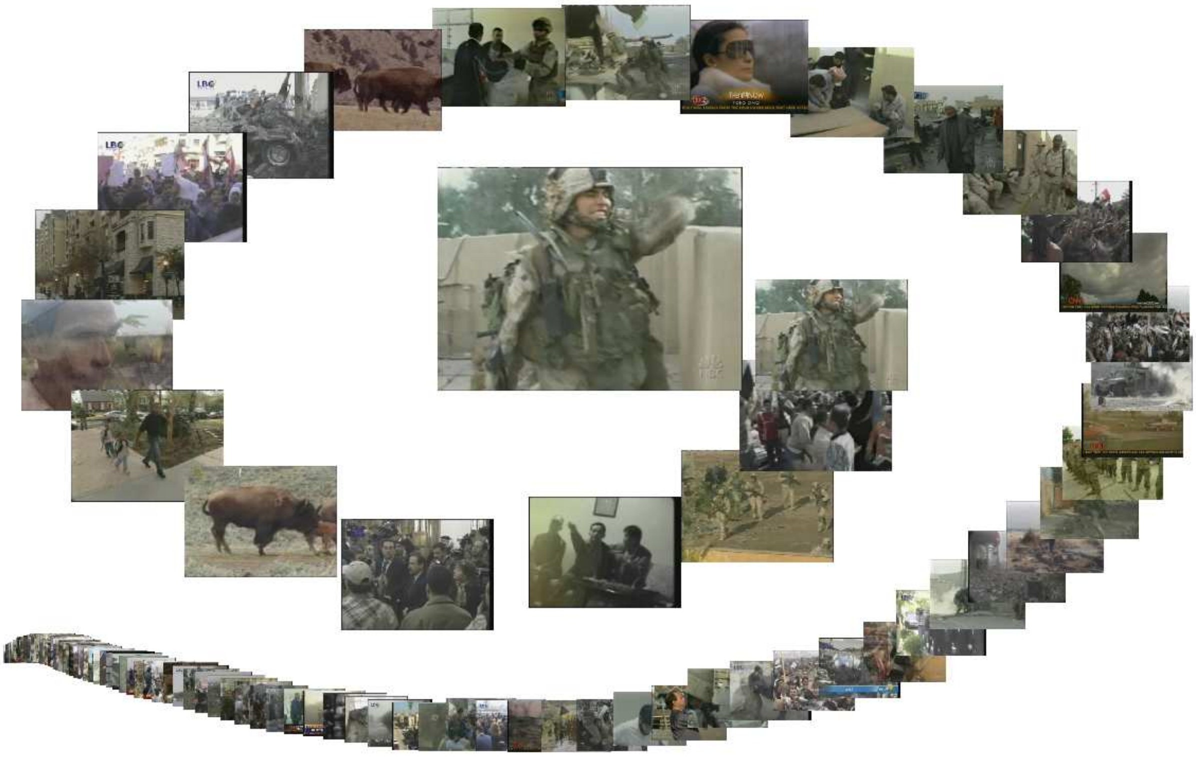
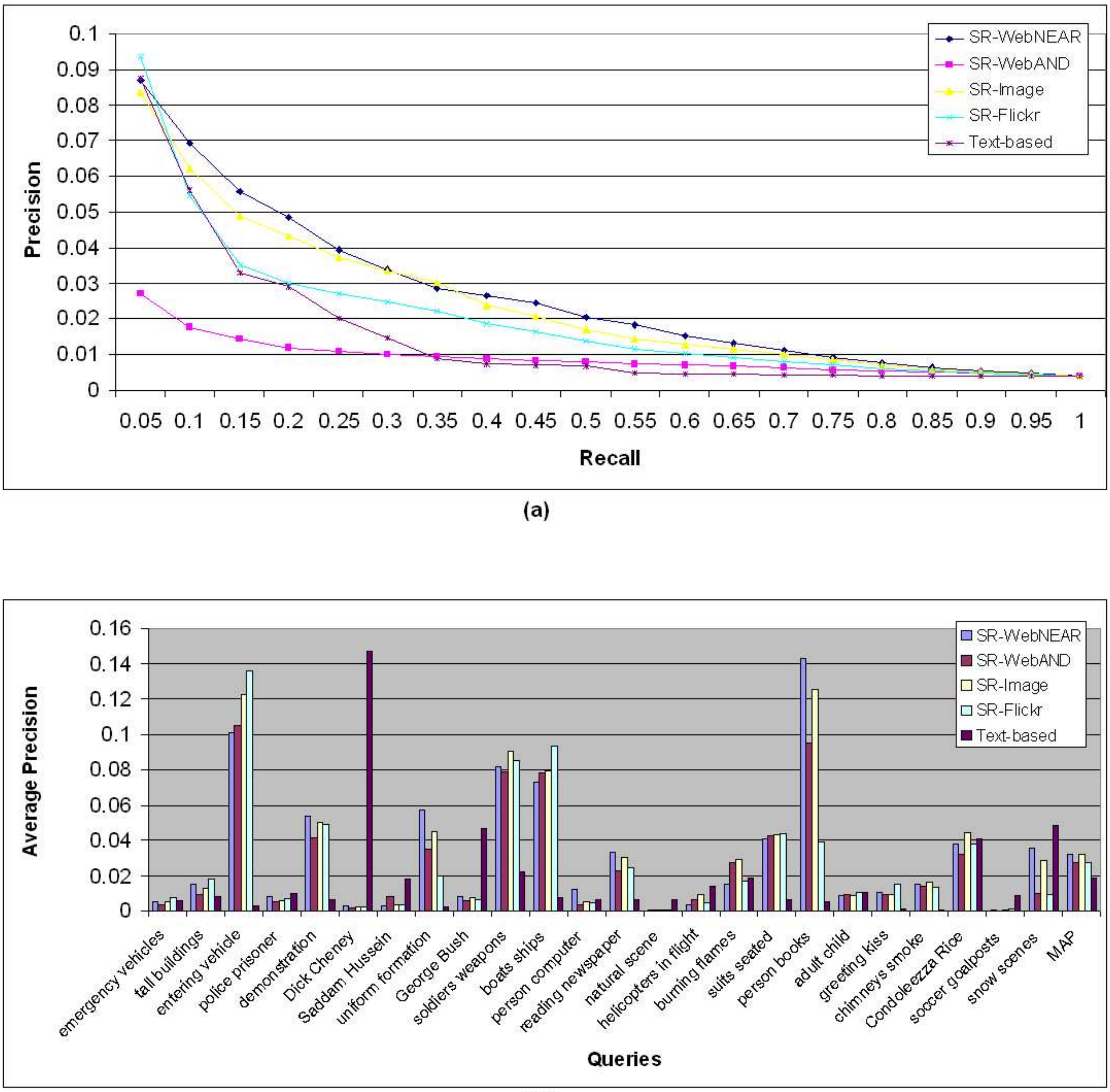


Figure 5.14: Top 100 retrieval results for the new concept ’war’.

5.3 The Semantic Retrieval Evaluations

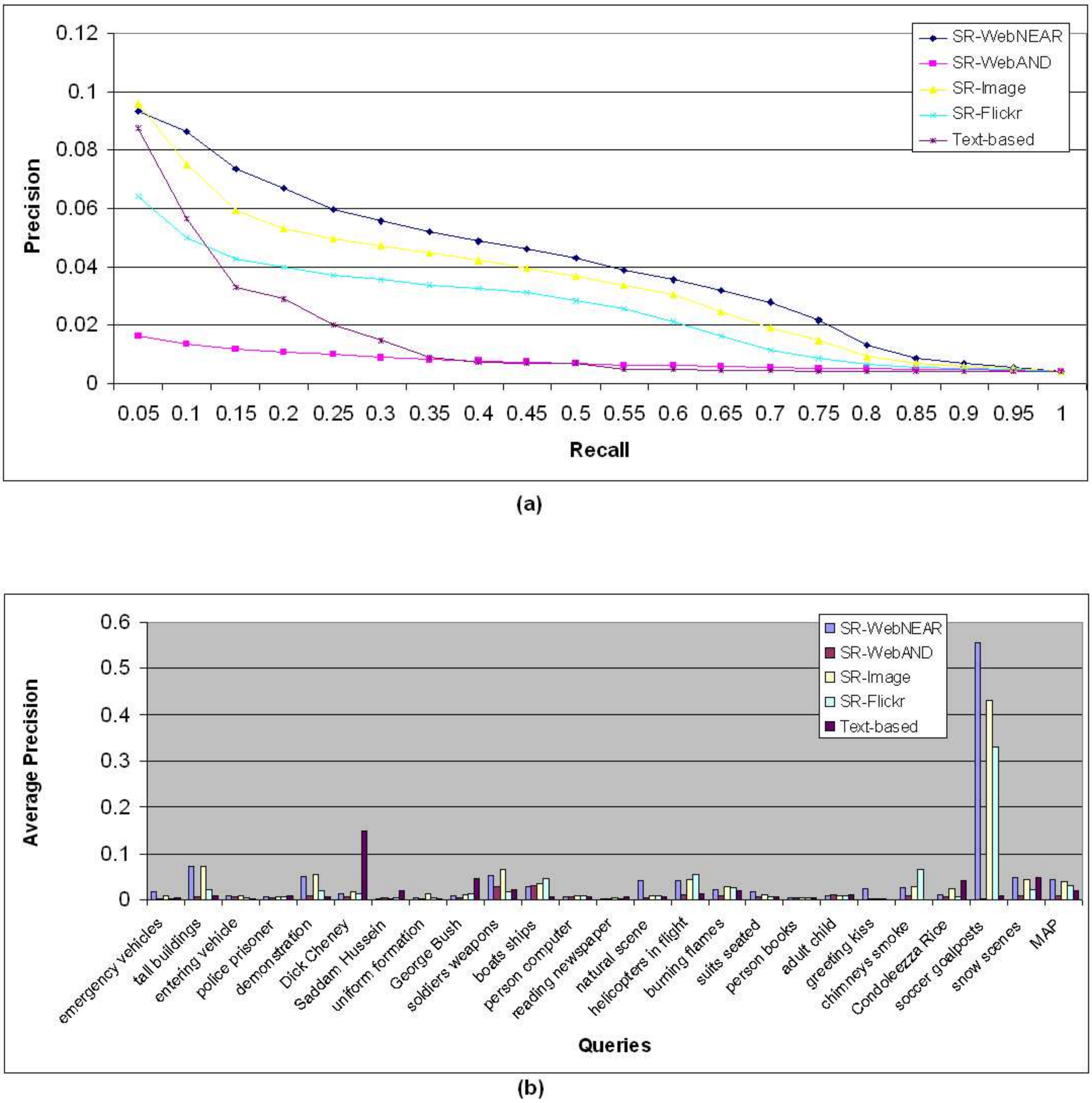
These evaluations are performed on TRECVID’06 and TRECVID’07 test sets, and associated queries of automatic search challenge. These 48 queries include events such as ’walking, greeting’, objects such as ’computer, book, door’ and also some named entities such as ’Saddam Hussein, Dick Cheney, George Bush’. The exact expressions of the queries can be found on the official website [23] 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. We tested VC-SR with four corpus-based similarity measures.



(a)

(b)

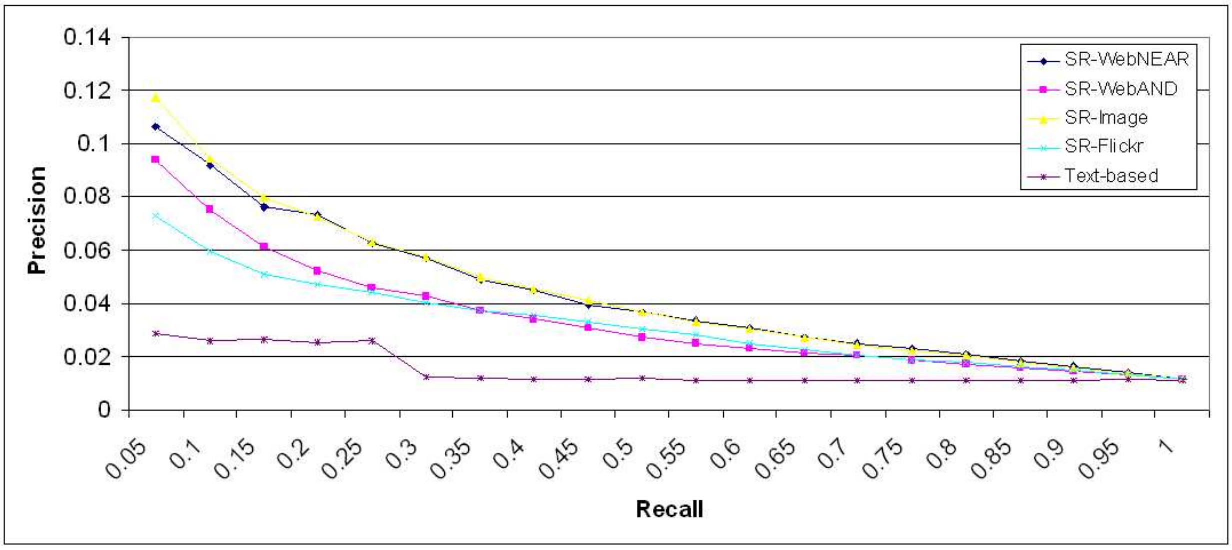
Figure 5.15: (a) demonstrates average precision-recall curves for different semantic retrieval methods on TRECVID'06 test data. (b) shows AP results for each query (represented by some selected words) using several semantic retrieval methods on TRECVID'06 test data. MAP (Mean Average Precision) is shown at far right. Both of the results are obtained from complete concept set and Columbia detectors.



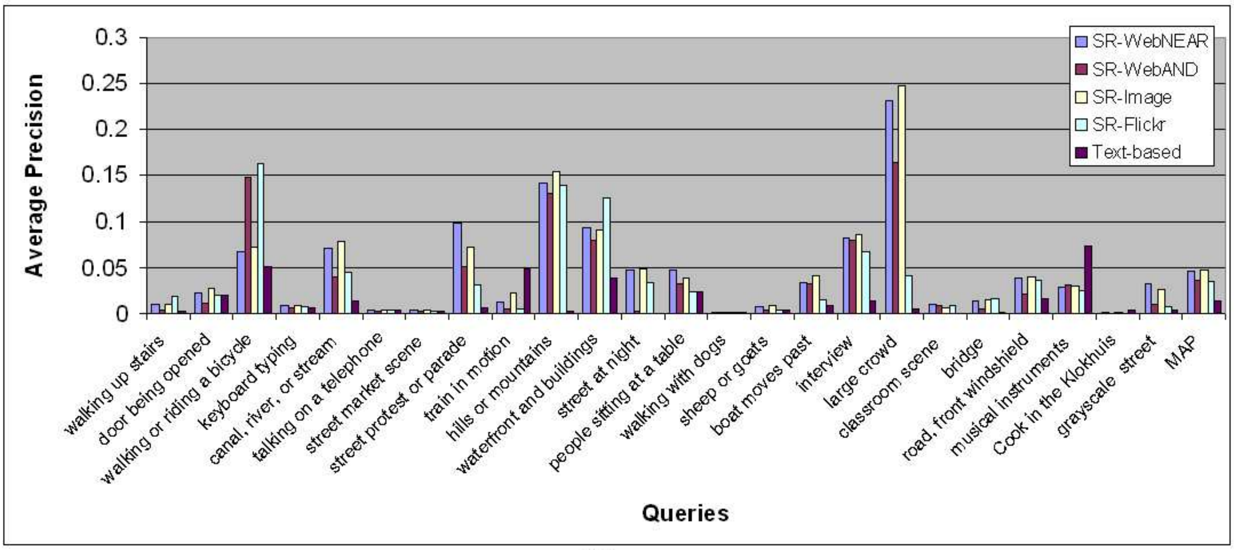
(a)

(b)

Figure 5.16: (a) demonstrates average precision-recall curves for different semantic retrieval methods on TRECVID'06 test data. (b) shows AP results for each query (represented by some selected words) using several semantic retrieval methods on TRECVID'06 test data. MAP (Mean Average Precision) is shown at far right. Both of the results are obtained from complete concept set and Vireo detectors.

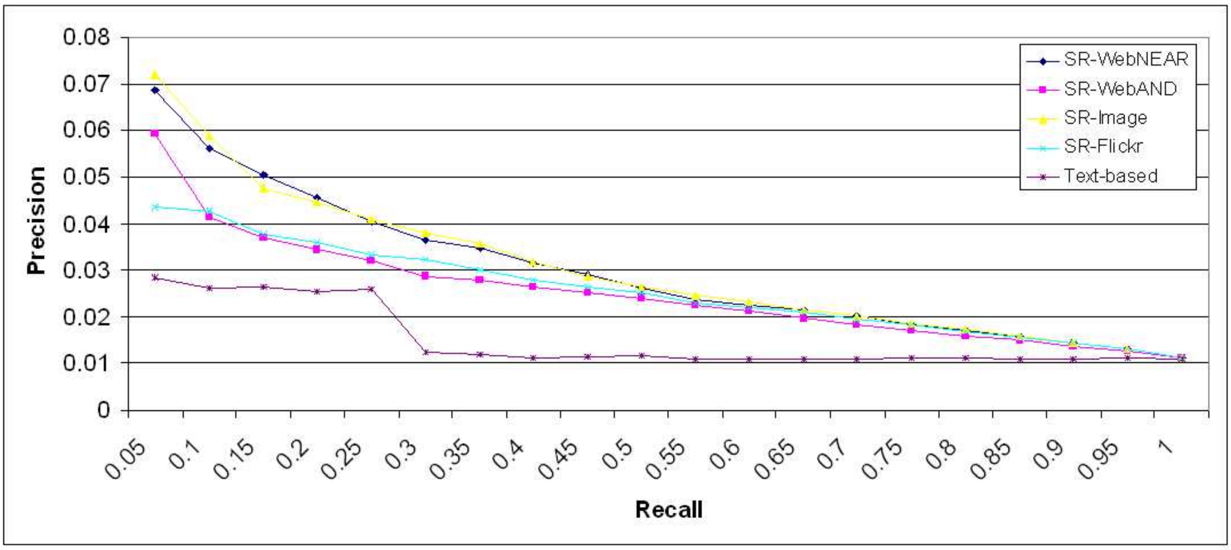


(a)

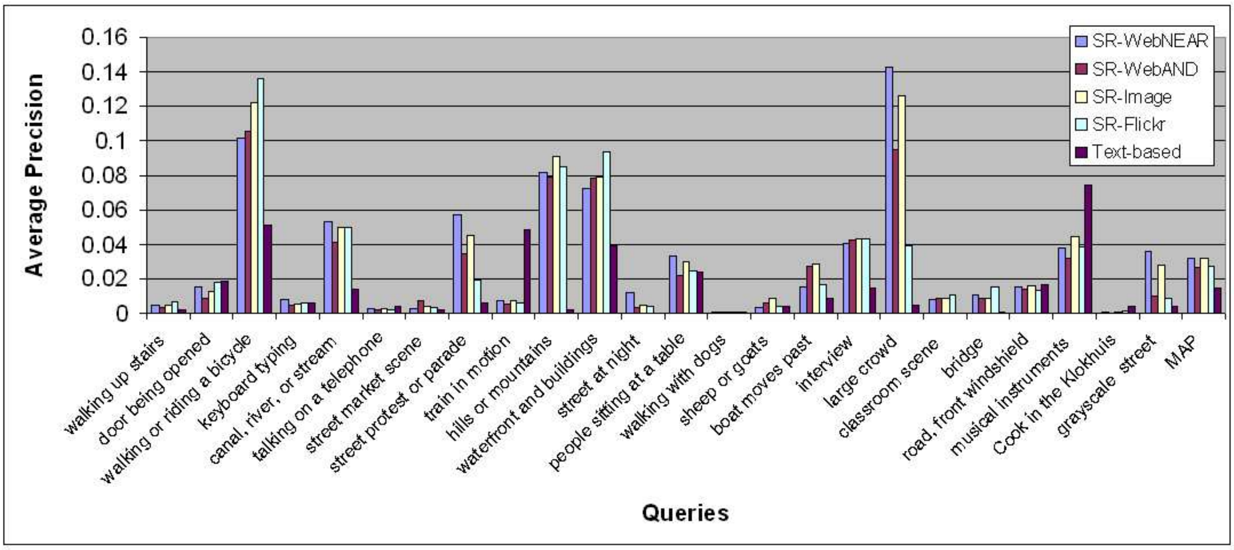


(b)

Figure 5.17: (a) demonstrates average precision-recall curves for different semantic retrieval methods on TRECVID'07 test data. (b) shows AP results for each query (represented by some selected words) using several semantic retrieval methods on TRECVID'07 test data. MAP (Mean Average Precision) is shown at far right. Both of the results are obtained from complete concept set and Vireo detectors.



(a)



(b)

Figure 5.18: (a) demonstrates average precision-recall curves for different semantic retrieval methods on TRECVID'07 test data. (b) shows AP results for each query (represented by some selected words) using several semantic retrieval methods on TRECVID'07 test data. MAP (Mean Average Precision) is shown at far right. Both of the results are obtained from complete concept set and Columbia detectors.

Method Name		Columbia Detectors		Vireo Detectors	
		MAP’06	MAP’07	MAP’06	MAP’07
Visual Content-based Semantic Retrieval	PMI-IR-WebNEAR	3.1%	3.2%	4.5%	4.6%
	PMI-IR-WebAND	1%	2.7 %	0.8%	3.7%
	PMI-IR-WebImage	2.8%	3.2%	3.9%	4.7%
	PMI-IR-FLICKR	2.4%	2.7%	2.9%	3.5%
Text-based Retrieval		1.9%	1.5%	1.9%	1.5%

Table 5.3: Overall MAP(Mean Average Precision) comparison of visual content-based semantic retrieval methods using corpus-based similarity measures on TRECVID’06 and TRECVID’07 data with complete concept set.

First, we tested the methods on TRECVID'06 test set. As it is shown in figure 5.15 and 5.16, the VC-SR had significantly better performance than the TEXT-SR on most of the queries. The MAPs of best VC-SR methods are 3.1% and 4.5% using Columbia detectors and Vireo detectors, respectively. The MAP of TEXT-SR is 1.9%. Overall, we had 63% and 136% performance increase over TEXT-SR method using Columbia and Vireo detectors, respectively. One drawback of VC-SR method is that it doesn't perform well on the named entities such as 'George Bush, Dick Cheney, Saddam Hussein, Condoleezza Rice '. For these queries TEXT-SR has better performance than VC-SR.

In TRECVID'07, the VC-SR had much better performance than the TEXT-SR. The MAPs of best VC-SR methods are 3.2% and 4.7% using Columbia detectors and Vireo detectors, respectively. The MAP of TEXT-SR is 1.5%. Overall, we had 113% and 213% performance increase over TEXT-SR method using Columbia and Vireo detectors, respectively. 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 results of TRECVID'07 evaluations are shown in figure 5.18 and 5.17. The overall MAP comparison for TRECVID'07 and TRECVID'06 data sets using both Columbia and Vireo detectors is shown in table 5.3.

From the retrieving new concept evaluations we know that Vireo detectors are better than Columbia detectors. Thus VC-SR using Vireo detectors performed better than Columbia detectors. 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 [43, 36, 37] 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.

5.4 Conclusion

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

Humans frequently use the context of known concepts in order to learn new concepts. The work in this thesis is motivated by this intuitive observation. We propose two different methods for semantic video retrieval using high level contextual relations between concepts. These relations are automatically extracted from available text (language) resources and hand crafted semantic networks. For both of the proposed methods we have promising results to pursue research on this newly emerging field.

In this proposal we demonstrate that retrieval using high-level context is feasible. And the high-level context can be learned from web and semantic networks. Even though we had significantly good results, our method for modeling context was trivial. More sophisticated methods such as probabilistic graphical models can be investigated for harnessing high-level context learned from web and semantic networks. In this thesis we didn't use spatial information due to the complexity of segmentation and localization tasks in TRECVID data sets. A more comprehensive context model which includes spatial relations can also be developed for better detection and retrieval results.

LIST OF REFERENCES

[1] Eakins,J. Graham,M. Content-based image retrieval. Technical Report, University of Northumbria at Newcastle, 1999.

[2] Chen,Y. Wang,J.Z. Krovetz,R. An unsupervised learning approach to content-based image retrieval. IEEE Proceedings of the International Symposium on Signal Processing and its Applications, 2003

[3] Smeulders,A.W.M. Worring,M. Gupta,A. Jain,R. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000

[4] Kennedy,L. et al. LSCOM Lexicon Definitions and Annotations Version 1.0. DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, 2006.

[5] Smeaton,A.F. et al. Evaluation campaigns and TRECVid. Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, 2006.

[6] Hauptmann,A. et al. Video Retrieval using Speech and Image Information. Electronic Imaging Conference (EI'03), Storage Retrieval for Multimedia Databases, Santa Clara, CA, 2003.

[7] Mihalcea,R. et al. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In Proc. of AAAI'06, 2006.

[8] Lin,D. An Information-theoretic Definition of Similarity. In Proc. of the ICML'98, 1998.

[9] Leacock,C. and Chodorow,M. Combining local context and WordNet sense similarity for word sense identification. In WordNet, An Electronic Lexical Database. The MIT Press, 1998

[10] Lesk,M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the SIGDOC Conference, 1986.

[11] Resnik,P. Using information content to evaluate semantic similarity. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995.

[12] Wu,Z. and Palmer,M. Verb semantics and lexical selection. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1994.

[13] Jiang,J. and Conrath,D. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the International Conference on Research in Computational Linguistics, 1997.

[14] Hirst,G. and St-Onge, D. Lexical chains as representations of context for the detection and correction of malapropisms. In WordNet, An Electronic Lexical Database. The MIT Press, 1998.

[15] Turney,P. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In Proc. of ECML'01, 2001.

[16] Miller,G. et al. WordNet: An electronic lexical database. MIT Press,1998.

[17] Pedersen,T. WordNet::Similarity - Measuring the Relatedness of Concepts. In Proc. of AAAI'04, 2004.

[18] Yahoo Web Search Engine. http://www.yahoo.com.

[19] Google Video Search Engine. http://video.google.com.

[20] YouTube Video Search Engine. http://www.youtube.com.

[21] Yahoo Image Search Engine. http://images.search.yahoo.com.

[22] Flickr Image Search Engine. http://www.flickr.com.

[23] The Offical Website of TRECVID Benchmark. http://www-nlpir.nist.gov/projects/trecvid/.

[24] Yanagawa,A. et al. Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts. Columbia University ADVENT Technical Report #222-2006-8, 2006

[25] Jiang,Y. et al. Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. ACM CIVR'07, 2007.

[26] Shotton,J. Winn,J. Rother,C. Criminisi,A. TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. European Conference on Computer Vision, 2006.

[27] Marszalek,M. and Schmid,C. Semantic Hierarchies for Visual Object Recognition. IEEE Conference on Computer Vision and Pattern Recognition, 2007

[28] Torralba,A. Fergus,R. Freeman,W.T. Tiny Images. Technical Report MIT-CSAIL-TR-2007-024, 2007.

[29] Iyengar,G. Nock,H. and Neti,C. Discriminative model fusion for semantic concept detection and annotation in video. ACM Multimedia, 2003.

[30] Snoek,C.G. et al. The mediamill TRECVID 2004 semantic video search engine. In Proc. of TRECVID, 2004.

[31] Jiang,W. et al. Active Context-based concept fusion with partial user labels. In IEEE International Conference on Image Processing (ICIP'06), 2006.

[32] Naphade,M.R. et al. Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems. In Proc. of ICIP, 1998.

[33] Naphade,M.R. and Smith,J.R. A Hybrid Framework for Detecting the Semantics of Concepts and Context. ACM CIVR'03, 2003.

[34] Yan,R. et al. Mining Relationship between Video Concepts Using Probabilistic Graphical Model. IEEE International Conference on Multimedia and Expo (ICME'06), 2006.

[35] Aytar,Y. et al. Improving Semantic Concept Detection and Retrieval using Contextual Estimates. In Proc. of ICME'07, 2007.

[36] Snoek,C.G. et al. Adding Semantics to Detectors for Video Retrieval. IEEE Transactions on Multimedia, 2007.

[37] Snoek,C.G. et al. Are Concept Detector Lexicons Effective for Video Search?. In Proc. of ICME'07, 2007.

[38] D. Wang et al. The Importance of Query-Concept-Mapping for Automatic Video Retrieval. In Proc. of ACM Multimedia, 2007.

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

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

[41] Rubner,Y. et al. A Metric for Distributions with Applications to Image Databases. In Proc. of ICCV'98, 1998.

[42] Lynch,P. et al. An Evaluation of New and Old Similarity Ranking Algorithms. In Proc. of ITCC'04, 2004.

[43] Hauptmann,A. et al. Can High-Level Concepts Fill the Semantic Gap in Video Retrieval? A Case Study with Broadcast News. IEEE Transactions on Multimedia, 2007.

[44] Lew,M.S. Sebe,N. and Huang,T. et al. Content-Based Multimedia Information Retrieval: State of the Art and Challenges. ACM Transactions on Multimedia Computing, 2006.

[45] Torralba,A. et al. Contextual Models for Object Detection using Boosted Random Fields. In NIPS'05, 2005.

[46] Torralba,A. et al. Context-based Vision System for Place and Object Recognition. In Proc. of ICCV'03, 2003.

[47] Amores,J. Sebe,N. et al. Context-Based Object-Class Recognition and Retrieval by Generalized Correlograms. IEEE TPAMI, 2007.

[48] Barnard,K. Duygulu,P. and Forsyth,D. et al. Matching Words and Pictures. Journal of Machine Learning Research, 2003.

[49] Sivic,J. and Zisserman,A. Video Google: A Text Retrieval Approach to Object Matching in Videos. In Proc. of ICCV'03, 2003.

[50] Liu,Y., Kender,J.R. Fast Video Segment Retrieval by Sort-Merge Feature Selection, Boundary Refinement, and Lazy Evaluation. Computer Vision and Image Understanding, 2003.

[51] Ponceleon,D Syeda-Mahmood,T. et al. CueVideo: automated multimedia indexing and retrieval. In Proc. of ACM Multimedia, 1999.

[52] Griffin,G., Holub,AD. and Perona,P. The Caltech-256, Caltech Technical Report, 2007.

[53] Everingham,M. Van Gool,L. Williams,C.K.I. Winn,J. and Zisserman,A. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, 2007.

[54] Russell,B.C. Torralba,A. Murphy,K.P. and Freeman,W.T. LabelMe: a database and web-based tool for image annotation. MIT AI Lab Memo AIM-2005-025, September, 2005.