Honours Year Project Report

Refining and Expanding WordNet for Video Retrieval

By

Zhao Jin

Department of Computer Science

School of Computing

National University of Singapore

2005/2006

Honours Year Project Report

Refining and Expanding WordNet for Video Retrieval

By

Zhao Jin

Department of Computer Science

School of Computing

National University of Singapore

2005/2006

Project No: H079150
Advisor: A/P Kan Min-Yen

Deliverable:
Report: 1 Volume

Abstract

Recent research in video retrieval has shown that correctly chosen high-level feature detectors are able to improve retrieval performance. This has motivated us to design and implement a system to deduce the correct high-level feature for a query with WordNet (reported in our paper submitted to CIVR 06); however, through our experience in building this system, we realize that the WordNet is still limited in terms of both the synsets and the relations it contains. Therefore, in this project, we refine and expand the WordNet using a two-phrase algorithm with a number of resources. Evaluations on this version of WordNet show that 1) it contains more useful synsets and relations 2) it can be used to perform high-level feature deduction better agreed by human judges and 3) it yields good improvement on the retrieval performance when integrated with the NUS PRIS video retrieval system.

Subject Descriptors:

H.3.3 Information Storage and Retrieval

I.2.4 Knowledge Representation Formalism

I.2.6 Learning

I.2.7 Natural Language Processing

Keywords:

Query Analysis, Visual Query

Implementation Software and Hardware:

Intel Pentium 4 1.7GHz, MS-Windows XP, Java JDK1.4.2, WordNet

Acknowledgement

I would like to thank my supervisor, Dr.Kan Min-Yen for his guidance throughout the year. He has been very patient, supportive and approachable whenever I had any difficulty in my research. Without him, the research would not have been so enjoyable.

I would also like to thank my research partner, Mr. Neo Shiyong for his assistance in running some of the evaluations and his valuable advice on my research.

Last but not least, I would like to thank my family for all the encouragement and support in all these years.

List of Figures and Tables

Table of contents

Chapter 1 Introduction

Traditionally, video retrieval has been largely based on textual description, ASR (Automatic Speech Recognition) and CC (Close-Caption). These sources of information have been proved useful especially for retrieving news videos with specific entities since the information of such entities are usually directly available in the sources; however, one major shortcoming of all of them is that they are unable to describe all the semantic concepts in the video. This makes it very difficult to retrieval videos with no specific entities because general entities, despite the fact that they appear frequently in video, are usually not directly available in all those sources.

As a solution to this problem, detectors for semantic concepts (high-level features) come into the picture. In the year 2002, TREC video track (TRECVID, 2002) included high-level feature extraction as one of its three main tasks. Since then more and more research effort has been made to build detectors for high-level features. One of such effort is the LSCOM annotation (LSCOM), a research effort to build detectors for around 200 high-level features. In 2005, (Christel and Hauptmann, 2005) proved that if the high-level features are chosen correctly for a text-based video query, they generally help in improving the retrieval performance of a video search.

A natural question that follows is how we can find the relevant high level features for a query. Our solution to this is to use WordNet and WordNet-Similarity to compute the similarity between the terms in the query and the description of a feature. The more similar they are, the more relevant we consider the feature is. This approach has been implemented and shown to be useful in our previous work (Neo, Zhao, Kan, and Chua 2006) submitted to CIVR 06. However, we have also realized through this research that WordNet has a number of limitations, i.e. insufficient synsets, insufficient relations, hidden information and static source of information. All of them are the factors that limit the efficacy of WordNet in

high-level deduction. This motivates us to refine and expand the WordNet to make it more suitable for high-level feature deduction in video retrieval.

The approach we take is a two-phrase, semi-automatic process with minimal human intervention. The reasons why we have it are briefly explained as follows:

1. Two important findings from our previous research are 1) indirectly related high-level features are harder to deduce 2) the gloss of the synsets in WordNet is also a good source of information we should not ignore. Therefore, in the first phrase, we refine the WordNet by 1) classifying the synsets to make the distinction between the synsets which are generally indirectly related to a query and the synsets which are generally directly related 2) building appropriate relations between the former and the latter based on the information from the gloss and other resources.

2. On the other hand, we have also realized through our previous research that lots of information about real-life happenings is missing in the WordNet. For example, there is no synset for “Hu Jintao”, who appears very often in the news video nowadays. Therefore, in the second phrase, we expand the WordNet to introduce addition synsets and relations based on the information from a news corpus.

There are two major advantages of this approach: the first advantage is that in this way, all the useful information is concentrated in the synsets and relations, which make it much easier to perform high-level feature deduction. The second advantage is that it alleviates the need to perform query expansion using other sources of information because much of such information would have been encoded into the WordNet through the expansion.

The rest of report is organized in this way:
Chapter 2 gives a review on the related works in the same area. Chapter 3 examines the limitations of the current WordNet. Chapter 4 describes the

methodology we used for refining and expanding the current WordNet in detail. Chapter 5 focuses on the evaluations on the expanded WordNet with the methodologies explained and the results listed and discussed. As the closing chapter for this report, Chapter 6 gives an account of the conclusion we have reached, followed by some possible directions for future work.

Chapter 2 Related Work

This chapter presents a summary of the related works and it is divided into four sections. The first section is concerned with query processing techniques in the current video retrieval systems, followed by the second section which is on the previous research effort on ontology building specifically for video retrieval. The third section gives a review on the different attempts to extend the WordNet while the resources we used in our project are introduced in the last section.

2.1 Query processing in the current video retrieval systems

Query processing has been one of the most important components in the current video retrieval systems. A number of systems like (Chua et al, 2004) (Kennedy, Natsev, and Chang, 2005) use a variety of resources like WordNet, Google News, OKAPI, etc. to perform query expansion; however, there is very little work on processing the queries specifically for high-level feature deduction. One of such paper is (Snoek et al, 2005) in which they map the concepts of the features into the synsets of WordNet manually. After the mapping, they calculate the similarity score between the concepts from the processed queries and the ones from the features and pick the one with the highest similarity score as the high-level feature to be queried on. Our work is different from all of the above in the sense that 1) we target at high-level feature deduction specifically and 2) we are doing it in a more automated and reusable way, i.e. using a semi-automated methodology to refine and expand a current ontology that can be reused by all the systems.

2.2 Ontology building in the context of video retrieval

The previous efforts of ontology building in the context of video retrieval have been focusing on the video annotation. (Hollink and Worring 2005) proposes the

four major requirement of a visual ontology for video annotation: Visuality, Generality, General-Visual Relations and Interoperability. Apart from this research on the requirement of a visual ontology, there are also other works (Tsinaraki, Polydoros, Kazasis and Christodoulakis, 2005) (Tsinaraki, Polydoros, Moumoutzis and Christodoulakis) which aim to build a visual ontology by integrating the existing framework like MPEG-7 and TV-anytime; however their work emphasizes on the specification of the ontology without explicitly mentioning a way of building it.

2.3 Previous work on WordNet expansion

Ever since its birth in the 1990's, WordNet (Miller G. 1995) has been widely used as a lexical database for all kinds of domains in information retrieval; however, due to the fact that WordNet is a general-purpose ontology, it is not possible for it to fit all kinds of domains equally well. Therefore, the research of how to enrich the WordNet has been an ongoing research area in information retrieval.

In the context of text retrieval, (Bentivogli and Pianta, 2003) enriches the WordNet with the idea of phrase sets; (Vossen and Piek, 2001) extends and trims the WordNet for cross-lingual retrieval and multilingual ontology building; (Mihalcea and Moldovan, 2001) works on enhancing WordNet both semantically and logically and (Gangemi, Navigli and Velardi 2002) enhances WordNet with domain concepts.

In the context of video retrieval, besides proposing the requirements for a visual ontology, (Hollink et al, 2005) also demonstrates a design of such ontology based on WordNet and MPEG-7. Their work is built upon the work by (Hoogs, Rittscher, Stein and Schmiederer, 2003) (Stein, Rittscher and Hoogs, 2003), which is also an extension to the WordNet for video retrieval. Since we aim to perform high-level feature deduction instead of semantic annotation, i.e. low

level features like visibility, colors, material are not the main concerns in our version of WordNet, our refinement and expansion to WordNet is more compact and informative than theirs.

2.4 Resources used in our project

Harvard IV-4 Dictionary: The Harvard IV-4 dictionary (Harvard) is a dictionary that list words according to their categories. For example, in the dictionary, “congress” is listed together with “democratic” and other terms about politics in the category “politics”. We use this dictionary to decide which category a concept belongs to.

AQUAINT Corpus: The AQUAINT corpus is a large text-corpus drawn from three news resources: the Xinhua News Service (People's Republic of China), the New York Times News Service, and the Associated Press Worldstream News Service. It is generally used in official benchmark evaluations conducted by National Institute of Standards and Technology (NIST) for text retrieval. This corpus is helpful in the sense that it provides information of the real life which WordNet is lacking of.

LSCOM dataset: The LSCOM dataset is a comprehensive list of approximate 200 concepts that can be used for video annotation. This list gives us a very good idea of what kinds of high-level features are being annotated as well as the type of links we should incorporate into the WordNet so that such high-level features can be deduced accurately.

Chapter 3 Limitations of the current WordNet

In this chapter, we examine the problems in the current WordNet in the context of video retrieval. Section 3.1 briefly introduces the limitations of the current WordNet based on our previous experience on using it for high-level feature deduction. Details of these limitations are then examined in depth in the following subsections.

3.1 Limitations in the current WordNet

Through our previous experience of using the WordNet for high-level feature deduction, we have discovered that the WordNet is limited in the following four ways:

1. The type of semantic relations encoded among the concepts is far from enough for high-level feature deduction;
2. Some concepts are not represented in WordNet;
3. Some information is hidden in WordNet;
4. The source of information of WordNet is rather static than dynamic.

3.1.1 Insufficient relations

In the current version of WordNet, the types of semantic relations encoded include: *antonym, hypernym/hyponym, holonym/meronym* and *domain (topic, region, usage)*. Among all these relations, the *hypernym/hyponym* relation is the most widely-used and most helpful for high-level feature deduction because it is the backbone of the noun hierarchy which allows us to know what concept is a kind of another concept. For example, when we want to search for shots about "basketball players on the court", we can limit our search to the shots about "sports" because we can discover through the *hypernym/hyponym* relations that "basketball" is a kind of "sports".

However, this is almost all the information we can get from the current WordNet: For high-level feature deduction, *antonym, holonym/meronym* are not particularly useful because *antonym* is exactly opposite to what we are looking for while *holonym/meronym* tends to give information that deviates from our original focus. For the three types of domain pointers: *topic* is useful but it is far from complete in the current WordNet, *region* mentions the location of a concept but the location indicated is usually a country which is too broad to be helpful while *usage* concerns only on grammar hence it is not related to high-level feature deduction at all.

If we try to classify the type of high-level features listed in the LSCOM dataset, we can see that there are 5 major types of them: *topic* (ex. politics, finances), *location/environment* (ex. office, sky), *subject/object/action* (ex. firefighters, plane, laughing). Therefore, in order to make an ontology useful for high-level feature deduction for video retrieval, we suggest that the following types of relations should be encoded:

1. *Topic*. Whether concept *A* is in the domain defined by concept *B*. Example: “U.S. President” is a topic in “Politics”.
2. *Location/environment*: Whether concept *A* usually appears/takes place in the location/environment defined by concept *B*. Example: “plane” usually appears in the “sky” while “swimming” usually takes place in “water”
3. *Subject/object/action*: Whether concept *A* is a kind of concept *B* or they usually appears together. Example: “fork” is a kind of “utensil” while “fork” and “spoon” usually appears together.

In addition, given the fact that video is an audiovisual medium, we propose encoding this relation as well.

4. *Sound*: Whether concept *A* makes the sound defined by concept *B*. Example: “guitar” makes the sound “strum”.

One last relation to be encoded is the general relation *related*.

5. *Related*: Whether concept *A* is related to concept *B* with a relation that has not been classified specifically. Example: “U.S. president” and “George Bush”.

3.1.2 Missing concepts:

Although WordNet contains around 120000 concepts in its database, there are still a number of concepts which are missing. This results in empty or incorrect results when we try to obtain information from the WordNet about such concepts. There are two major types of such concepts: video terminologies and proper nouns.

Video terminology is the first type of missing concepts in WordNet. For example: the user may want to see a video clip which is a “closeup” of a particular person. Instead, the concept of “closeup” defined in WordNet is in fact “a photograph taken at close range”, which is not really what the user meant when he/she specifies the query.

Proper noun concept is another type of missing concepts in WordNet. Although WordNet does contain a number of proper noun concepts, those are not really sufficient. For example, the concept “Hu Jintao” is not one of the proper nouns in the WordNet; however, given the fact that he appears very frequently in the videos, he should be included in the ontology as well.

3.1.3 Hidden information:

A lot of useful information in the WordNet is in fact hidden in the gloss of the synset. Take the synset “boat” as an example, the gloss of this synset is “a small vessel for travel on water”. This provides good information on the “location/environment” of the object “boat”. However, this piece of information is hidden in the gloss. In other words, in order to make use of such information, we would have to dig it out from all other less related terms every time we try to perform high-level feature deduction. Therefore, it would be very helpful if we

can extract such information and encode it explicitly into relations so that we can save the time and trouble of digging such information every time.

3.1.4 Static source of information

The source of information for WordNet is relatively static in the sense that there are very few changes or updates between editions of WordNet. This is also one important limitation of WordNet. The reason is that lots of things happen everyday in the real life with new concepts and new relations between the concepts emerging daily. If the source of information for a visual ontology is static, very soon it would not be helpful to answer the queries with the new concepts. Therefore, besides a static source of information as a base, a dynamic source of information should be included as well to keep the information updated.

Chapter 4 Refining and Expanding the WordNet

This chapter gives an overview of the structure of the refined and expanded WordNet in Section 4.1 with the detailed methodology we used to create this version of WordNet in Section 4.2.

4.1 Structure of the refined and expanded WordNet

In our version of WordNet, synsets are still classified based on their part-of-speech. In addition, the noun class is further classified into five subcategories: *topic, subject/object/action, location/environment, sound and camera position.* Note that these five subcategories are not mutually exclusive, i.e. one noun synset can be in multiple subcategories at the same time.

In terms of the relations encoded, all the original relations are preserved in the refined and expanded WordNet, with the following type of pointers added/enhanced: *topic, location/environment, sound, related.*

The structure of the refined and expanded WordNet is described in the following diagram and more details on the synset categories and relations are explained in the next two subsections.

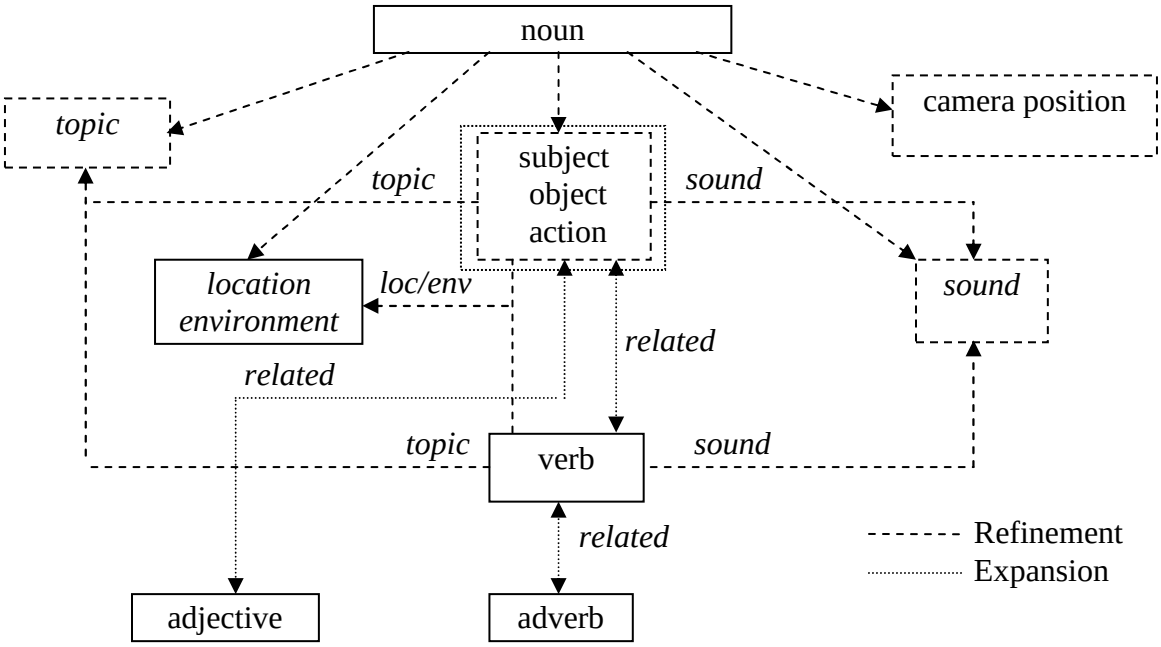

Fig 4.1 Structure of the refined and expanded WordNet

4.1.1 Synset categories in the refined and expanded WordNet

Altogether there are five sub-categories for the noun synsets:

- *Topic*: A synset in this category represents a concept that can be a topic of a video clip. Examples: “politics”, “sports”.
- *Subject/object/action*: A synset in this category represents a concept which is a subject or object or action. Examples: “firefigher”, “vehicle”, “swimming”
- *Location/environment*: A synset in this category represents a concept which can be a location or environment where certain things exist or certain actions take place. Examples: “sky” (for plane), “water” (for swimming).
- *Sound*: A synset in this category represents a concept which is a sound that can be produced by certain things or actions. Examples: “strum” (by guitar or by strumming)
- *Camera position*: A synset in this category represents a concept which is a description of a camera position relative to the video. Example: “closeup”

4.1.2 Relations in the refined and expanded WordNet

Besides the relations in the original WordNet, the following seven types of relations have been added or enhanced in the refined and expanded WordNet:

- *Domain of synset (topic)*: If synset *A* is in the domain (topic) defined by synset *B*, the relation *domain of synset (topic)* exists from *A* to *B*. This relation is inheritable.
- *Synset of domain (topic)*: If synset *A* is the domain (topic) of synset *B*, the relation *synset of domain (topic)* exists from *A* to *B*.
- *Location/environment of synset*: If synset *A* usually appears in or is executed in the location/environment defined synset *B*, the relation *location/environment of synset* exists from *A* to *B*. This relation is inheritable.
- *Synset of location/environment*: If synset *A* is the location/environment in which the concept defined by synset *B* exists or takes place, the relation *synset of location/environment* exists from *A* to *B*.
- *Sound of synset*: If synset *A* produces the sound defined by synset *B*, the relation *sound of synset* exists from *A* to *B*. This relation is inheritable.
- *Synset of sound*: If synset *A* is the sound produced by synset *B*, the relation *synset of sound* exists from *A* to *B*.
- *Related Synset*: If synset *A* is related to synset *B* but the type of relation is not clear, the relation *related synset* exists from *A* to *B* and from *B* to *A*.

4.2 Methodology for WordNet refinement and expansion

Our methodology for WordNet refinement and expansion is divided into two phrases:

- **Phrase I WordNet and dictionary-based refinement** is a semi-automated process in which we classify the synsets into the first four subcategories (all except the camera position subcategory), add in the missing synsets for the subcategory camera postionand build first six types of the relations (all except the *related synset* relation). The resources used are WordNet, Harvard IV dictionary and a list of video terminology.

- **Phrase II Corpus-based expansion** is a fully-automatic process in which we introduce more synsets for proper nouns and build the *related synset* relation with a dynamic source of information. The resource used in this phrase is the AQUAINT Corpus.

The detailed steps described and explained in the two subsections that follow.

4.2.1 Phrase I WordNet and dictionary-based refinement

In phrase I, the synsets are first classified into these four subcategories: *topic*, *subject/object/action*, *location/environment* and *sound*. Due to the fact that it is not practical to exhaustively list down every synset for each subcategory, we make use of the information in the original WordNet and the Harvard IV-4 dictionary to do a semi-automated classification. The actual method of classification and the reason why we do it in this way are as follows:

- ***Topic*:** Any synset that contains a *synset of domain (topic)* pointer is considered to be in this category.

  Reason: Although the synset "*topic*" exists in WordNet, there is no links to the synsets that can be a topic from there. Therefore, we use the *synset of domain (topic)* pointer as an indication of whether a synset can be a topic.

- ***Subject/object/action*:** Any synset which is a *hyponym* of the following synsets is considered to be in this category: {act, action, activity}, {animal, fauna}, {artifact},{body, corpus}, {food}, {natural object}, {person, human being}, {plant, flora}.

  Reason: These 8 unique noun beginners encompass all types of *subject/object/action* and hence we classify all its *hyponyms* to be in this category.

- ***Location/environment*:** In the Harvard IV-4 dictionary, there is a category for terms that are references to places, locations and routes between them. We map the terms which refer to actual location/environment from that category into the synsets in the WordNet and classify those synsets into *Location/environment* subcategory.

  Reason: There is in fact a unique beginner for "location, place" in WordNet; however, unlike the ones for the *subject/object/action* category, the *hyponyms* of this synset include too many synsets that are referring to an abstract location or environment while a lot of synsets which are indeed a physical location or environment missing from there. For example, the synset "office" is listed under "artifacts" but not "location, place". On the contrary, the terms listed in the Harvard IV-4 dictionary are more organized and comprehensive; therefore we decide to use the Harvard IV-4 dictionary as the source of information for this classification.

- ***Sound*:** Any synset which is a *hyponym* of the following synsets is considered to be in this category: "sound" -- sense 1,2 and 4.

  Reason: There are altogether eight senses for the noun "sound". Based on the definition we have for the subcategory *sound*, we select three senses (auditory effect, sensation of hearing and occurrence of audible event) out of the eight to be the roots of this subcategory.

As a second step, we add in the synsets for camera position which are missing in the original WordNet. For this subcategory, we use a list of terminologies for camera position as guidance for the addition.

- ***Camera position*:** We create one synset for each of the terms in the list of video terminologies for camera position.
  Reason: Since there is no information about camera position in the WordNet, the only way is to make use of the external information such as the list of video terminologies for camera position.

In the end, we combine the information from the classification and the gloss of WordNet to build relations among the synsets. This step can be further decomposed four smaller steps:

1. **We find out the adjective and adverb synsets that pertains to the noun synsets in each subcategory by following the *pertainym* pointer between adjective and noun and the derived pointer between adverb and adjective.** For example, we can get the adjective synset "political" and the adverb synset "politically" for the noun synset "politics" because "political" pertains to "politics" while "politically" is derived from "political". Together with the noun synset itself, we have a complete set of indicator synsets for any given noun synset in each subcategory.

2. **We then take out all the lemma from each indicator synsets.** Following the previous example, we can get the words "politics", "political" and "politically" as the indicators for the topic synset "politics".

3. **The third step for the relation building is to tag the gloss with a part-of-speech tagger (Infogistics, 2000) and filter out the stopwords with a stopword list of size 420 (Lextex, 2000).** This step is to prepare the

source of information, which is the gloss, for the matching and actual relation building in step 3.

4. **As the last step for the relation building, we scan through the remaining terms in the gloss. If in the gloss of a synset *A*, there is a term which is an indicator for a topic synset *B*, we consider the relation *Domain of synset (topic)* exists from *A* to *B* and the relation *Synset of domain (topic)* exists from *B* to *A*.** One such example is the synset for "hustings" whose gloss contains the indicator "political" (the activities involved in political campaigning, especially speech making) for the topic "politics". **We repeat this step for all three subcategories (*topic, location/enviroment* and *sound*) and build the six types of relations we need in phrase I.**

At the end of phrase I, we have partially overcome some of the limitations of WordNet: i.e. insufficient relations and missing concepts. In order to fully overcome all the limitations, we carry on with phrase II, corpus-based expansion.

4.2.2 Phrase II Corpus-based expansion

The source of information we use in phrase II is the AQUAINT corpus. This corpus is relatively noisier than the gloss in the WordNet because it is made up of millions of articles written by people from different background. In the light of this, we use a restricted version of mutual information together with TFIDF to analyze the corpus.

A restricted version of mutual information

A basic version of mutual information only takes into account the number of co-occurrence of two concepts in a sentence. We think this method is not strong enough because even if two terms co-occur within the same sentence, they may not be related at all. One of such example is that "ex-

president" and "Monday" co-occur in the sentence "Indonesia's ex-President Suharto turns 77 on Monday"; however, there is no relation between them at all. Therefore, in order to make the mutual information more accurate, we restrict the type of mutual information we want to look at in the corpus.

The restricted version of mutual information takes into account one more piece of information--sentence structure. Most of the English sentences contain three components: *subject*, *verb* and *object*, together with some modifiers for each component. Our idea is that the occurrence of two terms are counted only when 1) they come from the same component or modifier or 2) the components or modifiers they come from is one of the following pairs:

- *Subject* & *subject modifier*
- *Verb* & *verb modifier*
- *Object* & *object modifier*
- *Subject* & *verb*
- *Verb* & *object*

Under this restricted version of mutual information, the co-occurrences counted are more accurate. Let's take a look in the previous example, the co-occurrence of "ex-President" and "Monday" is no longer counted because "ex-President" comes from the *subject modifier* while "Monday" comes from the *verb modifier*; however, the co-occurrence of "ex-President" and "Suharto" is preserved as one comes from the *subject modifier* and the other comes from the *subject*.

In order to apply this version of mutual information, we tag the AQUAINT corpus with a Part-of-Speech tagger which gives not only the Part-of-Speech tags but also the phrase and clause tags as well. With all the tags, we analyze the sentence

structure, extract the key terms in each component and modifier and record the number of co-occurrences for each pair of terms.

After obtaining the pairs of terms and their number of co-occurrences, we classify pairs into three categories: 1) pairs of proper noun 2) pairs of normal word and 3) a mix of proper noun and normal word. The reason why we make this distinction is that proper nouns are much less common than normal words; as a result, the counts for the pairs with at least one proper noun are also much smaller than the ones with at lease one normal word; therefore, by classifying them into three category, we can in fact handle the counts in a relative way instead of an absolute way, which is more fair to the pairs with proper nouns.

These pairs are then further processed by modulating the counts using TFIDF, in which the counts of the pairs are discounted by the number of times each term occurs in any pairs of terms. The pairs with low counts are then filtered and the remaining ones are considered to be related, although how they are related is not analyzed. This concludes the analysis part for phrase II.

The actual synset addition and relation building start right after the analysis. For all the pairs of the terms, if any of the term is not in the WordNet yet, a synset is created to represent this missing term (which is usually a proper noun). A *related synset* pointer is then created between the first synsets of both terms. (Due to time constraint, no sense disambiguation is done here. We assume that the first synset, which is the most common meaning of a term, is correct.)

Phrase II ends when all the synsets are added and relations built. This not only overcomes the limitation of insufficient relations and missing concept, it also partially adds the dynamic element for the source of information in WordNet since the AQUAINT corpus are made up of news articles which describe the things that happen every day in the world. In order to fully overcome this limitation, we suggest that phrase II is repeated periodically with some source of

information similar to the AQUAINT corpus. In this way, the information of the ontology is always up-to-date to cater for the ever-changing need of the user in the search of videos.

Chapter 5 Evaluation

This chapter explains the evaluations we have done on the refined and expanded WordNet. An overview of all the evaluations is given in the first section. The details of the methodologies for the three evaluations are explained in the following three sections with the result listed and discussed.

5.1 Overview of evaluation

In order to fully evaluate the efficacy of the extended WordNet, we have designed three sets of evaluations based on different methodologies:

In the first evaluation, we take a look at what synsets and relations have been added by examining the content of the refined and expanded WordNet. This is to directly see how much more information has been included into the WordNet.

The second evaluation takes the idea of the human judges into the consideration. We reuse the query and high-level feature relevance ranking we obtained from the survey we conducted for the CIVR paper. This examines how well the high-level feature deduction with the refined and expanded WordNet matches with the thinking of the human judges.

Lastly, we have a retrieval-performance-based evaluation, which is concerned with whether a video retrieval engine can benefit from using the refined and expanded WordNet. We integrate the high-level feature deduction with different versions of WordNet into the NUS PRIS video retrieval system to obtain the Mean Average Precision (MAP) of the retrieval result and evaluate the efficacy from there.

For all three evaluations, we make the comparison among the original WordNet, the refined WordNet after phrase I and the refined and expanded WordNet so that we can clearly see the differences among them. We will refer to these three versions of WordNet as WNa, WNb and WNc respectively throughout this chapter.

5.2 Relation-based evaluation

For the relation-based evaluation, we have picked a few terms for a close examination based on 1) whether their relations have been enriched through the expansion or 2) whether they are missing in the original WordNet but are introduced in the expanded version. For each of the terms, we list out its relations with other terms which are introduced through the refinement and expansion.

The reason why we have this evaluation is that query expansion, an important component of query processing, relies heavily on the existing synsets and relations in the WordNet. Therefore, we devise this evaluation to show what kinds of synsets and relations have been introduced in the refined and expanded WordNet. Here are some examples of the synsets and relations introduced:

Term	**WNb**			**WNc**
	Top	**Loc**	**Snd**	**Related**
president	politics	nil	nil	Bill Clinton
basketball	sports	nil	nil	World Championship
twilight	nil	sky	nil	nil
Control tower	nil	airport	nil	nil
explosion (Action)	nil	nil	explosion (sound)	shooting
guitar	art	nil	sound	nil
Hu Jintao	nil	nil	nil	Communist Party, Beijing
Condoleeza Rice	nil	nil	nil	adviser National Security Council

Table 5.1 Sample synset and relations in different version of WordNet

As we can see from the table, the synsets and relations missing in WNa are added in WNb and WNc. For example, the relation between “president” and “politics”, which is not present in WNa is introduced in WNb while the synset Hu Jintao,

which is absent in WNa, is introduced in WNc and linked to “Communist Party” and “Beijing”.

In order to prove the efficacy of the WordNet, it is not enough just to show that how many synsets and relations are added. Therefore, we have also carried out two more evaluations: one to see whether the refinement and expansion increase the agreement of relevance ranking between the high-level feature deduction system and the human judges; another one to see whether they actually improve the retrieval performance.

5.3 Human-judges-based evaluation

In our previous research in the same area, we have conducted a survey in which 12 participants were asked to rank a set of high-level features according to their relevance for each of the 8 search tasks selected from TRECVID 2005.

The feature set we used is made up of the following 24 features:

- Topic: *anchorPerson, commercial, politics, sports, weather, financial*
- Subject/object/action: *face, fire, explosion, car, U.S. f lag, boat, aircraft, map, buildingExterior, prisoner; peopleWalking; peopleInCrowd*
- Location/environment: *waterscape; mountain; sky; outdoor; indoor; disaster; vegetation*

The search tasks we used include:

1. Find shots of a graphic map of Iraq, location of Baghdad marked - not a weather map
2. Find shots of one or more palm trees
3. Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people
4. Find shots of basketball players on the court
5. Find shots of a ship or boat
6. Find shots of Hu Jintao, president of the People's Republic of China
7. Find shots of George Bush entering or leaving a vehicle (e.g., car, van, airplane, helicopter, etc) (he and vehicle both visible at the same time)
8. Find shots of something (e.g., vehicle, aircraft, building, etc) on fire with flames and smoke visible

For each of the search tasks, the participant is asked to rank each of the features on a 5-point scale with 1 being very irrelevant and 5 being very relevant.

We have also implemented a WordNet-based high-level feature deduction system to produce relevance ranking for any set of features and search tasks. The algorithm of the system can be briefly describes as follows:

First of all, both the search task and the feature descriptions are tagged using a Part-of-Speech tagger and the stopwords in them are filtered using a stopword list. Then we expand the remaining terms in both sets by introducing the related terms in WordNet. The third step is to assign a weight to each term based on its part of speech and the relation it has with the original term. As the last step, we calculate the relatedness score between a search task and a feature with this formula:

$$Sim\_Lex(Q_j, HLF_k) = (\sum_{t_q \epsilon Q_j} \sum_{t_f \epsilon HLF_k} Resnik(t_q, t_f))/(|Q_j| * |HLF_k|))$$

where Resnik(t_q,t_f) is the Resnik measure from WordNet-similarity (Resnik 1995)
The final result is further normalized onto a 5-point scale so that it is comparable to the ranking from the human judges. (For more details of the survey, please refer to our paper in Appendix)

For this human-judge-based evaluation, we reuse the data collected from this survey and compare them with the output of the same system with different versions of WordNet.

The degree of agreement between the system and the human-judges measured by weighted Kappa coefficient is shown in the following table:

	WNa	**WNb**	**WNc**
1	0.077	0.252	0.228
2	0.200	0.219	0.293
3	0.002	0.002	0.003
4	0.152	0.08	0.068
5	0.092	0.111	0.111
6	0.081	0.086	0.086

7	0.315	0.344	0.338
8	0.258	0.236	0.207
Average	0.147	0.166 (+11.2%)	0.171 (+16.3%)

Table 5.2 Agreement between the human judges and the high-level feature deduction system with different versions of WordNet

As we can see from the table, the average degree of agreement measured by the Kappa coefficient increases in the order of WNa, WNb and then WNc. This result confirms that the expanded WordNet helps the query processing in the sense that it gives results that are better agreed by human judges; however, we can also see that the difference in agreement can be quite significant from query to query and the agreement values do not strictly followed the order in some of the queries. After studying the difference between the results from the system and the human judges, we have found out a few reasons for this:

1. When there is no directly co-related high level feature, the agreement value is lower. For example, for query 2, the feature “palm tree” is not part of the feature set in our system; therefore on average the agreement value is lower than all other search tasks.

2. When the indirectly co-related high level feature is totally disconnected from the key terms in the query, the agreement value is lower, too. If we compare the result from query 2 and query 3, although there is no direct co-related high-level feature for both of them, the fact that “indoor” is totally disconnected from “office” in any version of the WordNet makes the agreement value in query 3 even lower.

3. The fact that no sense disambiguation is done has also lowered the agreement value because any error in the meaning of the key terms, once occurred, are general passed along and amplified through the process of expansion. One such example is query 4, the term “court” is mentioned as a

venue for sports while its default meaning for it is the venue for legal issues. Therefore the agreement values in WNb and WNc are lower than the previous version because the terms introduced are not related to sports at all.

On seeing that on average the relevance rankings are better agreed by the human judges, we proceed to the third evaluation -- retrieval performance-based evaluation to see whether the expanded WordNet helps in actual video retrieval.

5.4 Retrieval-performance-based evaluation

As the last evaluation of our research, we look into whether the refined and expanded WordNet helps to improve the retrieval efficiency of a video retrieval system (NUS PRIS).

The methodology we use for this evaluation is that we factor the relevance ranking of the high level features into the score for each shot. For each of the shots in the database, if it contains the high level feature which has a high relevance ranking as decided by the system, it will receive a higher bonus mark compared to the one that contains no or a high level feature which has a lower relevance ranking. At the end of the ranking, all the shots with a score greater than a threshold are reported as the relevant shots for the given search task. The accuracy of this set of shots is then measured with the ground truth from TRECVID 2005. For this evaluation, we calculate the MAP over all the search tasks in TRECVID 2005. The result is listed in the following table:

Technique used:	**MAP**
Basedline: Heuristic weighting	0.104
Run1. Automated HLF query matching with WNa	0.110 (+5.7%)
Run2. Automated HLF query matching with WNb	0.113 (+8.6%)
Run3. Automated HLF query matching with WNc	0.115 (+10.1%)

Table 5.3 MAP of the baseline system with high-level feature deduction based on different versions of WordNet

Although we only make use of WordNet to perform the query processing for this part of evaluation, we can already see from the result that the refined and expanded WordNet is able to improve the MAP. In order to make these results comparable to the ones we obtained in our previous research, we have also put the query processing

result from WNb and WNc into our integrated system which includes temporal mutual information and confidence rating for the detectors. The results are further improved as listed in the following table:

Technique used:	MAP
Baseline: Automated HLF query matching with WNa + Temporal MI	0.113
Run4. Automated HLF query matching with WNb + Temporal MI	0.118 (+4.4%)
Run5. Automated HLF query matching with WNc + Temporal MI	0.120 (+6.2%)

Table 5.4 MAP of the integrated system with high-level feature deduction based on different versions of WordNet

Based on figures in both tables, we can conclude that the retrieval performances has benefited from the refined and expanded WordNet. In addition to discovering the directly related high-level features, which can be done with the original WordNet, the expanded WordNet is capable of discovering most of the indirectly related high-level features as well. This is fact has help to improve the retrieval performance of the general and sport queries whose retrieval relies heavily on high-level features.

Comparing with any other current strategy which in general improves the video retrieval performance by 5% to 8%, we believe that our improvement is quite competitive. Moreover, we also believe that the true potential of our WordNet has not been fully revealed due to the fact that the available high-level feature detectors are limited in number as well as their accuracy and coverage.

Chapter 6 Conclusion

As the closing chapter of this report, we briefly summarize what we have done and the conclusion we have reached through the research. In addition, we also discuss about the some possible directions for future research.

6.1 Summary of achievements and Conclusion

To summarize, in order to better associate a query with its relevant high level features, we have used a two-phrase approach to refine and expand the WordNet with a variety of information sources: In phrase I, we introduce and make explicit the relations between a term and its *topic*, *location/environment* and *sound*. In phrase II, we incorporate additional supports for proper noun and real life information.

The resulting WordNet is then evaluated with three sets of evaluations based on different methodologies, namely, relation-based, human-judge-based and retrieval-performance-based evaluation. In all three evaluations, we have validated that the refined and extended WordNet is better than the original version in the sense that 1) it makes more relations explicit 2) it increases the degree of agreement between the query processing result and the relevance ranking from the human judges 3) it improves the retrieval performance.

In conclusion, the limitations of WordNet can be overcome by incorporating information from different sources and the resulting WordNet is more efficient for high level feature deduction.

6.2 Directions for future research

In addition to what we have done to refine and expand the WordNet, some directions for future researches include:

- **Sense disambiguation:** Due to the complex nature of sense disambiguation, in our expansion process, we always use the first sense of the word whenever there is a need to choose among all the senses. In future, sense disambiguation techniques can be incorporated into the refinement and expansion process so that the relation building can be more accurate.

- **Dynamic source of information:** Another possible extension to our research is a daily update to the refined and expanded WordNet using the news articles. In this way the WordNet will always be updated and able to handle the new terms and relations that come up every day.

- **Time axis:** One shortcoming about keeping the source of information dynamic is that things changes as time goes by. The terms and relations introduced today may not be applicable for tomorrow. Therefore, just like the temporal mutual information we proposed in our previous research, if we can extend the WordNet by taking the time factor into consideration, the information would be a lot more accurate as well.

References

1. A. Hoogs, J. Rittscher, G. Stein, and J. Schmiederer. Video content annotation using visual analysis and a large semantic knowledgebase. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2003.

2. Aldo Gangemi, Roberto Navigli, Paola Velardi: The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet. CoopIS/DOA/ODBASE 2003: 820-838

3. Bentivogli, Luisa and Emanuele Pianta. Extending WordNet with Syntagmatic Information, In Proceedings of the Second Global WordNet Conference , pp. 47-53, Brno, Czech Republic, January 20-23, 2004. http://www.fi.muni.cz/gwc2004/proc/109.pdf

4. Chrisa Tsinaraki, Panagiotis Polydoros, Fotis Kazasis, Stavros Christodoulakis, "Ontology-based Semantic Indexing for MPEG-7 and TV-Anytime Audiovisual Content", Multimedia Tools and Applications, Vol 26, Issue 3, p299-325, 2005

5. Chrisa Tsinaraki, Panagiotis Polydoros and Nektarios Moumoutzis and Stavros Christodoulakis, Coupling OWL with MPEG-7 and TV-Anytime for Domain-specific Multimedia Information Integration and Retrieval, http://citeseer.ist.psu.edu/729681.html

6. Christel, M.G., Hauptmann, A.G.: The use and utility of high-level semantic features in video retrieval. In proceedings of the Conference on Image and Video Retrieval, Singapore (2005) 134–144

7. G.C. Stein, J. Rittscher, and A. Hoogs. Enabling video annotatation using a semantic database extended with visual knowledge. In Proceedings of ICME, 2003.

8. Harvard IV-4 Inquirer Dictionary http://www.wjh.harvard.edu/~inquirer/homecat.htm

9. Infogistic: NLPprocessor, http://www.infogistics.com/textanalysis.html (2000)

10. Kennedy, L.S., Natsev, A.P., Chang, S.F.: Automatic discover of query-class-dependent models for multimodal search. In: ACM Multimedia (MM '05). (2005) 882–891

11. Laura Hollink, Marcel Worring: Building a visual ontology for video retrieval. In proceedings of ACM Multimedia 2005: 479-482

12. Luisa Bentivogli and Emanuele Pianta. Beyond lexical units: Expanding WordNet with phrasets. In Proceedings of EACL'03: 10th Conference of the European Chapter of the Association for Computational Linguistics, pages 00–00, Budapest, Hungary, 2003.

13. Lextek: Onix text retrieval toolkit stopword list, http://www.lextek.com/manuals/onix/stopwords1.html (2000)

14. LSCOM annotation http://www.ee.columbia.edu/~lyndon/LSCOM/conceptsinfo.html

15. Miller, G.: Wordnet: An on-line lexical database. (1995)

16. Navigli, R. and P. Velardi, 2002. Automatic Adaptation of WordNet to Domains. In Proceedings of the OntoLex 2002. Las Palmas, Canary Islands, Spain.

17. Rada Mihalcea and Dan Moldovan, eXtended WordNet: Progress Report, in Proceedings of NAACL Workshop on WordNet and Other Lexical Resources, Pittsburgh, PA, June 2001.

18. Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 448–453.

19. Snoek, C.G.M., van Gemert, J., Geusebroek, J.M., Huurnink, B., Koelma, D.C., Nguyen, G.P., de Rooij, O., Seinstra, F.J., Smeulders, A.W.M., Veenman, C.J., ,Worring, M.: The mediamill trecvid 2005 semantic video search engine. In Proceedings of the 3rd TRECVID Workshop, NIST (2005)

20. Tat-Seng Chua, Shi-Yong Neo, K.Y.L.G.W.R.S.M.Z., Xu, H.: TRECVID 2004 search and feature extraction task by nus pris. In: TRECVID 2004 workshop, November 15-16. (2004) 159–170 Modeling News Video Retrieval with Semantic Features 13

21. Vossen, Piek. Extending, trimming and fusing WordNet for technical documents. In: *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June 2001. http://www.seas.smu.edu/~rada/mwnw/papers/WNW-NAACL-205.pdf.gz

Appendix: The paper resulted from this project

As we mentioned in the report, we have implemented a system to perform high-level feature deduction, which becomes the motivation of this current work. The details and evaluations of this system are reported in a paper we submitted to CIVR 06. A copy of this paper is attached here for easy reference.

Video Retrieval using High Level Features: Exploiting Query Matching and Confidence-based Weighting

Shi-Yong Neo*, Jin Zhao, Min-Yen Kan, and Tat-Seng Chua

Department of Computer Science, School of Computing,
National University of Singapore, Singapore, 117543
{neoshiyo,zhaojin,kanmy,chuats}@comp.nus.edu.sg

Abstract. Recent research in video retrieval has focused on automated, high-level feature indexing on shots or frames. One of the most important application of such indexing is to support precise video retrieval. We report on extensions of this semantic indexing on news video retrieval. First, we utilize extensive query analysis to relate various high-level features and query terms by matching the textual description and context in a time-dependent manner. Second, we introduce a framework to effectively fuse the relation weights with the detectors' confidence scores. This results in individual high level features that are weighted on a per-query basis. Tests on the TRECVID 2005 dataset shows that the above two enhancements yield significant improvement in performance over a corresponding state-of-the-art video retrieval baseline.

1 Introduction

News video retrieval systems often perform retrieval based solely on automatic speech recognition (ASR) results on the video's audio. This is because ASR, while not fully accurate, is reliable and largely indicative of the topic of videos. Such a transformation of the video retrieval problem into a text-based one has been shown to be effective [1].

To further increase the accuracy and resolution of video retrieval requires analysis and modeling of the video and audio content. The community has investigated this in part by developing specialized detectors that detect and index certain high-level features (HLFs; e.g., presence of cars, faces, buildings, etc). As such, research retrieval systems incorporate both standard text-based information (from ASR and/or closed captions) with results from an inventory of detectors designed to capture HLFs. In order to carry out a large-scale retrieval of video in a real time environment, most features have to be extracted and preprocessed during offline indexing. In our current state of the art, systems cannot detect and index (or even conceptualize) every possible useful high-level semantic feature. Therefore, it is necessary to carry out inference on a limited set of detectable HLFs that cover and support a wide range of queries. Thus we focus on using only ASR and the HLFs to support news video search.

We offer two extensions to this basic framework that enhance the contributions of HLFs, based on two observations. First, we note that many HLFs have a natural textual description (e.g., "car", "face") that have not been widely utilized for retrieval. We show

* Supported by Singapore Millennium Foundation (SMF).

how to match such feature descriptions with the user's textual query to enhance retrieval performance in a time-dependent manner. We approach this by employing morphological analysis followed by selective expansion using the WordNet [2] lexical database on both the feature descriptions and the user's query. The stronger the match between the descriptions and the query, the more important this HLF is to the query. However as queries are often time-sensitive (featuring new personas, corporations each day, using only the static information in WordNet is not enough. Thus we further employ the use of comparable news articles within the same period of time to further build and expand word-based relationships. Crucially different from previous work that only employs lexical expansion, our method fuses both static lexical information and with dynamic correlation by calculating time-dependent mutual information [3].

Secondly, as various HLF detectors vary greatly in performance, it is necessary to consider their accuracies in the fusion process. Currently, retrieval systems have used the output of such batteries of detectors "as-is", without considering the confidence of individual detectors. For example, detectors for faces are fairly mature and robust, whereas detectors for objects such as cars and animals have relatively low precision. We therefore introduce a performance-weighted framework which accounts for this phenomenon. Different from previous work, it evaluates the accuracy of individual high-level detectors during training/validation and utilizes probability of correct detection in feature weighting during testing.

We have validated our approach on the TRECVID 2005 dataset [4] and queries. Our experimental results show that the appropriate use of HLFs in retrieval can outperform text-based systems and improve results on a representative state-of-the-art multimodal retrieval systems in real-time.

2 Use of High-level Features in Video Retrieval

Starting from text-based search, video retrieval has incorporated the use of low-level video features (e.g., color, motion, volume) and, more recently, high level features for specific objects or phenomenon (e.g., cars, fire, and applause). To create such high level features, recent work has taken a machine learning approach, where each HLF detector is trained against an annotated corpus of video clips [4, 5]. A well-known example is the LSCOM set, which contains approximately 1000 concepts which can be used for video annotation. In TRECVID 2005, the LSCOM-lite set (a LSCOM subset of 39 interesting concepts) have been selected and tagged to provide training examples of approximately 50,000 shots or 70 hours of video. The detectors trained using these examples introduce useful and partial semantics to retrieval systems.

The IBM group used a fusion of low-level features and HLFs based on two learning techniques: Multi-example Content Based Retrieval (MECBR, a k-NN variant) and support vector machines (SVMs) [6]. Their system automatically maps query text to HLF models. The weights are derived by co-occurrence statistics between ASR tokens and detected concepts as well as by their correlations. They have shown that the use of HLFs significantly improves the retrieval accuracy.

Columbia University's team represented the text queries and subshots in an intermediate concept space which contain confidences for each of the 39 concepts [7]. The sub-

shots are represented by the outputs of the concept detectors for each concept, smoothed according to the frequencies of each concept and the reliability of each concept detector. The text queries are mapped into the concept space by measuring the similarity between the query terms and the terms in the concept's description. This approach was applied to automatic, manual, and interactive searches, yielding high performance for the few topics which have high-performing correlated concepts.

The MediaMill group [8] also extended the LSCOM-lite set by increasing the HLF pool to 101 features, some original as well as some recycled from the previous TRECVID tasks. This set provides a larger pool of semantic features for retrieval. Other top performing interactive retrieval systems from Informedia [9] and DCU [10] also show effective methods of integrating high level semantic features. One may conclude that even though the HLF detection accuracies are much lower than low level features, HLF have shown to be more useful for semantic queries.

In this work, we use a set of 25 HLFs for news video retrieval. Our primary reason for choosing this set is that the corresponding detectors are readily available and have been trained previously on both the TRECVID 2004 and 2005 HLF task. In addition, they have shown to be useful in retrieval in previous work [11][12]. These 25 features are targeted towards identifying the video genre, objects, backgrounds and actions, as shown in Figure 1. The underlying classification technique used can be found in [11]. The HLF task requires system to return ranklists of maximum 2000 shots for each HLF. Our system achieves a mean average precision (MAP) of 0.22. In order to maximize the detection accuracy, we propose to combine the best available HLF detection results from various participating groups. We only select ranklists which have a MAP $\geq .2$ and above (including IBM's HLF detector set [6], which has a .33 MAP). The score of shot S_c containing HLF_k is calculated using the following equation.

$$Score(S_c|HLF_k) = \alpha \sum_j Contains(S_c) + (1-\alpha) \sum_j \frac{maxPos - Pos(S_c)}{maxPos} \quad (1)$$

where $Contains()$ is an indicator function that checks whether a shot is present on the ranklist and the second term produces a normalized score in the range of $[0-1]$ that linearly weights the position (Pos) for the shot on the ranklist. The resulting ranked list achieves a MAP of 0.38.

- Genres: anchorPerson, $commercial, politics, \underline{sports}, weather, financial$
- Objects: $face, \underline{fire}, \underline{explosion}, \underline{car}, \underline{U.S.flag}, boat, aircraft, \underline{map}, \underline{buildingExterior}, \underline{prisoner}$
- Scene: $\underline{waterscape}, \underline{mountain}, sky, outdoor, indoor, disaster, vegetation$
- Action: $\underline{peopleWalking}, peopleInCrowd$

Fig. 1. High level features used by our system. The ten underlined features indicate the required features from the TRECVID HLFs; italicized features come from LSCOM-lite.

However, having a well-trained, accurate set of HLF detectors is not sufficient for precise retrieval. This is because each HLF detector models a specific phenomenon, and

which detectors are useful for particular queries varies greatly. Correctly determining and matching detectors to queries is therefore a critical task. Past systems have done this matching manually or using simple automated methods by unsupervised clustering or simple expansion using dictionaries. In this work, we leverage the textual descriptions of the HLF set for matching and also incorporate the confidence of the detectors in our fusion process. This is illustrated in Figure 2 which shows the placement of both of these modules in our processing framework for large-scale news video retrieval. We describe this two-fold approach in the following two sections.

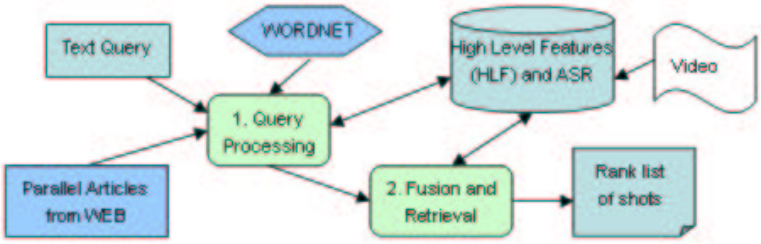

Fig. 2. Retrieval Framework

3 Query processing for HLF weighting

As user queries are usually short and contain insufficient context to perform a precise retrieval, we employ previous work on query expansion techniques using external resources [13] and query classification [14] during query processing to expand the user's original query (denoted Q_0) obtain an initial expanded query, Q_1.

WordNet has been a heavily utilized source of ontological lexical information in text retrieval. In text retrieval, systems relate terms by synonymy, hypernymy, hyponymy and overlap in definitions (gloss). We employ a similar technique, close in spirit to the MediaMill group [15] to determine the match between a detector and a query. Both the short one or two word original description of the detector and the user's expanded query (denoted as HLF_0 and Q_1, respectively) are expanded using WordNet. Both pieces of data are first tagged for part-of-speech using a commercial product, and then closed-class words and words on a 400+ word video domain stopword list [16] are removed.

Unlike previous work, we include terms from the WordNet gloss as we have found that the terms extracted from the gloss differs significantly from those extracted from the synonymy and the hyper/hyponym hierarchy. The former sometimes provides visual information about an object – its shape, color, nature and texture; whereas the latter only provides direct relations (e.g., *aircraft* & *airplane*; *fire* & *explosion*). For example, the word *boat* can not be related to *water* by virtue of any relationship link in WordNet, but by its gloss – *"a small vessel for travel on water"*.

The expanded terms (Q_2, HLF_1) are then empirically weighted based on an approximate distance from the original terms (Q_1, HLF_0). Expansion terms obtained from synonymy, hyponymy, gloss are weighted in this specific order, with the synonyms having the highest weight and gloss the lowest.

A final matching phase is done to determine which high level features are most relevant to the query. To match HLF_1 to Q_2 we use the information-content metric of Resnik [17] (as was done in [18]), which equates similarity with the information content of the pair of words' most specific common ancestor: $Resnik(t_i, t_j) = IC(lcs(t_i, t_j))$ where $lcs(t_i, t_j)$ is the most deeply nested concept in the *is-a* hierarchy that subsumes both t_i and t_j). Here, we factor in the expanded term weights from the previous step.

$$Sim\_Lex(Q_j, HLF_k) = (\sum_{t_q \epsilon Q_j} \sum_{t_f \epsilon HLF_k} Resnik(t_q, t_f))/(|Q_j| * |HLF_k|)) \quad (2)$$

After summing all such scores for each HLF, the k top scoring HLFs are taken with their weights and used in the final retrieval. When k is set to $|HLF|$, the total number of HLF detectors, all HLFs are used with their corresponding weights; when k is set to 1, only the highest scoring HLF is used for retrieval.

This framework would be fine for video in which the associated text information is aligned exactly to the clip. However, in professionally edited video, a time lag between the text content and video often occurs for stylistic reasons. And normally, the speech will come before the actual presentation of the visual information. We therefore carry forward n seconds of speech of each preceding shot to its succeeding shot.

Lexical similarity as computed from static dictionaries may not always be most suitable for news, especially because of news' transient nature. Aside from helping to increase the coverage of how Named Entities related to common words, it can also help refine the relations between words already linked together by WordNet. For example, the concept *fire* and *explosion* are associated in WordNet, but in news stories the relationship between fire and explosion varies greatly. Stories about a chemical factory explosion and terrorist bombings are likely to have both terms highly correlated, but a stories on forest fires are unlikely to contain the *explosion* concept. If a system relies solely on lexical links between words as processed from a dictionary, the systems may return forest fire shots when a user is searching for explosions. Another example are the HLFs of *car*, *boat* and *aircraft*. These three features are all related to each other as means of transportation, and thus are highly weighted and correlated to each other in a standard WordNet based expansion. As such it is possible a search where the *car* HLF is relevant would also shots with boats as relevant due to this problem. This is despite the fact that the three HLFs are largely mutually exclusive in actual text corpora and video data. To overcome these problems, we sampled external (e.g., non TRECVID) sources of news to model the dynamic weighting of similarity between HLFs across time. We use the external news articles to calculate the co-occurrence of $feature_1$ and $feature_2$ with respect to time. The relationship between $fire$ and $explosion$ is thus modified according to their co-occurrence in the external articles. If no news articles directly relate $explosion$ and $fire$ during a certain time period t, the link weight between $explosion$ and $fire$ is reduced accordingly. Equation 3 gives our time-sensitive similarity measure.

$$Sim\_Lex_t(Q_j, HLF_k) = \alpha Sim\_Lex(Q_j, HLF_k) + (1 - \alpha) MI(Q_j, HLF_k|t) \quad (3)$$

4 Confidence fusion and retrieval

The retrieval step is a text-based retrieval scoring function enhanced with HLF confidence scoring. We choose to use the text scoring function $Text(S_i)$ from Chua *et al.*'s work [11]. This scoring function utilizes the query's class and other additional contextual information to retrieve the relevant documents. It was shown experimentally in their work, that irrelevant segments were eliminated while recall was maintained.

As the HLF detectors perform at different accuracy levels, we must consider their precision in retrieval. Using the available training samples, we obtain accuracies for each detector using 5-fold cross validation, in terms of mean average precision. The final score of shot S_i with respect to query Q is given below.

$$Score(S_i) = \alpha * Text(S_i) + \beta * \sum_{HLF_k \epsilon S_i} Conf(HLF_k) \times Sim\_Lex_t(Q, HLF_k) \quad (4)$$

where $Conf(HLF_k)$ is estimated MAP of the $Detector_k$. The score for each shot will be computed base on the available textual features as well as the HLFs and their detection confidence with respect to the query.

5 Evaluation

The goal of our evaluations is to show the efficacy of both modules: HLF weighting and confidence-based fusion. For the weighting, we can measure how well our automatic weighting scheme agrees with the importance assigned to the HLFs by human subjects. For fusion, we measure the gain in retrieval performance when incorporating confidence in the HLF weighting scheme. We also measure the synergy when employing both modules together in the retrieval framework.

5.1 HLF weighting agreement with human subjects

We asked 12 paid volunteers to take a survey that assessed how they would weight HLFs in video retrieval. All participants were either university postgradute or undergraduate students and had not used textual descriptions to search for videos before. We selected 8 queries from past TRECVID queries that were representative of different semantic class (e.g., "George Bush", "Basketball players on court", "People entering and leaving buildings", etc), and asked the participants to first freely associate what types of HLFs would be important in retrieving such video clips, and second, to assign the importance (on a scale from 1-5) of the specific HLF inventory set used in our system (c.f., Figure 1) for the same 8 queries. "Important" (rating 5) here refers to a strong positive correlation between the HLF and the query, "unimportant" (rating of 1) a negative correlation. In total, we gathered approximately $25 \times 8 \times 12 = 2400$ judgments (some judgments were skipped by participants).

An analysis of the free association subtask shows that over 90% of the responses are concrete nouns, confirming earlier work that searchers focus on nouns as cues for retrieval. In addition, although only 5% of the features used in our experiments are mentioned explicitly in the free association task, some were later ranked as "Important" by

participants in the second subtask. Calculating the interjudge agreement using Kappa, we found only a low agreement ranging from 0.2 to 0.4, which varied with the search task. The low agreement may be partially due to the inexperience of the participants in searching video. Following the analysis given in Christel and Hauptmann's user study [19], we also calculated the standard deviation of a feature's score for each search task, shown in Table 1. Similar to their study, we also show a low level of agreement between judges, albeit lower than in their previous study.

Table 1. Importance ratings of features across all 8 search tasks. Blank cells indicate high standard deviations (above 0.7). Features sorted by standard deviation.

Feature (Avg. s.d.)	Search Task							
	Map	Tree	Office	Basketball	Ship	Hu Jintao	George Bush	Fire
Fire & Explosion (0.6)		1.4	1.4	1.3		1.3	1.4	
Car (0.7)		1.4	1.2	1.3	1.4	1.9	4.4	
Boat (0.8)				1.1		1.4	2.5	
Aircraft (0.9)	1.7			1.1		2.1		
Face (1.1)	1.2		3.5	3.0	2.0			2.0
US Flag (1.0)	2.2							
People Walking (1.2)								
Anchor person (1.3)								
Map (0.7)						1.5	1.6	
Prisoner (0.6)			1.1	1.1				
People in Crowd (1.2)								
Commercial (1.2)	1.5	2.6						
Politics (1.0)		2.4	1.3	1.8		4.3		
Sports (0.8)		1.5	1.4					
Weather (0.9)	1.6		1.4	1.6				
Financial (1.0)	1.5	3.3	1.5					
Disaster (0.9)	1.9	1.5						
Building Exterior (1.0)								3.5
Waterscape (1.0)				1.1	4.1			
Sky (1.0)			1.1	1.6		1.7	1.6	
Outdoor (1.0)	1.6		1.1					
Indoor (0.9)			4.5		1.5			
Vegetation (0.8)			1.3	1.3	1.5	1.5	1.4	

In some cases, the degree of agreement is high, especially when the search task mentions the feature directly (e.g., the "Basketball" query mentioning "Sports"). In fact, a trend of HLF rating stability was observed. Ratings for concrete nouns were most stable, followed by backgrounds and video categories, and with those describing actions being the most variable or unreliable. We also note that negative correlations (scores close to 1.0) are prominent in our dataset. We feel this is quite reasonable, as only a few HLFs are usually relevant per query. We have also computed the Kappa value between the HLF rankings from our system and the ones from the human judges. The value ranges between 0 to 0.25 with a mean value of 0.145. We believe this varying level of agreement is due to the fact that WordNet expansion works well for hypernym and hyponym relations, but less so for other relation types. As a result, the overall agreement is weak. We plan to look into the problem of how to enrich the WordNet so that it is capable of discovering other relations in the near future as an extension to this work.

5.2 Text retrieval and Query matching

We follow the evaluation standards in TRECVID 2005 automated search task. The task consist of 24 queries (e.g. Find shots of Condoleeza Rice; Find shots of an office set-

ting). A maximum of 1,000 shots are returned for each query and the performance is measured in MAP.

We leverage and modify an existing state-of-the-art retrieval system [11] to perform the required text retrieval. The text-based retrieval engine uses the text query to retrieve pre-segmented passages of text in the (possibly machine translated) ASR transcripts. These segments correspond to phrase level video segments in the corpus. The video segment associated with the matched phrase and the segment immediately afterwards are retrieved as the retrieved results. The reason that the segment afterwards is also included is because that audio descriptions of events often proceed the relevant video segments in edited video, such as news. This text-based baseline system also incorporates query expansion using external news resources. The resulting text retrieval system achieves an MAP of 0.063 based on the TRECVID 2005 dataset and queries. In comparison, the top three performers in TRECVID 2005's search tasks yield MAP of 0.67,0.62,0.61 respectively, showing that our text baseline is competitive.

To test the effectiveness of our query matching techniques, we further compare the performance to this system [11] which uses heuristics to weight HLFs to individual queries. When HLFs are integrated into the text-only system [11], the jump in MAP is significant (from 0.063 to 0.104) and validates earlier reported work. To test the effectiveness of the various components in the query matching module, 3 runs have been carried out. Run1 is based on the query-matching algorithm without using the WordNet glosses, while Run2 includes glosses. Run3 will include the use of glosses as well as temporal MI as stated in Eqn. 3.

Table 2. The performance in MAP combining textual features and HLF in retrieval. Percentages indicate performance gain over the baseline system.

Technique of Using HLF + Text	**MAP**
Baseline Heuristics weighting by [11]	0.104
Run1. Automated HLF query matching without Gloss	0.106 (+1.9%)
Run2. Automated HLF query matching with Gloss (Eqn.2)	0.110 (+5.8%)
Run3. Automated HLF query matching with Gloss + Temporal MI (Eqn. 3)	0.113 (+8.6%)

The Table shows that the use of HLFs during fusion have outperformed the text-based retrieval system by more than 50%. This is conclusive as textual feature alone are not reliable to pin-point shots which are relevant to the query. Run1 and Run2 indicates that the use of WordNet glosses is positive as the performance increases from the MAP of 0.106 to 0.110. Run3, which uses all the components obtain a MAP of 0.113 which is 8.6% better than the baseline system of 0.104. The main improvement comes from the sport and general queries. Queries which are directly or indirectly related to the available 25 HLFs benefits the most. This suggest that as more HLFs are added, a better performance can be obtained. The MAP performance is higher due to its re-ranking of relevant shots as it takes all HLFs into consideration during fusion (i.e., k==25). This performance is also better than a similar evaluation run (MAP of 0.070) submitted by IBM [6] which uses only text and HLFs.

5.3 Confidence-based fusion and A/V Integration

For the confidence-based fusion, we carried out 2 more runs to investigate the effects of considering HLF detection accuracy in the retrieval. As Run1 to Run3 uses HLF detection result without considering the accuracy of the various HLF detectors (normal fusion), we added Run4 which applies confidence-based fusion as in Eqn. 4. Run5 is designed to investigate the overall performance of the system by integrating other A/V features including low level features from [11]. The fusion is done by modifying the query-dependent multimodal fusion function in [11] to accommodate Eqn. 4. The results of these experiments are reported in Table 3.

Table 3. Aggregate MAP of the system. Percentages indicate performance gain over the baseline system.

Experiment	**Normal fusion**	**Confidence-based fusion using (Eqn. 4)**
Run4. Text + HLF	0.113 (+8.6%)	0.117 (+12.6%)
Run5. Text + HLF + A/V features[11]	0.127 (+22.1%)	0.131 (+25.9%)

The result shows that the use of confidence-based fusion yield significant improvement over normal fusion. Run4 based on confidence-based fusion is able to achieve a MAP of 0.117. This performance is statistically comparable to top performing submissions. The run that incorporate the rest of the A/V features obtains the MAP of 0.127 and 0.131 respectively, which is better than the best published MAP of 0.123 in TRECVID 2005 automated search task. The bulk of improvement come from the general queries as they depend largely on the use of HLFs as evidence of relevancy. Person-oriented queries on the other hand have less significant improvement as textual features and video OCR still constitute the main score. As the confidence-based fusion and the automated HLF to query matching affect different parts of the retrieval system, they can be combined easily, producing largely independent gains on MAP.

6 Conclusion

As video analysis has advanced to building high-level semantic features from low level ones, schemes that judisciously employ such HLFs are needed. We explore two distinct and complementary approaches to extend the current frameworks of such multimodal retrieval systems. We have investigated methods to automate and expand the matching of HLFs to user query terms. In particular, our query to HLF mapping methods examine 1) the use of dictionary definitions (WordNet's glosses) to help relate terms, and 2) time sensitive mutual information to make sure that the scores are sensitive to the timeframe and story distribution in the video corpus. Overall, our newly Text + HLF retrieval system is able to outperform baseline system and achieve similar results to top performing automated systems reported in TRECVID 2005. This framework is further tested by integrating other A/V features and the resulting performance is better than the best reported result.

References

1. Hauptmann, A., Chen, M.Y., Christel, M., Huang, C., Lin, W.H., Ng, T., Papernick, N., Velivelli, A., Yang, J., Yan, R., Yang, H., Wactlar, H.D.: Confounded expectations: Informedia at trecvid 2004. In: TRECVID, 2004. (2004)
2. Miller, G.: Wordnet: An on-line lexical database. International Journal of Lexicography (1995)
3. Neo, S., Goh, H., Chua, T.: Multimodal event-based model for retrieval of multi-lingual news video. In: IWAIT. (2006)
4. Over, P., Ianeva, T.: Trecvid 2005 an introduction. In: TRECVID, 2005. (2005)
5. Smeaton, A.F., Kraaij, W., Over, P.: Trecvid-an overview. In: TRECVID, 2003. (2003)
6. Amir, A., Iyengar, G., Argillander, J., Campbell, M., Haubold, A., Ebadollahi, S., Kang, F., Naphade, M.R., Natsev, A.P., Smith, J.R., Te?i?, J., Volkmer, T.: Ibm research trecvid-2005 video retrieval system. In: TRECVID, 2005. (2005)
7. Chang, S.F., Hsu, W., Kennedy, L., Xie, L., Yanagawa, A., Zavesky, E., Zhang, D.Q.: Columbia university trecvid-2005 video search and high-level feature extraction. In: TRECVID, 2005. (2005)
8. Snoek, C.G.M., van Gemert, J.C., Geusebroek, J.M., Huurnink, B., Koelma, D.C., andO. De Rooij, G.P.N., Seinstra, F.J., Smeulders, A.W.M., Veenman, C.J., Worring, M.: The mediamill trecvid 2005 semantic video search engine (draft version). In: TRECVID, 2005. (2005)
9. Hauptmann, A.G., Christel, M., Concescu, R., Gao, J., Jin, Q., Lin, W.H., Pan, J.Y., Stevens, S.M., Yan, R., Yang, J., Zhang, Y.: Cmu informedia's trecvid 2005 skirmishes. In: TRECVID, 2005. (2005)
10. Foley, C., Gurrin, C., Jones, G., Lee, H., McGivney, S., O'Connor, N.E., Sav, S., Smeaton, A.F., Wilkins, P.: Trecvid 2005 experiments at dublin city university. In: TRECVID, 2005. (2005)
11. Chua, T.S., Neo, S.Y., Goh, H.K., Zhao, M., Xiao, Y., Wang, G.: Trecvid 2005 by nus pris. In: TRECVID 2005. (2005)
12. Chua, T., Neo, S., Li, K., Wang, G., Shi, R., Zhao, M., Xu, H.: Trecvid 2004 search and feature extraction task by nus pris. In: TRECVID 2004. (2004)
13. Yang, H., Chua, T.S., Wang, S., Koh, C.K.: Structured use of external knowledge for event-based open-domain question-answering. In: SIGIR 2003, Canada, Jul 2003. (2003)
14. Neo, S., Chua, T.: Query-dependent retrieval on news video. In: MMIR'05 workshop in SIGIR'05. (2005)
15. Snoek, C.G.M., van Gemert, J., Geusebroek, J.M., Huurnink, B., Koelma, D.C., Nguyen, G.P., de Rooij, O., Seinstra, F.J., Smeulders, A.W.M., Veenman, C.J., , Worring, M.: The mediamill trecvid 2005 semantic video search engine. In: Proceedings of the 3rd TRECVID Workshop, NIST (2005)
16. Lextek: Onix text retrieval toolkit stopword list, http://www.lextek.com/manuals/onix/stopwords1.html (2000)
17. P., R.: Semantic similarity in a taxonomy: An information- based measure and its applications to problems of ambiguity in natural language. Journal of Artificial Intelligence Research,11 (1999) 95–130
18. Kennedy, L.S., Natsev, A.P., Chang, S.F.: Automatic discover of query-class-dependent models for multimodal search. In: ACM Multimedia (MM '05). (2005) 882–891
19. Christel, M.G., Hauptmann, A.G.: The use and utility of high-level semantic features in video retrieval. In: Conf. on Image and Video Retrieval, Singapore (2005) 134–144