# Using Semantic Contents and WordNet$^{TM}$ in Image Retrieval

Y. Alp Aslandogan[1], Chuck Thier[2], Clement T. Yu[1], Jon Zou[1] and Naphtali Rishe[3]

1: Department of EECS, University of Illinois at Chicago
(aslan,yu,jzou@dbis.eecs.uic.edu)
2: Tribune Media Services (CThier@tribune.com)
3: Florida International University

## Abstract

Image retrieval based on semantic contents involves extraction, modelling and indexing of content information. While extraction of abstract contents is a hard problem, it is only part of the bigger picture. In this paper we use knowledge about the semantic contents of images to improve retrieval effectiveness. In particular we use WordNet, an electronic lexical system for query and database expansion. Our content model facilitates novel uses of WordNet. We also propose a new normalization formula, an object significance scheme and evaluate their effectiveness with real user experiments. We describe the experiment setup and provide quantitative evaluation of each technique.

## 1 Introduction

Content based access to multimedia has recently gained much interest. Analysis of image, audio and video resources to bridge the gap between the media level content information and the high level abstract content is beginning to produce satisfactory results in certain domains. Image data is arguably the most commonly used multimedia data type. In this paper, we focus on techniques for effective image retrieval based on semantic contents.

A variety of features have been used to index and retrieve images based on contents. These include color, texture and primitive shapes [FBF+94, NBE+93, OS95, Gup95, GZCS94], user drawn shapes or sketches [NBE+93, KKOH92], keywords, textual description or captions [SQ96, RS95, CPLJ94, GZCS94, LW93, HCK90], user defined attributes [JG95], iconic object and relationship descriptions [ATY+95, CJ91]. Combinations of these methods were also used [LC96, HHTK96, WNM+95].

While some systems, notably [NBE+93, Gup95, SC96] perform automatic extraction of low level features others such as [ATY96, ATY+95, SQ96, CPLJ94, LW93, HCK90] require manually produced meta-data to be associated with each image.

Although automatic extraction of abstract contents from multimedia is in its infancy, some promising results are reported in restricted domains. These include human face detection [WKSS96] and recognition [BP94, SI92], identification of attributes such as age [KdV94], gender and facial expression for humans [YD94], basic movements such as entering and exiting a scene, bringing into or removing an object from a scene [Cou96], simple spatial relationships, etc. based on either static images or image sequences [SC96, AHKR96]. Analysis of the sound track and text recognition in TV and video provides yet another avenue for obtaining further content information [WKSS96, Lie96, ZTSG95]. It is likely that a combination of automatic and semi-automatic techniques will be utilized to provide a semantically rich description of visual data.

The purpose of this paper is to investigate the techniques which can provide effective retrieval, assuming that some meta-data have been obtained for the images. These techniques are applicable independent of whether the meta-data is generated automatically or manually. Our system is similarity based, i.e. with respect to a given query, each image is assigned a degree of closeness (or similarity). Images with higher similarities are retrieved ahead of images with lower similarities. The techniques we consider in this paper are the following:

1. **Normalization:** Consider two images which have the same set of semantic elements (entities or objects, attribute values and relationships) in common with a user query, but the two images have other elements as well. Normalization means that similarities are based on not only the set of matching elements, but also on those in the images as well. We study two ways of normalization and compare the results against no normalization. One of these normalization techniques (minimal normalization) is proposed by us and is very different from the normalization method which is commonly used in text retrieval.

2. **Object Significance:** Each object in an image may have a different degree of significance (importance) in characterizing the contents of the image. We assess the usefullness of a particular scheme where significance is determined by relative size and closeness to center.

3. **Automatic Use of an Electronic Thesaurus:** We use an electronic thesaurus (WordNet$^{TM}$) to expand both user queries and the meta-data associated with the images. We discuss how different features of WordNet can be incorporated to improve retrieval performance. Our content model based on objects, attributes and relationships allows us to use WordNet in a fully auto-



*SIGIR* 97 Philadelphia PA, USA

matic manner and improve retrieval effectiveness unlike in text retrieval.

We perform a number of experiments to determine the degrees of usefulness of these techniques individually and in combination. We find that the minimal normalization formula proposed here and specific uses of WordNet yield significant improvement in retrieval effectiveness.

The rest of the paper is organized as follows. In Section 2 we give an overview of the previous version of SCORE [ATY96, ATY+95, SYH94], briefly describing the system components and the retrieval process as performed by the system. The techniques of minimal normalization and object significance are discussed in Section 3. Section 4 discusses the use of an electronic lexical system (WordNet$^{TM}$) in image retrieval. The results obtained with this system are compared with those obtained with a custom-built thesaurus and those without using any thesaurus. In Section 5 we describe our experimental setup We present the results of these experiments in Section 6 and discuss whether and why these techniques yield improved retrieval effectiveness. We conclude in Section 7.

## 2 Review of the System

The SCORE system consists of two modules: A visual query tool and a search engine. The two modules interact with a simple protocol and hence can be tailored individually for specific purposes.

### 2.1 Visual Query Tool

The basic goal in designing the visual query tool, CANVAS, has been to facilitate intuitive querying for non-expert users. Since users of an image retrieval system may range from graphic artists to physicians and eventually house-holds, eliminating the need for a query language is essential. In our earlier experiments [ATY+95], real users were able to use the system effectively within minutes. With no manual and training session, most users quickly began querying with varying degrees of sophistication.

The user interface consists an icon palette and a query window. The icon palette contains icons denoting objects. Each object has a predefined set of attributes whose values can either be selected from a list or defined by the user. Figure 1 shows an example query (a man carrying a boy) and resulting thumbnail images returned by the search engine. Note that the first two images fully satisfy the query while the other three are only partial matches.

Relationships among objects which are classified into two types: *action* and *spatial*. Spatial relationships are predefined to facilitate inference [SYH94].

The user interface is extensible in the sense that users can add new object types and select appropriate attributes for them easily. By providing icons for objects of interest the user can tailor the palette to their needs. A set of 30 attributes are predefined with their input methods and default values where applicable. The users can choose a subset of these attributes to describe new object types or define new attributes based on the input methods.

Other features of the user interface include

- Object exclusion (negation),
- Viewing the contents of returned images,

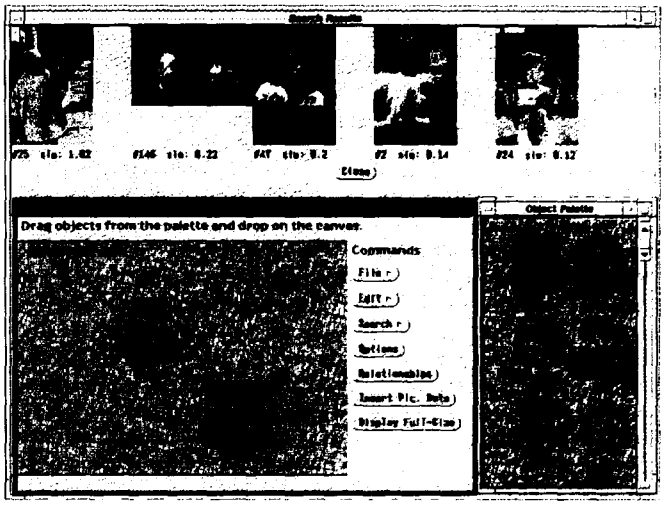


Figure 1: A sample search and resulting thumbnails.

- Defining exact or fuzzy values for attributes where applicable.

### 2.2 Search Engine

The search engine is written in ANSI-C and runs on top of UNISQL/X, a hybrid DBMS with object oriented features and an SQL-like query language. The search engine and the visual query tool interact through a simple interface. Hence it is possible to use the same search engine with a different visual query tool and vice versa. A version of our search engine on top of the object oriented database system $O_2$ is also operational. The meta-data which forms the basis for the retrieval is constructed with the same visual query tool (described in subsection 2.1) and stored in various classes.

### 2.3 The Retrieval Process

We compute a similarity value between the user's specification and an image which provides a measure of the closeness between them. The $k$ closest images will be retrieved, where $k$ has a default value of 10 and can be changed by the user. The similarity value is defined to be the sum of the similarities of the entities (together with their attribute values) and their relationships in one image with the corresponding entities and relationships in the user's query. Matching of entities and relationships is by names, synonyms and IS-A relationships. Matches on rare objects and relationships yield higher similarities than those on popular objects and relationships, as governed by the inverse document frequency weight formula [Sal89]. The similarity of a query term (entity name, relationship name or attribute value to a database term is given by:

$$\mathrm{sim}(t_q,t_{db})=\mathrm{idf}(t_q)/(\mathrm{distance}(t_q,t_{db})+1) \qquad (2.1)$$

where $t_q$ is the query term, $t_{db}$ is the database term, $\mathrm{idf}(t_q)$ is the inverse document frequency weight of the query term and

$$\mathrm{distance}(t_q,t_{db}) = \left\{ \begin{array}{ll} 0 & \text{if } t_q \text{ and } t_{db} \text{ are identical} \\ 1 & \text{if } t_q \text{ and } t_{db} \text{ are synonyms} \\ \#\mathrm{edges} & \text{if } t_q \text{ and } t_{db} \text{ are related} \\ & \text{through the IS\_A hierarchy} \end{array} \right\}$$

Matching of attribute values is by both exact matches and fuzzy matches. Exact matches yield higher similarity than fuzzy matches. Relationships are either spatial or non-spatial. Matching of two sets of objects involved in two non-spatial relationships can be inexact, as long as there is a subset of two or more elements which are in common between the two sets. Matching of spatial relationships may involve deduction and reduction using indices [SYH94]. In our earlier experiments involving images extracted from TIME magazine and some travel agency brochures, users familiar with Entity-Relationship concepts obtained very good retrieval results [ATY+95] in the sense that most of the desired images were always ranked among the top three. Retrieval results by users who are not as knowledgeable were not as good. It was also observed that in general, the *inverse document frequency* method, which assigns higher weights to entities, attribute values and relationships occurring less frequently than those occurring more frequently, yields better retrieval results than assigning equal weights to all entities, attribute values or relationship types.

## 3 Techniques for Effective Retrieval

A primary goal in content-based retrieval is to match a user's query with a stored description which may differ from the query in various ways and degrees. In addition, the user's criteria may not be very clear or completely thought out [HCK90]. To facilitate successful retrieval in the presence of such obstacles, we employ the following techniques.

### 3.1 Normalization

For each user query, there may be a large number of images in the database with one or more entities or relationships matching those in the query. Normalization rewards those images where the proportion of entities or relationships that match the query description is relatively high. As an example, suppose that the user builds a query where there are a man, a child and the man is holding the child. There may be many images in the database having a man, a child, and the relationships between the man and the child. Among these, the ones with extra entities or relationships that are not in the description will be assigned lower similarities.

Normalization of term weights is known to yield improved results in standard text retrieval [Sal89]. Different flavors of normalization are used in different systems. Two new normalization techniques for text retrieval are introduced and compared with cosine normalization in [SBM96]. The Xenomania system [BPJ93] normalizes the refined attribute values (as opposed to image similarities) based on the values in the meta-database. Although the names are the same, the concepts are quite incompatible. Because of different modeling paradigms, there are no experimental results showing the usefulness of entity-level normalization in image retrieval. Our aim is to examine this issue in the context of semantic content based image retrieval.

#### 3.1.1 Full Normalization

Each image in the database may have one or more entities and zero or more relationships. By full normalization, the similarity due to entities and relationships in the database image that match those in the query is divided by the total number of entities and relationships in the image. This ratio is then used in the computation of the final similarity. This is similar to normalization performed in text retrieval [Sal89].

#### 3.1.2 Minimal Normalization

In our previous experiments [ATY+95] we have observed that the user queries tend to be briefer than database descriptions. Analysis indicated that certain relevant images suffered from extra entities/relationships in them which are not matched with the query. This observation led to exploration of better normalization formulas which would not punish well-described relevant images. Accordingly, we propose the following new normalization formula.

Let items denote both objects and relationships. Let $C(Q,P)$ be the set of items in a query $Q$ which can be matched with those in the image $P$. Here, the corresponding items in $P$ and $Q$ need not be identical due to IS-A hierarchy matching, fuzzy attribute value matching or inexact relationship matching. Suppose the matching of the i-th item in $Q$ with itself and the matching of the i-th item in $Q$ with the corresponding item in $P$ yield respectively the values $w_i$ and $w_i'$ respectively. It is clear that $0 \leq w_i' \leq w_i$, since exact match yields higher weights than inexact or fuzzy match. Let $M(Q,P)$ denote the total matching value between the items in $Q$ and those in $P$, i.e., it is the sum of the $w_i'$. A similarity value, $SIM(Q,P)$, between $P$ and $Q$ can be defined to be

$$\frac{M(Q,P)}{M(Q,Q)} * \frac{1}{(1+\frac{(M(P-Q,P)*minmatch)}{(M(P,P)*M(P,Q))})}$$

where $P-Q$ denotes the set of items which are in $P$ but do not have corresponding matching items in $Q$ and *minmatch* is the smallest matching value due to the items in $C(P,Q)$ i.e. the smallest $w_i'$. The usage of the various symbols in the formula are as follows.

$M(Q,Q)$ is to ensure that the similarity value, $SIM(Q,P)$, is between 0 and 1. The similarity has the value 1, when the image $P$ is identical to the query $Q$. This can be observed when $P-Q$ is null and therefore $M(null,P)$ is 0. When there are items in P but not in Q, then the denominator is increased by $\frac{M(P-Q,P)}{M(P,P)}$, which is a weighted proportion of unmatched items, multiplied by $\frac{minmatch}{M(P,Q)}$, which is a proportion of matched value. Two properties of this similarity function are:

Let $P_1$, $P_2$ and $P_3$ be three images and $Q$ be a query containing items $I_1$,$I_2$, ..., $I_k$,$I_{k+1}$, ..,$I_s$.
**(1)** If $P_1$ and $P_2$ have the same set of items but $P_2$ has a larger set of items which can be matched with Q than $P_1$, then the similarity between Q and $P_2$ is higher than that between Q and $P_1$. More precisely, if $P_1$ and $P_2$ have the set of items $I_1'$, $I_2'$, ..., $I_k' \cup NC_1$ and $I_1'$, $I_2'$, .. ,$I_k'$,$I_{(k+1)}' \cup NC_2$ respectively, where $I_j$ matches $I_j'$, $1 \leq j \leq k+1$, and $NC_1$ and $NC_2$ are two sets of items which do not have any corresponding matching items with Q, then $SIM(Q,P_2) > SIM(Q,P_1)$.

**(2)** If $P_3$ and $P_1$ have the same set of matched items with Q, but $P_3$ has additional non-matched items, then the similarity between Q and $P_1$ is higher than that between Q and $P_2$. More precisely, with $P_1$ as given in (1) and $P_3$ given as $P_1 \cup NC_3$, where $NC_3$ is different from $NC_1$ and each item in $NC_3$ has no corresponding matching item in Q, then $SIM(Q,P_3) < SIM(Q,P_1)$.

Essentially, these two properties imply that matching of items is much more important than mismatches of items to the extent that any additional match yields a higher similarity (property (1)). Mismatches can be used to differentiate two images which have identical matches but different mismatches ( property (2)). In comparison with full normalization, less emphasis is placed on mismatches. We investigate the extent to which this similarity function reflects user's relevance judgements.

### 3.2 Object Significance

The *object significance* feature is introduced to the system to indicate the importance of an object in an image. Meta data for each object includes a value for the attribute *significance* which indicates the subjective relative importance of the object in the image. In this paper, the significance of an object is approximated by the size (area) of the object and its location (closeness to the center, bacgrsound vs foreground). Clearly, it is possible that a user may be more interested in a small object than a large one. However, in the absence of other subjective information, our method assigns higher significances to entities with larger sizes than entities of smaller sizes in the same image. We will examine whether this method is reasonable.

If objects can be automatically or semi-automatically extracted, object significance given by our definition can be automated. With semi-automatic object identification techniques like "flood-fill" and "active-contours" [NBE+93] the area information is readily available.

## 4 Using an Electronic Thesaurus for Query and Database Expansion

WordNet is an electronic lexical system developed by George Miller and his colleagues at Princeton University [MBF+90]. The noun portion of WordNet is designed around the concept of *synset* which is a set of closely related synonyms representing a word sense (meaning). Every word that is in the WordNet has one or more senses and for each sense it has a distinct set of synonyms, and a distinct set of words related through other relationships such as hypernyms/hyponyms (IS_A relation), holonyms (MEMBER_OF relation) and meronyms (PART_OF relation). In our experiments we have used the noun and verb portions of WordNet (Version 1.5).

### 4.1 Different Uses of WordNet

We have compared the effects of three parameters in the use of WordNet. The parameters are the number of senses used, the types of relationships used and whether expansion is performed for meta-data in the database as well as the queries. These are compared against the results obtained with the custom-built thesaurus and against results where neither custom built thesaurus nor WordNet was used.

1. **Single sense versus multiple senses:**
   The first sense returned by WordNet is the most common sense of a word[1]. We have compared using first sense versus using all senses obtained from WordNet.

2. **Word relationship types:**
   We have compared three cases: using synonyms alone, the combination of synonyms, hypernyms and holonyms, and finally using all relationships (synonyms, hypernyms, holonyms, meronyms). Combining options 1. and 2. would give numerous possibilities. We have only performed experiments for some of these possibilities.

3. **Database and/or query expansion:**
   A strategy for database expansion was designed and compared against query expansion alone. This is explained in more detail in the following subsection.

In all retrievals using WordNet, the search engine optional parameters *entity significance* and *minimal normalization* (explained in subsections 3.1.2 and 3.2) were turned on.

[1] G. Miller, personal communication.

### 4.2 Database Expansion

Expanding the meta-data in the database facilitates two new ways of matching query and database entity names which are not possible with query expansion alone. These are:

1. **Through sibling relationship (IS_A hierarchy)**
   Query entity and database entity are descendants of the same entity (see Figure 2).
   Example: Query term `high-rise` matches with database term `building` through common ancestor `structure`.

2. **Through MEMBER_OF relation and IS_A hierarchy**
   Query entity and database entity share a common ancestor through a combination of IS_A and MEMBER_OF relationships.
   Example: Query entity `musician` is a member of a `musical organization` (see Figure 2). Database entity band IS_A `musical organization`.
   During the experiments (including ones not reported here), we have observed that users tended to describe collective entities differently. Some users preferred to describe the whole entity while others preferred to describe the individuals forming the collection. Examples of such differences in description include musician(s) versus a band, player(s) versus a team etc.

These cases are illustrated in Figure 2.

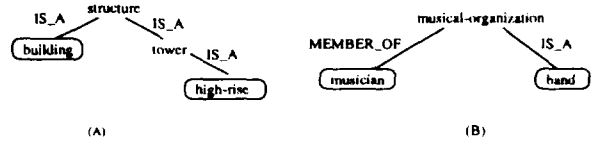


Figure 2: Different ways of matching query and database terms.

The similarity of a query term to a database term after expansion is again computed according to formula (2.1) with some generalization(See section 2). We repeat it here for convenience:

$$\mathrm{sim}(t_q,t_{db})=\mathrm{idf}(t_q)/(\mathrm{distance}(t_q,t_{db})+1)$$

however, in this case the distance($t_q$,$t_{db}$) is

=0 if the two terms are identical or in the same *synset* (a set of close synonyms representing a word sense in WordNet)
=1 if the terms are synonyms (which are not as close as the words in a synset)

=the number of edges along the path between the query term and the database term in a IS_A hierarchy or the combined IS_A and MEMBER_OF hierarchy if they are related through such a hierarchy.

Initially, a threshold on the distance between the query term and the common ancestor was applied as the benefits of such a strategy was demonstrated for caption based image retrieval [SQ96]. In other words, only words t edges or less apart from the query word are searched, where t is the distance threshold. Using thresholds for word-to-word similarities in caption retrieval incorporating WordNet was found to be useful [SQ96]. However, observing that a strict threshold also causes some desired matches to be missed, we have examined an alternative strategy as explained in the following subsection.

### 4.3 Entity Category Verification with WordNet

One parameter that has to be determined in query and database expansion using WordNet is the extent to which the IS_A hierarchy will be traversed. The words far above from the actual query or database terms in these hierarchies are typically very general and can match practically anything [SQ96]. Using a low threshold improves precision of results but certain desired matches may be missed. To illustrate this point suppose that a threshold of 2 is used for expanding query and database terms through the IS_A hierarchy. Consider a query where there is a *building*. In one of the desired images in the database, there are *high-rises*. Clearly, it is desirable to match these two terms. However, the common ancestor of *building* and *high-rise* (*structure*) is 1 edge away from building and 2 edges away from high-rise (see Figure 3). In other words the distance between *high-rise* and *building* is 3. Since we assumed a IS_A distance

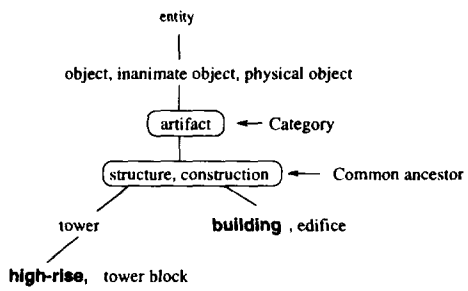


Figure 3: Matching *high-rise* with *building* through the IS_A hierarchy.

threshold of 2, this match is rejected. An alternative is to use either a very high distance threhold or no threshold at all. However, in our experiments, we have observed that such a strategy yields many more undesired matches than the desired ones. So there is a need to balance the flexibility in traversing the IS_A hierarchy with the high precision requirement. We have applied the technique of entity category verification for this purpose. While we totally eliminate the distance thresholds when expanding the query and database terms, the candidate matches are verified by comparing their categories. For instance, the match of *building* with *high-rise* is accepted since they both belong to the category of *artifact* while the match of *car* with *building* is rejected since *car* belongs to the category *vehicle* [2].

When the entity types are known apriori, the verification of categories is easy. However, since we do not limit the entity types a user can chose, a more general method is required. One way of doing this is to use the noun classes of WordNet. Since every noun in WordNet is in an IS_A hierarchy (indeed possibly more than one hierarchies), for every user defined entity, the category could be determined as long as the entity type is found in WordNet.

WordNet divides nouns into 25 classes. These include act, animal, artifact, attribute, body, ...etc. According to the authors, these classes are "organized in such a way as to make the statement of an adjective's selectional preferences as simple as possible" [MBF+90]. Based on our sample queries and the database contents, we have added two new classes (*vehicle* and *body of water*) to the above and use them as our entity categories. Since all the category terms are from the IS_A hierarchies of WordNet, categories are identified automatically.

Figure 4 illustrates the category verification process. In

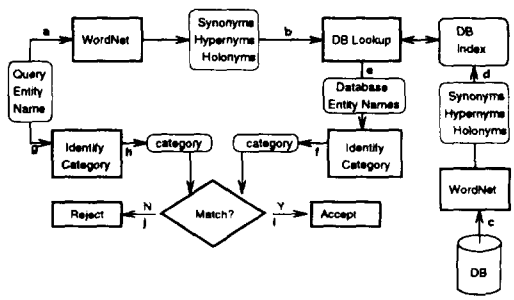


Figure 4: Verifying entity categories.

Figure 4 the query entity name is input to WordNet to get its synonyms, hypernyms and holonyms (step a). These words are then used as index keys in retrieving entity ids from the database (step b). On the database side, the database entity names are input to WordNet (Step c) and the resulting synonyms, hypernyms and holonyms are used to index the database (step d). When the expanded query and the expanded meta-data are matched (step e) the categories of matching database entities are determined (step f) and compared with that of the query term (step h). If the categories are the same, the candidate match is accepted (step i). In case of a category mismatch the database entity is eliminated from further consideration as a candidate match for the query entity (step j).

The choice of category words is in fact domain dependent. For domains where a particular terminology is used (for instance medical or jounalistic image databases) selection of categories can be a non-trivial task. For our image database, which consists of campus life, business and nature images, we have found the WordNet noun classes to be satisfactory with little modification [3].

## 5 Experiment Setup

We have conducted experiments with 6 different volunteer subjects. The subjects consisted of the following: Two middle aged non-university educated males with little or no

[2] Note that a common ancestor does not need to be the category of either term.

[3] Specifically, we have added two new categories to the noun classes of Wordnet

background in computers, one middle aged university educated female[4] with experience in photographic databases and familiarity with general computer use and three middle aged university educated males with good background in computers.

The subjects were given an initial description of the system and allowed to play with it for 5-10 minutes before starting the experiment. Each subject performed retrievals separately and did not observe other subjects.

A sample of 250 images consisting of a variety of domains (nature, office-business, sports, construction, campus life etc.) was collected from commercially available CDs (TIFF format) and two UIC photoshop CDs (KODAK Photo-CD format). The meta-data describing these images was entered into the database by using the same interface used for querying (Figure 1). The meta-data consisted of approximately 1000 objects and 350 relationships with an average of 5-6 attributes per object.

A total of 32 queries were submitted by the subjects. Each subject submitted either 5 or 6 queries. On the average, the queries included 2-3 objects and 4-5 attributes per object. The average number of relationships per query was less than one. Since all images have to be evaluated against each query in determining their relevance, it is not practical to perform experiments of a larger scale. The relevance was determined in a two-step process. During the experiments, the subjects' responses to the retrievals were observed. The reason(s) why they found an image to be relevant or irrelevant were recorded. The subjects determined the relevance of only the retrieved images as it would take too long for the subjects to rate all the images. After analyzing these reasons, two of the authors determined the remaining relevant ones among all the images that were not retrieved and hence not seen by the subjects.

The subjects posed three types of queries:

**Concept Query:** The user thought of a concept (typically expressed by a phrase) and searched for all relevant images. An example concept query is "people eating lunch on campus".

**Specific Query:** The user posed a query to retrieve a particular image. To select these images the user was shown a thumbnail catalog of the image database.

**Memorized Query:** The user chose an image from a thumbnail catalog shortly before the experiment. He was later asked to remember the image and retrieve it.

With a few exceptions, users constructed two queries of each type. The users freely constructed the queries with no instructions. The goal of the experiments is to examine the degrees of usefulness of the techniques individually and in combination.

We use the standard measures *recall* and *precision* [Sal89] to evaluate the effectiveness of our system. All experimental results reported in Section 6 are results averaged over the 32 user queries.

In the first part of our experiments (Subsections 6.1 and 6.2) we have used a custom-built thesaurus constructed as follows: Entity names, relationship names and attribute values from all queries were compared with entity names, relationship names and attributes of the meta-data for corresponding relevant database images. If a pair of words (one from query and one from the database, respecting their types) were actually synonyms or closely related through IS_A hierarchy then the pair was entered into the thesaurus stored in the database.

[4] An employee of UIC photoshop which provides publication related services to many other departments of the university.

## 6 Experimental Results

In the following subsections, precision values are tabulated against recall ranges. The top cell (recall range of 0.1 to 0.2) indicates the proportion of relevant images to total number of retrieved images when only ten to twenty percent of relevant images are retrieved. The last cell (recall range of 0.9 to 1.0) indicates the precision when all of the relevant images are retrieved. A precision value of 1.0 indicates that all of the retrieved images were relevant while a value of 0.5 indicates that only half of them were relevant. A high precision value for all ranges is desirable with the top few ranges being most important.

### 6.1 Plain Retrieval versus Full and Minimal Normalization

Table 6.1 shows that the minimal normalization method is much better than the full normalization method which is in turn better than the plain search (as indicated above, for a given recall range higher precision values denote better retrieval performance). The minimal normalization method has an average precision value of 0.573 while the corresponding values for the full normalization and the plain search are 0.380 and 0.267 over the entire recall range interval. When the user's query is simple, with the plain retrieval option quite a few images have the same similarity. These images which include both relevant and irrelevant ones have the same likelihood of being retrieved. As a result, the precision suffers. The usual normalization method avoids this problem by assigning higher similarities to those images which have fewer content elements (entities, attribute values and relationships) than those having more elements if all these images have the same set of elements in common with the query. One drawback of this method is that since the database descriptions vary from one image to another, images satisfying the query quite well may suffer due to many extra content elements which do not match the query. This situation becomes particularly visible for those images which have detailed descriptions. The new normalization method places less emphasis on the mismatches, yet avoids the problem encountered by plain retrieval. Please refer to Table 6.1 for detailed results.

In tables 6.1 through 6.5 precision values for the recall range 0.0-0.1 were not shown because the number of relevant images for each query is less than 10.

| Recall | pl | fn | mn |
|---|---|---|---|
| 0.1-0.2 | 0.22 | 0.35 | 1.0 |
| 0.2-0.3 | 0.41 | 0.68 | 0.75 |
| 0.3-0.4 | 0.46 | 0.65 | 0.79 |
| 0.4-0.5 | 0.35 | 0.46 | 0.78 |
| 0.5-0.6 | 0.15 | 0.17 | 0.18 |
| 0.6-0.7 | 0.22 | 0.32 | 0.58 |
| 0.7-0.8 | 0.17 | 0.21 | 0.36 |
| 0.8-0.9 | 0.12 | 0.12 | 0.19 |
| 0.9-1.0 | 0.32 | 0.47 | 0.53 |
| Avg. | 0.268 | 0.381 | 0.573 |

**Table 6.1** Comparison of normalization techniques.

**Legend :** pl: plain retrieval (no normalization); fn: full normalization; mn: minimum normalization

### 6.2 Use of Significance

By comparing Table 6.2 with Table 6.1 it is clear that by employing entity significance, improvement in retrieval effectiveness is obtained in each of the three cases of plain

retrieval, full normalization and minimal normalization. In most cases, objects of interest in an image are those which occupy a significant portion of the image. Although a particular user may be interested in small objects, our experiments suggest that the degrees of importance of objects can be approximated by their relative areas. With automatic object recognition, it is easy to assign significance based on this principle. Techniques like "flood-fill" and "active-contours" [NBE+93] facilitate semi-automatic object identification and the area information can readily be obtained. Using size based significance in retrieval is simple and yet effective.

| Recall | s | s,fn | s,mn |
|---|---|---|---|
| 0.1-0.2 | 0.43 | 0.53 | 1.0 |
| 0.2-0.3 | 0.65 | 0.60 | 1.0 |
| 0.3-0.4 | 0.56 | 0.68 | 0.78 |
| 0.4-0.5 | 0.41 | 0.55 | 0.75 |
| 0.5-0.6 | 0.21 | 0.25 | 0.12 |
| 0.6-0.7 | 0.41 | 0.47 | 0.57 |
| 0.7-0.8 | 0.22 | 0.27 | 0.48 |
| 0.8-0.9 | 0.16 | 0.16 | 0.13 |
| 0.9-1.0 | 0.39 | 0.46 | 0.53 |
| Avg. | 0.382 | 0.441 | 0.595 |

**Table 6.2** Results using "object significance"

**Legend :** s: significance; fn: full normalization; mn: minimum normalization

### 6.3 Use of WordNet

Results using WordNet are shown in Tables 6.3 and 6.4. In all experiments minimal normalization and object significance are employed.

#### 6.3.1 Query Expansion

Table 6.3 shows that retrieval performance when WordNet was used for query expansion only.

| Recall | No Th. | 1/syn | 1/A | 1/shh | A/syn | A/A |
|---|---|---|---|---|---|---|
| 0.1-0.2 | 0.58 | 0.58 | 0.58 | 0.58 | 0.58 | 0.58 |
| 0.2-0.3 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| 0.3-0.4 | 0.78 | 0.78 | 0.73 | 0.78 | 0.78 | 0.70 |
| 0.4-0.5 | 0.62 | 0.66 | 0.66 | 0.62 | 0.73 | 0.69 |
| 0.5-0.6 | 0.21 | 0.21 | 0.25 | 0.19 | 0.20 | 0.12 |
| 0.6-0.7 | 0.08 | 0.09 | 0.27 | 0.35 | 0.45 | 0.37 |
| 0.7-0.8 | 0.04 | 0.04 | 0.15 | 0.45 | 0.35 | 0.44 |
| 0.8-0.9 | 0.02 | 0.02 | 0.05 | 0.30 | 0.11 | 0.20 |
| 0.9-1.0 | 0.06 | 0.09 | 0.35 | 0.58 | 0.52 | 0.55 |
| Avg. | 0.376 | 0.385 | 0.448 | 0.538 | 0.524 | 0.516 |

**Table 6.3** WordNet Results with **query expansion** only

**Legend** - No Th: No Thesaurus; 1/syn: first sense, synonyms; 1/A: first sense, all relationships; 1/shh: first sense, synonyms, hypernyms and holonyms; A/syn: All senses, synonyms; A/A: All senses and all relationships (hypernyms, holonyms, meronyms etc).

The second column of Table 6.3 (labeled No Th.) shows the results when neither the custom-built thesaurus nor WordNet was used. The third column shows the results of using the synonyms in the first sense of query words. The fourth column is for using all relationship types in the first sense. The fifth column is for the combination of synonyms, hypernyms and holonyms in the first sense (i.e. meronyms are excluded). The sixth column is for synonyms in all senses of a word. Finally the seventh column is for using all senses and all relationship types in each sense.

By comparing the results under column 3 with those under column 2, we observe that using the synonyms in the first sense of the query words improves retrieval effectiveness only marginally. This is due to the fact that although matches using these synonyms are accurate in general, there are insufficient matches. When more senses or more relationships are used (columns 4,5,6) there are significant improvements. However, when all senses and all relationships are used, undesirable matches begin to cause degradation in performance. Thus, when queries are expanded, using synonyms of all senses (column 6) and using synonyms, hypernyms and holonyms in the first sense (column 5) tend to give better results than using all senses and all relationships (column 7).

Columns 5,6 and 7 of Table 6.3 suggest that using the combination of synonyms, hypernyms and holonyms in the first sense (column 5), using synonyms in all senses (column 6), and using all senses and all relationships (column 7) each improve retrieval significantly when compared to not using any thesaurus. These improvements are results of facilitating new, desired matches between query entities and database entities, which are not possible with synonyms in the first sense alone. The trade-off here is that there may be spurious (unwanted) matches. Indeed, this was the case with some of our queries. But the overall effect was positive. The reason for this is that in our system the words have clear roles (entity name, relationship name, attribute value). Hence, a match between entity names is either supported or unsupported by attribute matches. As an example suppose a "man" matches a "person"(synonym in sense 4 of WordNet). This is a desired match and is likely to be supported by additional matches on attributes such as height, age etc. On the other hand "man" also matches "game-equipment" (synonym in sense 10 of WordNet). In this case, however, the attributes are dissimilar. Hence both images will get an increase in similarity but the desired image will get a much bigger impact. Secondly, if the match obtained via WordNet is a desired one, then this will likely support a relationship match if one is defined in the query. For instance, if a "man" is "holding a baby" in the query, (assuming the baby was already matched) matching the "man" with "person" will enable the relationship match as well.

#### 6.3.2 Database Expansion

As pointed out in Section 4.2 certain matches which are not possible with query expansion alone can be facilitated when both queries and image descriptions in the database are expanded. Table 6.4 shows the results of two such expansions. As indicated in Table 6.3, the use of all relationships in all senses does not give the best performance. Thus this option is not used in getting the results shown in Table 6.4. Retrieval results were improved by expanding the database descriptions as well as the query as can be seen by comparing Table 6.4 to Table 6.3.

| Recall | No Th. | Custom Th. | WN1 | WN2 |
|---|---|---|---|---|
| 0.1-0.2 | 0.58 | 1.0 | 1.0 | 1.0 |
| 0.2-0.3 | 1.0 | 1.0 | 1.0 | 1.0 |
| 0.3-0.4 | 0.78 | 0.78 | 0.78 | 0.78 |
| 0.4-0.5 | 0.62 | 0.75 | 0.61 | 0.63 |
| 0.5-0.6 | 0.21 | 0.12 | 0.21 | 0.19 |
| 0.6-0.7 | 0.08 | 0.57 | 0.41 | 0.41 |
| 0.7-0.8 | 0.04 | 0.48 | 0.45 | 0.45 |
| 0.8-0.9 | 0.02 | 0.13 | 0.29 | 0.29 |
| 0.9-1.0 | 0.06 | 0.53 | 0.51 | 0.52 |
| Avg. | 0.376 | 0.595 | 0.584 | 0.585 |

**Table 6.4** WordNet results with **query and database expansion**

**Legend** - No Th.: Results with no thesaurus. Custom Th: Best results with the custom-built thesaurus (using object significance and minimal normalization). WN1: Expand query with synonyms and hypernyms/holonyms of first sense. Expand database with synonyms, hypernyms and holonyms of first sense. WN2: Expand query with synonyms of all senses and hypernyms/holonyms of first sense. Expand database with synonyms, hypernyms and holonyms of first sense. Both WN1 and WN2 use object significance and minimal normalization.

When implementing our database expansion method, two decisions were made. The first is to limit the relationship types we use. Based on the results shown in Table 6.3, the combination of synonyms, hypernyms and holonyms were selected.

The second is to balance the need for matching the desired terms which are far apart in the in the IS_A and MEMBER_OF hierarchies with the need for high precision. In order to address this issue we first applied a distance threshold (of 2, determined empirically) on both query and database expansion. Observing that some desired matches were missed, we then lifted the threshold. Although this improved the recall, the overall results degraded due to undesired matches. Finally, the method of entity category verification was applied to avoid spurious matches while preserving desired matches. With category verification, we obtained better results than using a low threshold or no threshold at all. In all of the above cases, the distance between the words that are matched through the hierarchy was incorporated into the similarity formula to give less weight to distant words.

For query expansion, in one run we use synonyms, hypernyms and holonyms in the first sense and in the other run synonyms in all the senses and synonyms, hypernyms, holonyms in the first sense. It is found that these two runs have comparable performance and significantly better than the performance obtained by query expansion alone.

### 6.3.3 Category Verification

The results of using category verification is compared with the best previous results in Table 6.5. For easy reference, the previous recall-precision values are repeated here. The first result column (labelled No Th.) is the result with no thesaurus used. The second column (labelled Custom Thes.) is the best result with the custom-built thesaurus. The column labeled WN2 shows the best result with WordNet using distance thresholds and no categroy verification. The next column shows the effect of lifting the threshold. Due to spurious matches between undesired entities, the overall results are significantly worse than those with using a threshold. The results of applying category verification is shown in the last column (labeled WN4). Note that the improvement in precision as compared to the best previous results with WordNet is achieved at low recall intervals, which is a desired property. This shows that using the category verification process provides a fine balance between recall and precision and hence improve overall effectiveness.

| Recall | No Th. | Custom Th. | WN2 | WN3 | WN4 |
|---|---|---|---|---|---|
| 0.1-0.2 | 0.58 | 1.0 | 1.0 | 0.67 | 1.0 |
| 0.2-0.3 | 1.0 | 1.0 | 1.0 | 0.85 | 1.0 |
| 0.3-0.4 | 0.78 | 0.78 | 0.78 | 0.64 | 0.82 |
| 0.4-0.5 | 0.62 | 0.75 | 0.61 | 0.61 | 0.71 |
| 0.5-0.6 | 0.21 | 0.12 | 0.21 | 0.20 | 0.23 |
| 0.6-0.7 | 0.08 | 0.57 | 0.41 | 0.34 | 0.50 |
| 0.7-0.8 | 0.04 | 0.48 | 0.45 | 0.37 | 0.45 |
| 0.8-0.9 | 0.02 | 0.13 | 0.29 | 0.22 | 0.13 |
| 0.9-1.0 | 0.06 | 0.53 | 0.52 | 0.45 | 0.56 |
| Avg. | 0.376 | 0.595 | 0.585 | 0.483 | 0.600 |

**Table 6.5** Comparison of WordNet results without and with **category verification**

**Legend** - WN2: Best WordNet result with distance threshold and without category verification (multiple senses, query and database expansion). WN3: WordNet results with no distance threshold and without category verification. WN4: WordNet results with no distance threshold and with category verification.

Comparing columns for WordNet results and those for custom-built thesaurus in Table 6.5 reveals that proper use of WordNet may bring retrieval results very close to the level of a manually constructed thesaurus. It should be noted that there are several other combinations of WordNet options which we have not tested. Some of these combinations may yield better retrieval performances than the ones reported here.

## 6.4 Discussion and Review of Related Work in Uses of WordNet

Ellen Voorhees [Voo93] used the context of queries/documents and the IS_A hierarchies in the noun portion of WordNet to automatically disambiguate word senses in **text** retrieval The result of sense disambiguation is a sense identifier for each query word. The same procedure is applied to the database terms. The expanded query consists of both query words and sense identifiers. These are matched with those of the database. Although sense disambiguation helped in certain queries, the overall results indicated that unexpanded queries containing only query terms were superior.

In a study involving retrieval of image captions [SQ96], WordNet was used to compute semantic distances between words. Instead of matching words of query with those of the database regardless of their intended meaning, the authors computed word-to-word distances using a concept hierarchy derived from WordNet's IS_A hierarchy. The words in the queries and in the caption database were *manually* disambiguated by the authors [SQ96]. By using word-to-word similarities, the total similarity of a query to an image caption was determined. Based on this general idea, different similarity formulas and thresholds were studied. An improvement of over 40 % in retrieval effectiveness for certain combinations of their parameters are reported.

The key difference between our system and the systems described in [Voo93, SQ96] are three-fold:

**The use of E-R model:** The structure of our queries and image meta-data in the database follow the object-attribute-relationship (E-R model) and therefore reduces the negative effect of spurious matches between undesired pairs of words (see subsection 6.3.1). The trade-off is the query and meta-data construction time. Our average meta-data generation time of 2 minutes per image however, compares favorably with the average of 2.5 minutes for a person to write a caption reported in [SQ96].

**Classification of Terms** We have used WordNet noun classes for verifying potential matches between query and database terms. In [SQ96], a concept hierarchy derived from the WordNet IS_A hierarchy was used for disambiguating the word senses while Voorhees [Voo93] defined and used the term *hood* for this purpose.

**Results of fully automatic use of WordNet:** In [SQ96] sense disambiguation was done *manually*. In [Voo93] automatic sense disambiguation was unsuccessful due to insufficient context provided by short queries. In that work, a mismatch occurred in three cases: (a) The disambiguation algorithm resolved the meaning to two different senses in the query and the document even though both were the same sense, (b) the algorithm was not able to resolve a sense at all in either the query or the document, and (c) the words had different roles such as verb v.s. noun. It was observed that while sense resolution worked well for document words, it either failed or produced incorrect senses in case of short queries. In our experiments, we have observed significant improvement with specific, *fully automatic* uses of WordNet. We have tried first and all senses of words, incorporated a distance based weighting scheme and category verification. Analysis indicated that the main reasons for the success were the following: 1. Since our expansion scheme never excluded exact matches, no matches were missed due to insufficient context or incorrect disambiguation. 2. For both queries and database words, the first sense was indeed the correct sense for more than 90% of the words. Hence using first sense alone for expansion was satisfactory for most cases. 3. When we used all senses of a word, correct matches had an advantage due to positive contribution from matching attributes and relationships.

Different normalization formulas were proposed and tested for text and image databases [SBM96, BPJ93, Sal89]. In [SBM96] two new normalization techniques based on the document length and relevance were shown to be superior to cosine normalization in text retrieval. The normalization applied in [BPJ93] is based on numerical values of object (face) features such as length of the mouth. Our normalization formula on the other hand is based on object similarities and hence not directly comparable to methods presented in these works.

## 7 Conclusion

Recent developments in automatic extraction of abstract contents opens a new avenue for content based image retrieval. In this paper we have presented techniques for improving retrieval effectiveness based on semantic contents of images. These include a new normalization scheme, object significance and selective uses of WordNet for query and database expansion. We have conducted experiments with real users to demonstrate the effectiveness of these methods individually and in combination. With a general data model involving objects, their attributes and relationships, these techniques are applicable in a variety of domains.

Our experiments with WordNet indicated that specific uses of an electronic thesaurus can provide significant improvement over not using any thesaurus. In fact, our preliminary results show that a proper use of WordNet yields a performance which approximates that of a custom-built thesaurus. Although WordNet was used for query expansion in text retrieval with varying degrees of success, our modelling of image contents via objects and attributes facilitate more effective *automatic* query and database expansion mechanisms. Entity category verification using WordNet noun categories help increase recall while avoiding undesired matches.

Future work includes using WordNet adjective antonyms for conflict elimination, investigation of better uses of WordNet involving other combinations of search options, associating "concepts" with images via automatic procedures and experimentation with other image collections.


### Acknowledgements

We would like to thank Ellen Voorhees for her explanation of sense disambiguation and query expansion schemes. This research was supported by the following organizations: NSF grant under IRI-95 09253, NASA under NAGW-4080 and by ARO under BMDO grant DAAH04-0024.


### References


[AHKR96] P. Alshuth, Th. Hermes, J. Kreyb, and M. Roper. Video Retrieval with IRIS. In *Proceedings of ACM Multimedia Conference, Boston MA*, page 421, 1996.

[ATY+95] Y. A. Aslandogan, C. Thier, T. C. Yu, C. Liu, and K. Nair. Design, implementation and evaluation of SCORE (a System for COntent based REtrieval of Pictures). In *IEEE-ICDE-11*, March 1995.

[ATY96] Y. Alp Aslandogan, Chuck Thier, and Clement T. Yu. A System for Effective Content Based Image Retrieval (Prototype Demonstration). In *Proceedings of ACM Multimedia Conference, Boston MA*, pages 429–430, 1996.

[BP94] M. Bischel and A. Pentland. Human Face Recognition and Face Image Set's Topology. *CVGIP: Image Understanding*, 59(2):54–261, March 1994.

[BPJ93] Bach, J. R., Paul, S., and Jain, R. A Visual Information Management System for the Interactive Retrieval of Faces. *IEEE-TOKDE*, 5(4):619–628, August 1993.

[CJ91] Chang, S.K. and Jungert, E. Pictorial data management based upon the theory of symbolic projections. *Journal of Visual Languages and Computation*, 2:195–215, 1991.

[Cou96] Jonathan D. Courtney. Automatic, Object-Based Indexing for Assisted Analysis of Video Data. In *Proceedings of ACM Multimedia Conference, Boston MA*, pages 423–424, 1996.

[CPLJ94] Chua, T., Pung, H., Lu, G., and Jong, H. A Concept Based Image Retrieval System. In *IEEE Int'nl Conf. on system Sciences*, pages 590 598, January 1994.

[FBF+94] Faloutsos C., Barber R., Flickner M., Hafner J., Niblack W., Petkovic D., and Equitz W. Efficient and Effective Querying by Image Content. *Journal of Intelligent Information Systems*, 3(1):231–262, 1994.

[Gup95] Gupta, A. Visual Information Retrieval Technology, A VIRAGE Perspective. White paper, Virage Inc., 1995.

[GZCS94] Gong, Y., Zhang H., Chuan, H.C., and Sakauchi M. An Image Database System with Content Capturing and Fast Image Indexing Capabilities. In *Proceedings of Intl. Conf. on Multimedia Computing Systems*, pages 121–128, May 1994.

[HCK90] Halin, G., Crehange, M., and Kerekes, P. Machine Learning and Vectorial Matching for an Image Retrieval Model: EXPRIM and the system RIVAGE. In *Proceedings of ACM-SIGIR 1990, Brussles, Belgium*, pages 99–114, 1990.

[HHTK96] K. Hirata, Y. Hara, H. Takano, and S. Kawasaki. Content-Oriented Integration in Hypermedia Systems. In *Proceedings of ACM Conference on Hypertext*, 1996.

[JG95] Jung, G.S. and Gudivada, V. Adaptive Query reformulation in Attribute based Image Retrieval. In *Intelligent Systems*, pages 763–774, 1995.

[KdV94] Y. Kwon and L.N. da Vitoria. Age Classification form Facial Images. In *Proceedings of CVPR, Seattle.*, pages 762–767, 1994.

[KKOH92] Kato, T., Kurita, T., Otsu, N., and Hirata, K. A Sketch Retrieval Method for Full Color Image Database, Query by Visual Example. In *IEEE-IAPR-11*, pages 530–533, August-September 1992.

[LC96] W.S. Li and K. Candan. SEMCOG: Integration of SEMantics and COGnition Based Approaches for Image Retrieval. Technical report, NEC, 1996.

[Lie96] Rainer Lienhart. Indexing and Retrieval of Digital Video Sequences based on Automatic Text Recognition. In *Proceedings of ACM Multimedia Conference, Boston MA*, pages 11–20, 1996.

[LW93] Lum, V.Y. and Wong, K. A Model and Technique for Approximate Match of Natural Language Queries . 1993.

[MBF+90] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[NBE+93] Niblack W., Barber R., Equitz W., Flickner M., Glasman E., Petkovic D., Yanker P., Faloutsos C., and Taubin G. The QBIC Project: Querying Images By Content Using Color, Texture, and Shape. In *SPIE Storage and Retrieval for Image and Video Databases*, pages 173–187, February 1993.

[OS95] Ogle, V. E. and Stonebraker, M. Chabot: Retrieval from a Relational database of Images. *IEEE Computer*, 28(9), 1995.

[RS95] Rohini and Srihari. Automatic Indexing and Content Based retrieval of Captioned Images. *IEEE Computer*, 28(9), September 1995.

[Sal89] Salton, G. *Automatic Text Processing*. Addison Wesley, Mass., 1989.

[SBM96] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted Document Length Normalization. In *Proceedings of ACM SIGIR 96*, pages 21–29, August 1996.

[SC96] John R. Smith and Shih-Fu Chang. VisualSEEk: a fully automated content-based image query system. In *Proceedings of ACM Multimedia Conference, Boston MA*, pages 87–98, 1996.

[SI92] A. Samal and P. A. Iyengar. Automatic Recognition and Analysis of Human Faces and Facial Expressions. *Pattern Recognition*, 25(1):65–77, February 1992.

[SQ96] Alan F. Smeaton and Ian Qigley. Experiments on Using Semantic Distances Between Words in Image Caption Retrieval. In *Proceedings of ACM SIGIR Conference*, 1996.

[SYH94] Sistla, P.A., Yu, C., and Haddad, A. Reasoning about Spatial Relationships in Picture Retrieval Systems. In *Proceedings of the 20th VLDB Conference, Santiago, Chile*, pages 570–581, 1994.

[Voo93] Ellen M. Voorhees. Using WordNet to Disambiguate Word Senses for Text Retrieval. In *Proceedings of ACM SIGIR Conference*, pages 171–180, 1993.

[WKSS96] Howard D. Wactlar, Takeo Kanade, Michael A. Smith, and Scott M. Stevens. Intelligent Access to Digital Video: Informedia Project. *IEEE Computer*, pages 46–52, May 1996.

[WNM+95] J.K. Wu, A. Desai Narasimhalu, B.M. Mehtre, C.P. Lam, and Y.J.Gao. CORE: a content-based retrieval engine for multimedia information systems. *Multimedia Systems*, 3:25–41, 1995.

[YD94] Yasser Yacoob and Larry Davis. Recognizing Facial Expression by Spatio-Temporal Analysis. In *Proceedings of IEEE ICPR*, pages 747–749, 1994.

[ZTSG95] H.J. Zhang, S. Y. Tan, S. W. Smoliar, and Y. Gong. Automatic Parsing and Indexing of News Video. *Multimedia Systems*, 2:256–266, 1995.