AN APPROACH TO SEMANTICS-BASED IMAGE RETRIEVAL AND BROWSING*

Jun Yang[1], Liu Wenyin[2], Hongjiang Zhang[2], Yueting Zhuang[1]

[1]Department of Computer Science
Zhejiang University
Hangzhou, 310027, China
i_jyang@yahoo.com, yzhuang@cs.zju.edu.cn

[2]Microsoft Research China
49 Zhichun Road
Beijing, 100080, China
{wyliu, hjzhang}@microsoft.com

ABSTRACT

The state-of-the-art image retrieval approach is to incorporate image semantics with low-level features to enhance retrieval performance. Although many systems annotate images with descriptive keywords and retrieve images by keyword-based search, they have not explored the full potentials of semantics. This paper exploits the power of thesaurus-aided approaches to facilitate semantics-based access to image collections. The contribution of our work includes constructing a dynamic semantic hierarchy (DSH) to support flexible browsing of images by semantic subjects, as well as formulating a semantic similarity metric to improve the accuracy of semantic matching. Both approaches are seamlessly integrated into a unified framework for semantics- and feature-based image retrieval. Experiments conducted on real-world images demonstrate the effectiveness of our approaches.

1. INTRODUCTION

With the explosive growth of digital images, effective and efficient access to image database has recently gained much research interest. In this filed, content-based image retrieval (CBIR) is devised to search images based on the similarity of low-level features, such as color and texture [1][2]. The current CBIR techniques are still at very low performance level, because the semantic content of the image are not necessarily captured by its low-level features. Keyword-based image retrieval is therefore more preferable, since it matches the query keywords against the image annotations directly at the semantic layer. However, this approach entails a tedious and labor-intensive process of manual annotation, which may also introduce errors due to the human subjectivity. As a current trend among many state-of-the-art image retrieval systems, image semantics (represented in keywords) is often combined with low-level features within a unified framework, allowing them to benefit each other to yield better performance and efficiency.

At the semantic layer, a variety of measures have been proposed to evaluate semantic similarity between images. Lu et al. [3] use the number of keywords in common between query and the image annotation as the similarity degree. Paek et al. [4] present a novel TF*IDF vector-based approach to match accompanying text of photographs in order to classify them. Although proven to be effective, both approaches overlook the similarity between different words and thus are vulnerable to the problems posed by the richness of natural language, such as synonyms, polysemy and other complex word relevancy. In comparison, the thesaurus-based keyword similarity measure proposed by Smeaton et al. [6] is a good way to solve this problem. However, their approach ignores the visual aspect of images. Moreover, as another important functionality required in image database systems, the capability of convenient and flexible image browsing is not well supported.

To address the limitations of the current image retrieval systems, we investigate the power of thesaurus-aided approaches to facilitate image retrieval and browsing. In particular, we utilize an electronic thesaurus WordNet™ [7] to interactively construct a *dynamic semantic hierarchy* (DSH) to support flexible image browsing. Furthermore, a novel *semantic similarity metric* is formulated based on WordNet to improve the accuracy of semantic matching. Both methods are seamlessly integrated into our former image retrieval system, *iFind*, to enhance its performance and facility.

The rest of this paper is organized as follows. In Section 2, we briefly review the image retrieval system

* A short version of this paper will be presented in ICME01, August 2001, Japan.

iFind. Section 3 presents the details of our approaches. The implementation issues discussed in Section 4 demonstrate how the proposed approaches are integrated into *iFind*. Experimental evaluation is presented in Section 5. Finally, concluding remarks are given in Section 6.

2. REVIEW OF OUR SYSTEM

We have developed a prototype system *iFind* for image retrieval, which can accommodate both image semantics and visual features. The unique features that *iFind* had implemented includes a semi-automatic image annotation strategy [9] and a unified framework for semantics- and feature-based image retrieval and feedback [3].

2.1 Semi-Automatic Annotation Strategy

Firstly a semantic network is constructed for the images in the database. In this network each image is linked to a set of keywords relevant to its content. A weight is assigned to each link to show the descriptive power of the corresponding keyword to the image. As such image-keyword links may not be available initially, they can be obtained interactively, either by manual annotation or more preferably by the semi-automatic annotation strategy during relevance feedback. Whenever the user provides a set of relevant/irrelevant images to the input query keywords, an underlying voting scheme is triggered to propagate or update the image annotation: 1) For each relevant image, if it has not been annotated with the query keyword, create a link between them with an initial weight (e.g., 1.0 in our implementation). Otherwise the keyword weight is increased by a given increment (e.g., 1.0). 2) For each irrelevant image, if the query keyword has been linked to it, decrease its weight by some degree (e.g., one fourth). 3) Check all the links and delete those with weights below a given threshold (e.g., 1.0). In this way, the keywords are propagated and updated in a hidden manner in the course of user interaction, and hence improves both the coverage and quality of the annotation among the images. Performance evaluation shows that this strategy outperforms manual annotation in terms of efficiency and automatic annotation in terms of accuracy. See [9] for detailed description and user study result of the semi-automatic annotation strategy.

2.2 Unified Framework for Semantics- and Feature-Based Retrieval

We then implemented a unified framework in *iFind*, under which semantics and visual features are seamlessly incorporated to benefit each other to yield better performance. Each image in the database is indexed by its visual features of color and texture, as well as by the keyword annotation obtained semi-automatically using the strategy described above. In the retrieval phase, the similarity of each image to the query is calculated by combing the visual feature similarity with the semantic similarity. The relevance feedback process is performed respectively at the semantic level and the feature level in a parallel manner. The query is then reevaluated based on the updated features and semantics to improve the retrieval results. Experimental results manifest that, by using this framework higher retrieval accuracy is achievable with less iterations of feedback than using traditional relevance feedback methods [10].

However, there are also severe limitations with *iFind*. On the one hand, it lacks the support to image browsing/navigation functionality, which requires the images to be explicitly organized by their semantic subjects. Unfortunately, in *iFind* the images are implicitly annotated with keywords and therefore remain unorganized no matter how much annotation is available. Meanwhile, a number of predefined categories currently available in *iFind* are too rough and inflexible to facilitate convenient browsing. Because image browsing is among the most popular user behaviors, it is of great importance to be integrated into the system. On the other hand, semantic similarity is measured as the number of keywords in common between the query and image annotation. This measure ignores the similarity between different words, which is no doubt a flaw when considering the users' different preferences on use of keywords, as pointed out by Bates [12]: "the probability of two persons using the same term in describing the same thing is less than 20%".

Hence, we propose two thesaurus-aided approaches to solve the aforementioned problems. The *dynamic semantic hierarchy* (DSH) is developed to facilitate flexible and convenient browsing. The *semantic similarity metric* is formulated as a delicate estimation of semantic similarity.

3. THE PROPOSED APPROACHES

WordNet [7][8] is an electronic thesaurus that models the lexical knowledge of English language. Information in WordNet is organized around logical groupings called synsets, each of which consists of a set of closely related synonyms representing the same word sense. There are also various types of semantic links among synsets, which constitute a highly interconnected network of synsets in WordNet. Particularly, in the noun portion of WordNet, semantic links can be of the type of Hyponym/Hypernym (IS_A relationship) or Meronym/Holonym (Part-of or Member-of relationship).

For the sake of simplicity, we apply the following restrictions when using WordNet, i.e., 1) we use only the noun portion of WordNet, since nouns are more intensively used to describe images than other classes of words; 2) for the word with multiple word senses (i.e. polysemy), take the first sense as the user-intended sense, so that we can use the term synset and word interchangeably in this paper because each word is mapped to exact one synset; 3) among

various semantic links between noun synsets, we only use Hyponym/Hypernym relationships. (Hypernym is a generic concept of hyponym, while hyponym is a specific concept of hypernym, e.g. *tree* is the hypernym of *oak*, while *oak* is a hyponym of *tree*.) The above simplifications reduce WordNet to a number of semantic hierarchies of synsets, each of which goes from generic concept at higher levels to specific concept at lower levels. Each hierarchy is a well-defined tree structure without loop in it. The root concepts of these hierarchies include *action*, *living form*, *object*, *place*, *event*, *phenomenon*, *group*, *possession* and *condition*, which are defined as the beginning concepts of some generic domain in WordNet. These hierarchies have a broad coverage of all the nouns and no overlap between each other — no noun can belong to more than one hierarchy. We refer to them as WordNet hierarchies, a fragment of which is exemplified in Figure 2.

3.1 Construction of the Dynamic Semantic Hierarchy (DSH)

As described above, the *iFind* system has applied the semi-automatic annotation strategy, so that the annotating keywords of images can be obtained and updated interactively. As keywords are readily available, our key idea is to construct hierarchical categories from all these keywords to support image browsing. Although the WordNet itself provides well-structured and comprehensive semantic hierarchies, it is not applicable to image browsing given its huge size (WordNet noun portion contains approximately 48,800 synsets). Therefore, we have devised the dynamic semantic hierarchy (DSH) as a set of sub-hierarchies of WordNet hierarchies, which can be expanded interactively and progressively from a few predefined root concepts. This compact DSH offers greater convenience, flexibility and efficiency for image navigation and browsing than the original WordNet hierarchies.

As there is no keyword initially available in *iFind*, DSH is an empty hierarchy with only the root concepts described above available. A virtual root "everything" is created and all the root concepts are inserted as its offspring. DSH will be expanded each time a new keyword is identified by the system, such as annotating images with new keyword, or marking images as relevant to the query with new keyword. These operations will cause the synset of the new keyword to be inserted into the proper position in DSH. We also check to see whether the "lowest" common ancestor of this new keyword and any DSH node exists in DSH. If not, this ancestor node will be inserted. The insertion operation will guarantee the structure of DSH to be in full conformity with WordNet hierarchies. The detailed algorithm of inserting a new keyword *w* into DSH is given in Figure 1. Since a vast amount of nouns in WordNet are very infrequently occurring words that are unlikely to be used to describe images, they should not appear in DSH. An infrequent noun is identified if its frequency calculated from a huge English text corpus is below a given threshold. We preclude such nouns before building DSH, so that the algorithm in Figure 1 only concerns those frequent nouns.

Step 1: *Find the corresponding synset S_n of w in WordNet.*

Step 2: *Start from S_n , trace bottom- up along the links in WordNet hierarchy, until the first ancestor synset of S_n that has already existed in DSH is reached. This ancestor synset is denoted as S_a.*

Step 3: *Set { S_1, S_2 ,… , S_M } to be the direct children of S_a in DSH, where M is the number of children.*
For i=1 to M
Find the lowest common ancestor synset S_{co_a} of both S_i and S_n in WordNet hierarchy.
If S_{co_a} <> S_a , go to Step 5.
End For

Step 4: *Insert S_n into DSH as a direct child of S_a.*
Exit the algorithm.

Step 5: *Insert S_{co_a} into DSH as a child of S_a.*
Remove S_i as the child of S_a and then insert it as a child of S_{co_a}.
If S_n <> S_{co_a}, insert S_n into DSH as a child of S_{co_a}.
Exit the algorithm.

Figure 1: Insertion algorithm for DSH.

Insertion operations using this algorithm are exemplified in Figure 2. The hierarchy shown in Figure 2 (a) is a tiny fragment taken from the original WordNet hierarchy, with the infrequent nouns shown in dashed rectangles. Figure 2 (b) gives the initial look of DSH with only the root concept *living form* available. Then a set of keywords is inserted into DSH under *living form* in the order of *parrot*, *horse*, *hen* and *mammal*. The resultant DSH after each insertion is displayed in Figure 2 (c)~(f), respectively, with the keywords inserted at each step shown in bold. As can be easily seen, DSH is semantically well-structured and relatively compact compared with the original WordNet hierarchy.

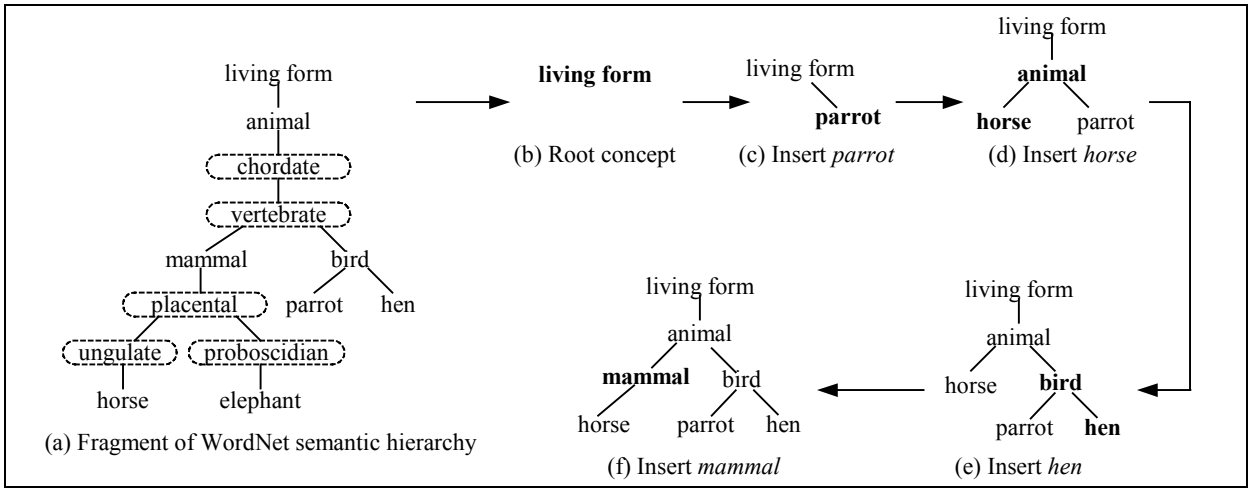

Figure 2: Example of insertion operations.

We take the insertion of *horse* (see (d)) as an example to demonstrate how our algorithm works. Firstly, *horse* is mapped to the synset S_n (step 1) and its first existing ancestor (in DSH) S_a is located in the WordNet hierarchy, which is *living form* in this case (step 2). As *parrot* is the only child of *living form* in DSH, we find its common ancestor S_{co_a} with *horse* (step 3), which is *animal*. Please note that *chordate* and *vertebrate* are ignored by the algorithm as infrequent nouns. Since S_{co_a} is not equal to S_a, we jump from step 3 and to step 5. Finally in step 5 we insert *animal* below *living form* and attach both *parrot* and *horse* as its children.

A likely concern of this algorithm is the loop in step 3, which will be stopped as soon as the first common ancestor of S_i and S_n that is not equal to S_a is found. One may argue that there can be more than one such common ancestor and in that case the algorithm fails to structure DSH correctly. However, this cannot actually happen because our algorithm guarantees that S_a is always the lowest common ancestor of any two of its children and each node has only one parent.

3.2 Semantic Similarity Metric

As stated above, *iFind* matches query with image annotation by counting the number of keywords in common, which is referred to as exact keyword match scheme. This scheme ignores the semantic similarity between different words and consequently fails to address the following particular issues:

- Unable to match closely related synonyms, e.g., a query of *soccer* cannot match the images labeled with *football*.
- Unable to match generic concept with its specific concepts, e.g., the query of *sports* is unable to capture the *basketball* or *football* images.
- Unable to return promising candidates in case of no exact keyword match. For instance, if the query is *football*, *iFind* will return a random list of images if no image is annotated with *football*. However, it is more reasonable to put *sports* images (if there is any) in top ranks since they are semantically closer to the query than random selected images.

Many efforts have been made to utilize thesaurus-based semantic similarity measure in traditional Information Retrieval (IR) research and Smeaton [6] concluded that these approaches worked best when both query and document are short. Since the query and image annotations are usually brief, we rely on WordNet to define a quantitative semantic similarity metric and use it for image retrieval. This metric is defined in the following two steps:

3.2.1 Measuring Word-Word Similarity

In the first step, we define the word-word similarity. Since the WordNet cover all the nouns and accommodates rich semantic links among them, it provides a thorough and domain-independent knowledge base for semantic distance estimation. The word-word similarity is transformed to the similarity between their corresponding synsets. The similarity between synset s_1 and s_2 in the same WordNet hierarchy is determined by the depth of their lowest common ancestor synset s_a, i.e., the number of links from root to it in the hierarchy. The similarity is then normalized by dividing the maximum depth possible in the WordNet hierarchies and thus results in [0, 1], given as:

$$Sim(s_1, s_2) = \frac{depth(s_a)}{depth_{\max}} \quad (1)$$

Two particular situations have to be addressed: First, if s_1 and s_2 is the same synset, their similarity is set to one; Second, if they belongs to different hierarchies, their

similarity is set to zero since they are too far away from each other to have any semantic relevancy.

This semantic similarity metric can be explained by the observation of WordNet hierarchy as an inheritance system, in which the properties of an ancestor are inherited by all of its descendants. In addition, the hierarchy goes from generic concept at higher level to specific concept at lower level. Therefore, the lowest common ancestor of two synsets represents their common property and its depth implies how specific such properties are. The more specific (thus more informative) their common properties are, the more similar the two synsets are, and vice versa. For instance, in Figure 2 (a), determined by the depth of their common ancestor, *horse* is more similar to *elephant* than to *parrot*. This makes sense intuitively because both *horse* and *elephant* are *mammals*, whereas the property *horse* and *parrot* have in common is that they are both *animals*.

Aslandogan [5] proposed to measure distance between synsets as the number of links along the shortest path connecting them in WordNet. This approach can be arguable since it presumed that all the links indicate the same semantic distance. The similarity measure suggested by Smeaton [6] is in essence very close to ours, which approximate synsets similarity using the "information content" of their common ancestor synset. Compared with simply using the depth of common ancestor in our approach, information content theoretically gives more precise estimation of conceptual similarity but is also computationally much more expensive.

3.2.2 Measuring Query- Annotation Similarity

The word-word similarity is then utilized to measure the similarity between query and image annotation. In *iFind*, both the image annotation and the query are represented by weighted keyword sets. The image annotation is expressed as $A=\{< a_1, w_{a_1} >, \cdots, < a_m, w_{a_m} >\}$, where a_i is the keyword (synset) in image annotation with w_{a_i} being its weight. Similarly, the query is denoted as $Q=\{< q_1, w_{q_1} >, \cdots, < q_n, w_{q_n} >\}$. All the keywords in the user-submitted query will have the same weight initially (e.g. one), but can be set differently during reevaluation of query in relevance feedback, as shown later.

Here we adopt the algorithm suggested by Smeaton et al. [6] to extend the similarity metric between two single words to that between two word sets, as follows:

$$Sim(Q,A)=\frac{\sum_{i=1}^{n}\max_{j=1\cdots m}\left[Sim(q_i,a_j)\cdot w_{q_i}\cdot w_{a_j}\cdot\delta(q_i,a_j)\right]}{n}$$

$$\text{where}\quad \delta(q_i,a_j)=\begin{cases}1, & \text{if } a_j \text{ is a descendent of } q_i\\ 0.5, & \text{otherwise.}\end{cases} \qquad (2)$$

where $Sim(q_i, a_j)$ is the similarity between q_i and a_j, calculated using the similarity metric defined in (1). This approach finds the best-matching keyword in the image annotation for each query keyword and computes the average of these maximum similarities as the final similarity between the query and the image. Incorporation of keyword weight w_{qi} and w_{aj} in (2) is very intuitive, since the keyword with higher weight has greater descriptive power and should contribute more to the similarity metric than other words. Please note that each weight is normalized into [0,1] by Gaussian normalization. Therefore, the similarity given by (2) will also result in values between [0,1].

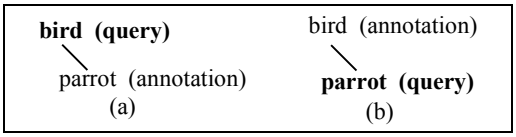

Figure 3: Example of anomaly.

The last term $\delta(q_i, a_j)$ is a weighting factor addressing the relationship between query keyword q_i and annotation keyword a_j. Since the word-word similarity defined in (1) does not distinguish between query and annotation keyword, it evaluates the two queries shown in Figure 3 to be of the same similarity. However, this causes an anomaly in a retrieval context: While it is reasonable to retrieve *parrot* image for query of *bird* (see (a)), the *bird* image does not necessarily satisfy query of *parrot*. $\delta(q_i, a_j)$ is added to attack this problem by emphasizing the case when the query keyword is a generic concept over the annotation keyword, but not the opposite case.

4. IMPLEMENTATION ISSUES

So far we have presented the details of constructing *dynamic semantic hierarchy* (DSH) and defining *semantic similarity metric*. In this section, we demonstrate how they are integrated to *iFind* system to improve its performance and facility.

4.1 DSH-driven Browsing Tool

The main interface of the updated *iFind* system is shown in Figure 4. The left-bottom part is the query interface in which user can input keywords or submit example images as the query. The retrieved images are shown in the

scrollable image browser on the right side as a ranked list of thumbnails. The user can also indicate positive/negative examples by clicking the hand-like icons at the lower part of each image. Our proposed DSH is visualized by the tree control at top-left pane of the interface. As the rendering tool, this tree control keeps in conformity with the DSH. At first it contains only items denoting the root concepts of the DSH. Whenever a keyword is inserted into the DSH, a corresponding item will be created in the same position of the tree control, representing a new category. All the images annotated with that keyword will be put under this new category. By clicking the symbol "+" and "-" left to each item, user can expand/close the item to display/hide all its children categories. With the help of such hierarchical structure, the user can recursively trace down the hierarchy to find out the category to his/her interest and then browse all the images in the category by double-clicking the item. Since this browsing structure is semantically well-structured, the users is able to quickly locate the intended category in a naturally step-by-step manner.

As we can see, such DSH-driven hierarchical categories provide a more convenient and flexible means for image browsing, compared with the browsing structure using predefined categories. It is capable of learning new keywords and adding them dynamically as a category into the appropriate position in the DSH. In addition, it is carefully tailored to include only the used keywords and is therefore very compact and efficient. The weakness of our approach lies in its vulnerability to some puzzling organizations of words in WordNet. For example, the word *man* is conceptually a kind of *people* and should be inserted as its sub-category, whereas in WordNet they belong to different hierarchies. As the result, the DSH and in turn the browsing tool will have this confusing structure, which may bring inconvenience to users.

Figure 4 Main user interface of the updated *iFind*

4.2 Integration of Semantics with Visual Features

The semantic similarity metric is integrated into *iFind* by replacing the previously used exact keyword match scheme. The working flow of the updated *iFind* consists of three phases as the query phase, the feedback phase and the refinement phase, each of which employs an integration of semantics and visual feature to retrieve images. In the query phase, the system accepts a set of keywords and matches them with the annotation of each image using the semantic similarity metric defined in (2). The semantically matched images will be retrieved if there are any, otherwise a random list of images is returned. If the semantic matches are less than a certain number, the system will take the top match as the example image and trigger a content-based retrieval to find images similar to it in terms of visual features.

After the first batch of retrieved images are displayed in the browser pane, the system moves to the feedback phase, during which the user is allowed to mark the retrieved images as either relevant or irrelevant ones. Upon the acceptance of these feedbacks, both the image visual feature representations and the semantic representations are updated in a parallel manner. At the feature level, the traditional re-weighting method suggested by Rui and

Huang [10] is adopted to adjust the weight of each visual feature. At the semantic level, the semi-automatic annotation strategy [9] is utilized to propagate and update the keywords among the images. Finally in the refinement phase, the improved retrieval results will be calculated as follows: Collect all distinct keywords that ever appeared in the annotations of all the positive feedback images to compose pseudo-query Q_p as a weighted keyword set, with the keyword weight set to the numbers of its occurrence in such annotations. A similar pseudo-query Q_n is constructed from negative feedback images. The final similarity S_i of the ith image in the database to the query is calculated based on the updated image visual features and semantics using a modified form of Rocchio's formula [11].

$$S_i = sim(Q, A_i) + \alpha(1 + sim(Q_p, A_i)) \sum_{k \in Np} S_{ik} - \beta(1 + sim(Q_n, A_i)) \sum_{k \in Nn} S_{ik} \quad (3)$$

where N_p and N_n are the total number of positive and negative feedbacks respectively, S_{ik} is the visual similarity calculated from the Euclidean distance of low-level features between ith image and kth feedback example. The other two parameters α and β are constants reflecting the contribution of the positive/negative feedbacks, which are assigned to 1.0 in the current implementation of the system for the sake of simplicity. Obviously, this formula provides a comprehensive similarity metric that addresses both the semantic aspect and the visual feature aspect of images.

5. PERFORMANCE EVALUATION

To show the effectiveness of our approach described in Section 4.2, we conduct some experiments on the ground-truth image database, which is constructed using images from the Corel Image Gallery. We select totally 5,000 images, which are classified into 50 categories with exactly 100 images in each category. Images within the same category are regarded as relevant to each other and can be described by one keyword, which is exactly the category name. Thus, if the category name is used as query keyword, all the images within this category are expected to be retrieved by the system.

We devise a semi-automatic experimentation method to avoid the tedious manual efforts of submitting queries and feedbacks. The process starts by using one category name as the query to match the annotation of each image in the database, using the semantic similarity metric defined in (2). The first 100 images ranked top in their similarity to the query are returned. Among these 100 images, the system automatically marks those belonging to the intended category as positive examples and the rest as negative ones. By performing feedback based on these assumed feedback examples, the system improves the retrieval results by bringing out more relevant images. The same process is repeated in the further iterations of feedbacks and the statistics (hit and miss) at each iteration are recorded.

Figure 5 shows the performance of the system on 8 random selected queries, given that no images are initially annotated with any keywords. Since only the top 100 images are retrieved for each query, and there are exactly 100 relevant images to each query (because each category has exactly 100 images), the value of precision and recall is the same. We use "retrieval accuracy" in this paper to refer to both of them and plot it against number of iterations. As we can see, the retrieval accuracy improves rapidly as feedbacks goes on, achieving an average of about 70% after only 3 feedback iterations and about 90% after 10 feedback iterations. We also examine the case when there are 10% images in each category being initially annotated with the category name and compare the performance with the case without initial annotation, as shown in

Figure **6**. The two curves shown are the average retrieval accuracy calculated from 12 queries respectively. The performance with initial annotation available is considerably higher than that without initial annotation, especially in the initial several iterations. These experiments demonstrate that the proposed semantic similarity metric, together with integrated feedback scheme, provides an effective image retrieval approach.

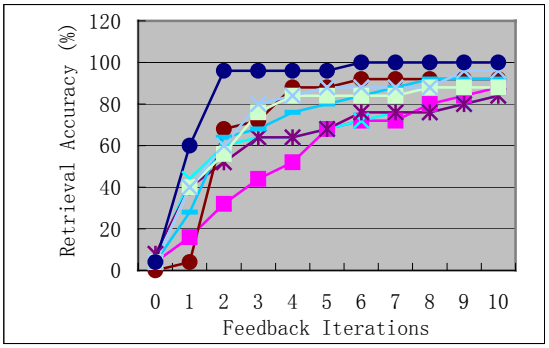

Figure 5: Performance on 8 random queries.

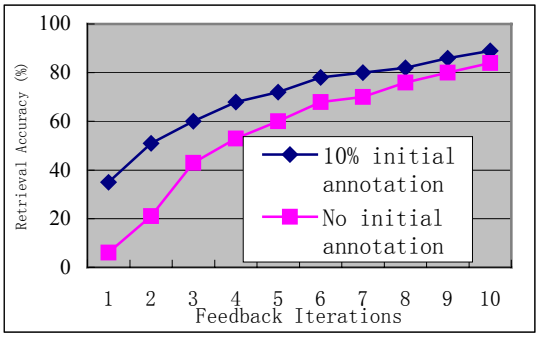

Figure 6: Performance comparison: 10% initial annotation versus no initial annotation.

A particular experiment is conducted to manifest the advantage of our semantic similarity metric over the exact keyword match scheme, in terms of capturing the variety of user preferences on use of descriptive keywords. This experiment is performed with the aid from 20 human subjects who has no knowledge on image retrieval. At first each subject is asked to browse through all the images by categories without knowing the category name. Later they are required to submit query intending to search for images from exact one category, with their own keywords that they thought might be descriptive to the category. In *iFind* there are 10% images in each category that have been annotated with the category name. The same automatic feedbacks strategy described above is utilized to track the performance of each query. We asked each human subject to submit five random queries (totally 100 queries) and recorded the average retrieval accuracy achieved by using our semantic similarity metric. In comparison, we conduct the same experiment using the exact keyword match scheme. As shown in Figure 7, our approach outperforms the exact keyword match scheme by an average of 10% of accuracy at each iteration.

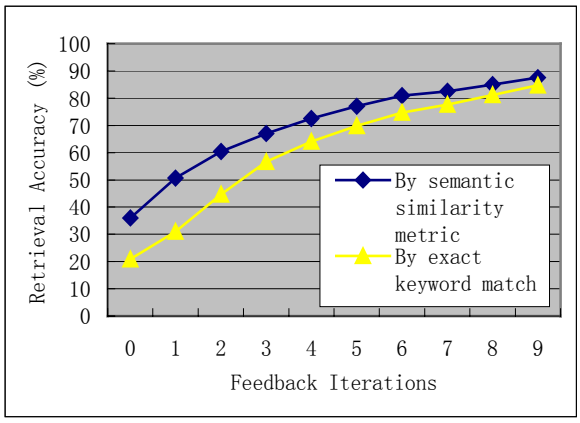

Figure 7 Performance comparison: semantic similarity metric versus exact keyword match.

In the experiment, it is noticed that the probability of the query keyword chosen by human subject coinciding with the category name is 58%, in which case our approach reduces to the exact keyword match scheme. In the remaining 42% cases, the query word is different with but closely related to the category name. In this case, the exact keyword match scheme works as if no initial annotation is available (because it measures the similarity between different words as zero), while our semantic similarity metric can still match the initially annotated images and consequently yield better performance. Furthermore, since all the images in Corel Image Gallery are clearly pre-classified, the users are more likely to use identical keyword for the same category. When conducting this experiment on randomly selected real-world images, the possibility of using the same descriptive (query) keyword can be even lower. Therefore, a more significant improvement of retrieval accuracy can be expected by using our approach over the exact keyword match scheme.

6. CONCLUSIONS

In this paper, we present the power of thesaurus-aided approaches to facilitate semantic access to image database. We construct the *dynamic semantic hierarchy* (DSH) interactively and progressively from WordNet , which is then visualized in *iFind* as a hierarchical category-based browsing tool that features flexibility and convenience. We also formulate a novel *semantic similarity metric* as a delicate measure of keyword similarity. This approach outperforms the exact keyword match scheme in terms of addressing the variety of relevant keywords used in image annotations and queries. It is seamlessly incorporated with visual features under a unified framework and helps it achieve a higher performance using less feedback iterations.

Currently our approaches are tailored to and incorporated into *iFind*. However, it is actually general enough to be used in other systems, provided that the semantic content of images is available. Since both approaches rely on WordNet as their knowledge base, they inevitably suffer from some ill organizations of words in WordNet. Hence, we attempt to use other thesauruses to improve our approach in the future work.

7. ACKNOWLEDGEMENT

The work presented in this paper was performed at Microsoft Research China and supported by Microsoft Visual Perception Lab of Zhejiang University.

8. REFERENCE:

[1] Yong, R., Huang, T. S., Chang, S. F. “Image Retrieval: Current Techniques, Promising Directions and Open Issues”, *Journal of Visual Communication and Image Representation*, Vol. 10, 39-62, 1999.

[2] Flickner, M. et al. “Query by image and video content: The QBIC system.” Computer. Vol 28, pp23-32, 1995.

[3] Lu. Y, Hu, C. H., Zhu, X. Q., Zhang, H. J., Yang, Q. “A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems”, *ACM Multimedia*, pp 31-38, 2000.

[4] Paek, S., Sable, C.L., Hatzivassiloglou, V., Jaimes, A., Schiffman, B.H., Chang S. F., Mckeown K.R, "Integration of Visual and Text-Based Approaches for the Content Labeling and Classification of Photographs", *ACM SIGIR* 1999.

[5] Aslandogan, Y. A., Their, C., Yu, T. C., Zou, J., Rishe, S. “Using Semantic Contents and WordNet in Image Retrieval”, *ACM SIGIR*, 1997

[6] Smeaton, A. F., Quigley, I. “Experiments on Using Semantic Distance Between Words in Image Caption Retrieval”, *In Proc. of the 19th International Conference on Research and Development in Information Retrieval,* Zurich, Switzerland, 1996.

[7] Miller, G. A., Beckwith, R., Felbaum, C., Gross, D., Miller, K. “Introduction to WordNet: An On-line Lexical Database”,

International Journal of Lexicography, Vol.3, No.4, pp. 235-244, 1990.

[8] Miller, G. A, "Nouns in WordNet: A Lexical Inheritance System", *International Journal of Lexicography*, Vol.3, No.4, pp. 245-264, 1990.

[9] Liu, W. Y., Sun, Y. F., Zhang, H. J. "MiAlbum—A System for Home Photo Management Using the Semi-Automatic Image Annotation Approach", *Technical Report,* Microsoft, 2000.

[10] Rui, Y., Huang, T. S. "A Novel Relevance Feedback Technique in Image Retrieval", *ACM Multimedia*, 1999.

[11] Rocchio, J. J. "Relevance Feedback in Information Retrieval". In *The SMART Retrieval System*, pp. 313-323. Prentice Hall, 1971.

[12] Bates M. (1986). "*Subject Access in Online Catalogs: A Design Model"*, Journal of the American Society for Information Science, 11, 357 - 376.