Building Semantic Indexing for Image Retrieval Systems

Chin-Hwa Kuo, Yu-Tao Huang, Yung-Hsiao Lan[+], and Tzu-Chuan Chou[*]
Computers And Networking Laboratory (CAN Lab.)
Dept. of CSIE, Tamkang University, Taiwan, R.O.C.
+886-2-2623-8784, chkuo@mail.tku.edu.tw

Abstract*-In this paper, we discuss the semi-automatic annotation that can build the semantic index for image retrieval systems. Every image will be annotated at annotators' private and partial wording, and then we use WordNet to add extra relative keywords to increase the searched probability of each image. Unfortunately, blind expansion will make the index of images unreasonable and unwise due to the common polysemous-characteristic of English words. Therefore, we use a modified Word Sense Disambiguation technique for the image annotation to identify the sense of every annotative keyword. After examining the results, we concluded that with the proposed mechanism, the precise rate of the image retrieval system has a dramatic improvement.*

Keywords: Semantic image indexing, Image Retrieval System, Polysemy, Word Sense Disambiguation, and WordNet.

1. Introduction

Capturing the semantic of an image is an interesting and challenging issue. The access to visual information is not only performed at a perceptual level, but also at a conceptual level [1]. To achieve this objective, one of the research directions in semantic modeling and representation makes use of semantic networks to retrieve target images [2]. Many image database research projects have devoted lots of efforts in this field. For example, MediaNet [3] uses partially annotated collections of multimedia data to enhance the retrieval of multimedia data. The concepts and relationships between the concepts are defined and exemplified by multimedia such as text, images, video, and audio-visual descriptors. In addition, Yang et al. proposed thesaurus-aided approaches to facilitate semantics-based access to images [4]. They constructed the semantic hierarchy, which supports flexible image browsing by semantic subjects. In the above two mentioned papers, both of them aimed to build a connection between semantic concept and image low-level features to provide a semantic retrieval function on a CBR image database system.

Our approach here is in the same direction as semantic networks, but we attempt to use a semantic concept in the indexing phase instead of using a low-level feature. In this paper, we extend our previous work in Can-Find [8], a semantic image indexing and retrieval system that mainly uses keywords to interchange information and aspects between users and machines. In order to search images by keywords, all images in the database systems must be annotated by relative words or phrases according to the content of each image automatically or manually. To annotate automatically by machines is not an easy goal that can be reached in one step. On the other hand, the annotation will be biased and partial towards the private weakness of the annotators. In this paper, we discuss the semi-automatic annotation that can build the semantic index for image retrieval systems. That is to say, every image will be annotated by the annotators' perception at their pleasure. Even if the annotative words or phrases are based on annotators' private and partial wording, we use lexical reference database, e.g. WordNet [6], to expand the annotative keywords about each image, and then to increase the searched range of each image. Unfortunately, blind expansion will make the index of images unreasonable and unwise due to the common polysemous-characteristic of English words. Therefore, we use a modified Word Sense Disambiguation (WSD) technique for the image annotation to identify the sense of every annotative keyword. We expect that with the WSD technique, we can expand every annotative keyword based on its correct sense according to the context, i.e. all annotated keywords of the image and according to the definition and sample sentences of every sense of the keyword provided by WordNet. The prototyped platform provides image retrieving functions for users in semantic level. After examining and analyzing the results, we concluded that through expansion by the selected sense, the image retrieval system gives a dramatic improved in precise rate as a whole.

This paper is outlined as following. The semantic index subsystem will be introduced in section 2. The details of keyword expansion and the proposed algorithm of Word Sense Disambiguation will be also discussed there. The results will be

[+] Yung-Hsiao Lan is now currently with Institute for Information Industry, Taiwan, R.O.C.
[*] Tzu-Chuan Chou is now currently with Institute of Information Science, Academia Sinica, Taiwan, R.O.C.

shown in section 3. Finally, the conclusion will be presented in section 4.

2. Semantic Indexing Subsystem

As shown in Fig.1, the fundamental building blocks of proposed semantic indexing subsystem include keyword extraction, Word Sense Disambiguation (WSD) and keyword expansion mechanisms. In the keyword extraction phase, the stop words and punctuations inside the annotated text are eliminated. After extracting the rest of the words, we use the word lemmatization method for each word to gain its normalized form. For instance, *working* is in the same class with *works*, *worked*, *work*, as the words with the normalized form *work*. In the WSD phase, the sense of each normalized-from word is decided according to all the annotative words belong to the same image and to the definitions and sample sentences of each sense provided by WordNet. In the keyword expansion phrase, every extracted word is expanded by the relationships supplied in WordNet with the decided sense in the WSD phrase. Consequently, all the extracted words and expanded words become the semantic index of the annotated image. Those who use the extracted words and expanded words to research in the proposed image database system will obtain corresponding annotated-images. The details of keyword expansion and WSD blocks of the indexing subsystem will be discussed in this section.

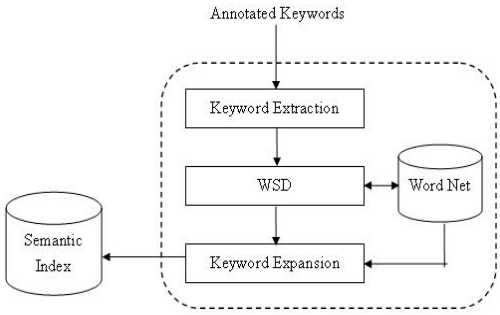

Figure 1. Semantic indexing subsystem

2.1 Keyword Expansion

After images are annotated, uses of the image retrieval system can search the corresponding images by given keywords. But in the case that the habitual wordings of the annotator and users are different, users cannot gain all the desired images. Even those retrieved images are annotated by relative keywords. For instance, if an image is only annotated by the word "sea" but not "ocean", it can be only searched by the word "sea" but not "ocean". In order to solve this problem, we have to add extra keywords to increase the searched probability and broaden the searched scope of each image. The extra keywords must be related to the content of images. A direct approach is to add keywords that are relative to the annotated keywords. Therefore, if the annotated words or phrases are based on the annotators' private and partial wording, we use lexical reference database, e.g. WordNet [6], to expand the annotative keywords related to the image, to increase the searched range of each image.

WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. There are 152,059 unique strings (words and phrases) and 115,424 synsets (synonym sets) in the WordNet 2.0. Different relations link the synsets. There are 16 different relations tables defined in the WordNet 2.0, e.g. hypernym/hyponym, derivation, cause of, attribute, region/usage/category domain classification, attribute, antonym, participle, member/substance/part meronym, etc. [6].

The main purpose of the keyword expansion is to use the relationships defined in the WordNet to broaden the searched scope of each image in the annotation phase. To simplify the expansion process, we use the most common relations, synonym and hypernym. For instance, the word "*human*" has synonyms such as {*person, individual, someone, and somebody, mortal, soul, homo, man,* and *human being*} and hypernyms such as {*organism, being, living thing, animate thing, causal agent, cause, casual agency, hominid, animal, animate being, beast, brute, creature, fauna,* etc}. In the case of simple keyword expansion, given an annotated keyword, the related synonyms and hyponyms are selected. On the other hand, in the case of complete keyword expansion, exploring all relationships in WordNet may be very useful, but it might lose focus. Hereafter we only discuss the simple keyword expansion, but the follow-up mechanisms are also fit for the complete keyword expansion.

We expect that the extended keywords generated in the keyword expansion phase might build a more complete conceptual map of an image, however blind expansion will make the index of images unreasonable and unwise due to the common polysemous-characteristic of English words. In fact, there are 81,795 single words (not including 70,264 phrases) and there are 23,722 polysemous single-words in the WordNet 2.0. Through our statistic, those 58,073 monosemous single-words' average tag_count is only 0.46, but the polysemous single-words' averge tag_count is 91.96. (tag_count is the word frequency information supplied in WordNet database, it indicates how common a word is in relation to a text. The number is equal to the times the word was found in the test corpus. Therefore the higher the number, the more common the word.) Thus as it can be seen, the more common words

usually have more than one sense. For this reason, comprehensive keyword expansions will add lots of unconcerned keywords into the index of the image database. For instance, if a landscape image including some plant is annotated as tree, flower, and plant, and we want to expand the word "*plant*", the expansion can be from the sense as "*a living organism lacking the power of locomotion*" or from another sense as "*buildings for carrying on industrial labor*." Obviously, the second sense is not suitable to the annotated image. If we expand from the second sense, some improper keywords, e.g. *works*, *industrial plant*, *building complex*, *complex*, *structure*, and *construction*, will be added into the index of the landscape image. This kind of result does not match our expectation. Nevertheless, once we consider the other annotative keywords of the same image, the meanings of the expanded words will be limited. How do we know the meaning (sense) of "*plant*" here for this image? We need a technique, like Word Sense Disambiguation (WSD), to make sure the sense of the keyword for the current annotated image. By looking at the other annotative keywords for the same image, the extended keyword will be more correct and useful. The details will be discussed in the following section.

2.2. Word Sense Disambiguation

Lots of word sense disambiguation research has been proposed. Dictionary based disambiguation uses definition and examples of each sense of target word in dictionary or thesaurus to identify the sense of a word. Lesk exploits the sense definitions in the dictionary directly [11]. Yarowsky shows how to apply the semantic categorization of words, which are derived from the categories in Roget's thesaurus, to the semantic categorization and disambiguation of contexts [10]. In Dagan and Itai's method, translations of different senses are extracted from a bilingual and their distribution in a foreign language corpus is analyzed for disambiguation [12]. Much of the research on automatic word sense disambiguation has been corpus base. Corpus-based approaches typically train a statistical classifier on contexts containing a polysemous word in a known sense. Based on what it has found in training, the classifier assigns a sense to novel occurrences of the polysemous word. Many resources with sense tagged information have been used, for example, SemCor [9] (Landes), Senseval (Kilgarriff and Rosenzweig; Pedersen) [13], DSO (Ng and Lee) [14], etc.

In order to identify the senses of the annotative keywords to expand the keywords from the correct meaning about the annotated images, a WSD process is implemented before the keyword expansion approach. Our approach is similar to the mentioned corpus-based WSD method, but we use the definition and sample sentences of each sense supplied by the WordNet database instead of the content of sense-tagged corpora and use the statistic result of SemCor, i.e. the tag count in the test corpus, as the training data. The training process applies Naïve Bayes approaches to build a word sense identifier [10], and the basic concept is explained as follows. To obtain the most possible sense is equal to gain the *S* defined as (1).

$$S = \arg_i \max P(S_i \mid C) \quad \text{.....(1)}$$

S is the most proper sense of the keyword. S_i is one of the senses of the target polysemous word. *C* is the set of contextual features compose by the keywords, i.e. the other keywords annotated to the same image. Under most circumstances, we cannot find the value of $P(S_i \mid C)$, but through the use of Bayesian Theorem, we can obtain the following formula:

$$P(S_i \mid C) = \frac{P(C \mid S_i)P(S_i)}{P(C)} \quad \text{.....(2)}$$

In order to attain the most probable sense for the targeted keyword, we can just ignore the value of *P*(*C*), because the value of *P*(*C*) related to the targeted keywords is always constant.

$$\begin{aligned} S &= \arg_i \max P(S_i \mid C) \\ &= \arg_i \max \frac{P(C \mid S_i)P(S_i)}{P(C)} \quad \text{.....(3)} \\ &= \arg_i \max P(C \mid S_i)P(S_i) \end{aligned}$$

Hence, we utilize the tag_count value supplied within the WordNet database to be the frequency of specific sense to obtain the probability of the targeted keyword's definition appearing, e.g. $P(S_i)$ as (4).

$$P(S_i) = \frac{freq_{S_i}}{\sum_{j=1}^{N} freq_{S_j}} \quad \text{.....(4)}$$

where $freq_{S_i}$ represents the frequency of the sense S_i, and $\sum_{j=1}^{N} freq_{S_j}$ is the sum of the frequencies of all the senses of the targeted keyword. To calculate the value of $P(C \mid S_i)$, we use the words in the definitions and sample sentences defined in the WordNet database as the occurring of the conditional event and assume all the words that lies in the annotations are conditionally independent, so the $P(C \mid S_i)$ can be calculated by (5).

$$P(C \mid S_i) = P(\{K_j \ in \ C\} \mid S_i) = \prod_{K_j} P(K_j \mid S_i) \quad \text{.....(5)}$$

where K_j is part of the keyword set.

Unfortunately, amount all meanings of the keyword, if one of them doesn't appear in the definition and sample sentence of the sense supplied in WordNet, the value of the term $P(K_j \mid S_i)$ will be zero. So the value of $P(C \mid S_i)$ is zero too, and then $P(C \mid S_i)$ cannot compare with each other. To avoid this situation, we change the formula of $P(C \mid S_i)$ that being smoothing.

$$P(C \mid S_i) \approx \frac{1}{V} + \frac{V-1}{V}\left(\sum_{j=1}^{M} P(K_j)\begin{cases} K_j \subset C\ ,\ P(K_j) = \dfrac{1}{M} \\ K_j \not\subset C\ ,\ P(K_j) = 0 \end{cases}\right) \quad \text{.....(6)}$$

where K_j is the j^{th} keyword defined of the target meaning, S_i, M is the total number of definition keyword, C is the set of contextual features compose by the keywords, and V is the average of the individual meanings defined in WordNet of the keyword. In our statistic, the value of V is 6.099. Let us have a insight into (6), if all the keywords of the definition and sample sentences of the specific sense do not appear in the annotation of the targeted image, the value of $P(C \mid S_i)$ can approximated as $1/V$, and then $P(S_i)$ will have the power to make decisions. That is to say, if we have no idea about correct sense, we trend to choose the most frequent sense to expand the keywords. On the other hand, if any keyword of the definition and sample sentences of the specific sense does appear in the annotation of the targeted image, it will directly influence decision result. The above discussion concludes that this approach is very reasonable. Based on (4) and (6), we can calculate the value of $P(S_i)$ and $P(C \mid S_i)$. Therefore, we can obtain the best sense of each annotative keyword according to the current annotated image by (3), and finally, we can expand each annotative keyword from its correct sense.

3. Results

To illustrate our idea and contributions, we have implemented a prototype. The system includes both indexing subsystem and query subsystem. In the query subsystem, simple keyword matching and ranking mechanism are built. The user interface is similar to our previous work [8]. To experiment the processes, we set up a small search engine with a collection of 1211 annotated images to test the three different indexing methods and examine the precision and recall of each method. There are three different types of indexing methods:

1. Without Keyword Expansion.
2. With Keyword Expansion.
3. With WSD and Keyword Expansion

During the experiment, we used ten different keywords to capture the search results as shown in figure 2. With the results, we analyzed the recall and the precision of the search. We discovered that keyword expansion mechanism does greatly increase the recall, but will cause a negative impact on the precision of the search. If we use both WSD and keyword expansion, the recall doesn't decrease too much, but the precision is better than that we only use keyword expansion. If the keyword we query has multiple senses and being imprecise, then the word sense disambiguation technique will improve the recall and the precision. On the other hand, if it is precise, word sense disambiguation will have less of an effect, as shown with "mouse" and "snow" topics. However, if we acquire the definition of the keyword before doing keyword expansion, and then use the definition we obtained to do keyword expansion, this will help improve the overall precision and recall of the search. Because People usually have different descriptions for the same image, it probably loses some information. For this reason, we can use the technique we present to avoid this condition and have a good performance in image retrieval.

4. Conclusions

Retrieving an image in semantic level is not only challenging but also interesting research issue. In the past few years, many researches have been devoted to this subject. Every image will be annotated at annotators' private and partial wording, and then we use lexical reference database, e.g. WordNet, to increase searched probability of each image. Unfortunately, blind expansion will make the index of images unreasonable and unwise due to the common polysemous-characteristic of English words. In this paper, we use a modified Word Sense Disambiguation technique for the image annotation to identify the sense of every annotative keyword and we discuss and experiment the three different types of indexing methods to obtain images through the annotated information. The precision and recall of the search is examined. After examining and analyzing the results, we concluded that with the proposed mechanism, the precise rate of the image retrieval system has a dramatic improvement. This is the main contribution of this paper to the image retrieval systems.

Acknowledgements

The present work is supported by National Science Council Taiwan ROC under the contract No. NSC92-2520-S-032-002.

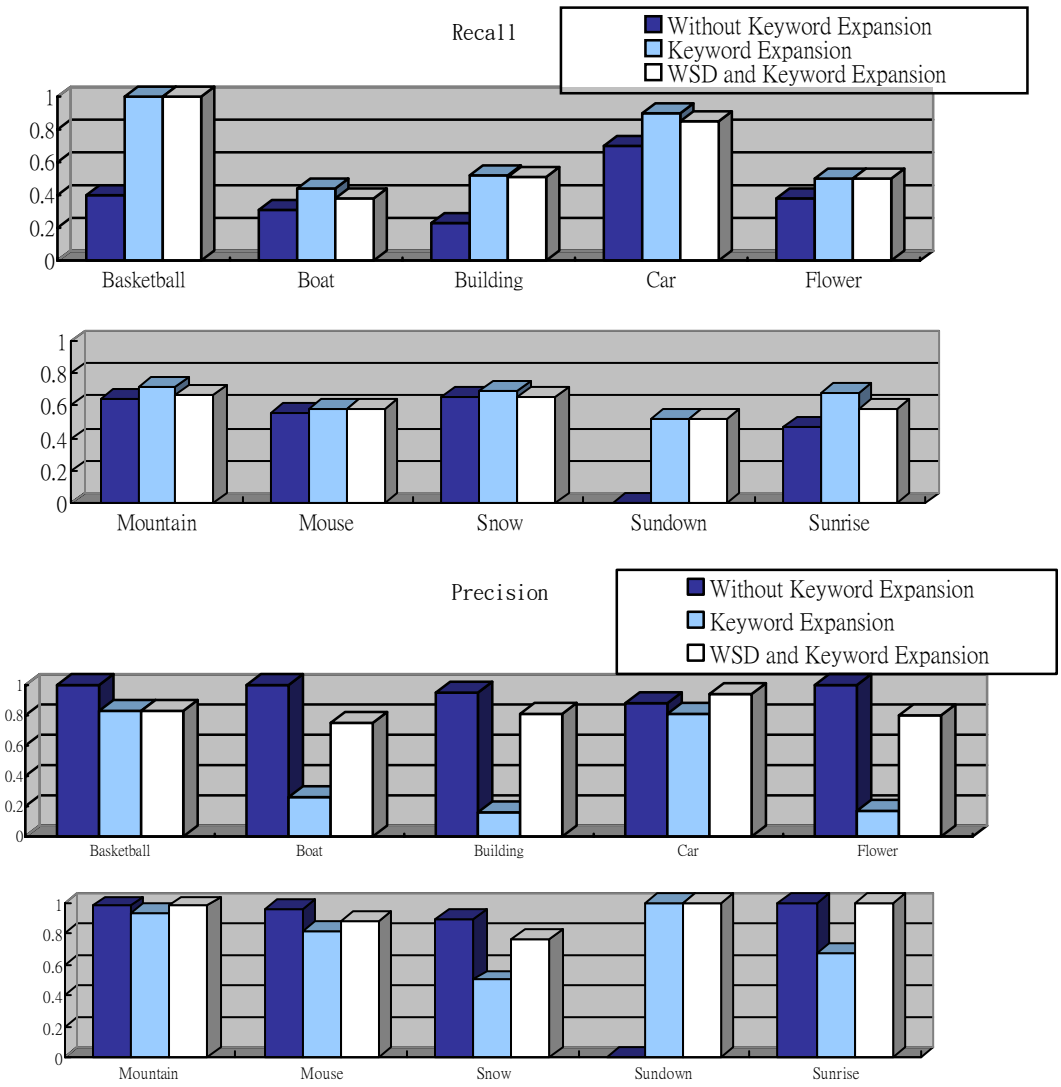

Figure 2. Recall and Precision on different indexing Scheme.

References

[1] Bimbo, Alberto, “Visual Information Retrieval,” Morgan Kaufmann Publishers, 1999.

[2] Wasfi Al-Khatib, Y. Francis Day, Arif Ghafoor, and P. Bruce Berra, “Semantic Modeling and Knowledge Representation in Multimedia Databases,” IEEE Transactions on Knowledge and Data Engineering, 1999, pp. 64-79.

[3] A. B. Benitez, J. R. Smith, and S. -F. Chang, “MediaNet: A Multimedia Information Network for Knowledge Representation,” Proceedings of IS&T/SPIE 2000 Conference on Internet Multimedia Management Systems, Vol. 4210, Boston, MA, Nov. 6-8, 2000.

[4] Jun Yang, Liu Wenyin, HJ Zhang, YT Zhuang, “Thesaurus-Aided Image Browsing and Retrieval.” Proceedings of IEEE International Conference on Multimedia & Expo (ICME) 2001, pp. 313-316.

[5] Jurafsky, Dan, "Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition," Upper Saddle River, N. J., Prentice Hall 2000.

[6] WordNet, http://www.cogsci.princeton.edu/~wn/, 2004

[7] British National Corpus, http://www.hcu.ox.ac.uk/BNC/, 2004.

[8] Chin-Hwa Kuo, Tzu-Chuan Chou, Nai-Lung Tsao, and Yung-Shiao Lan, “CanFind – A Semantic Image Indexing and Retrieval System,” ISCAS2003.

[9] Landes, Shari, Claudia Leacock and Randee I. Tengi. 1998. “Building Semantic Concordances,” in *WORDNET, AN ELECTRONIC LEXICAL DATABASE*, edited by Christiane Fellbaum, The

MIT Press, Cambridge, Massachusetts, London, England.

[10]Yarowsky, David. 1992. "Word sense disambiguation using statistical Models of Roget's Categories trained on Large Corpora," in COLING 14, pp. 454-460.

[11]Lesk, Michael. 1986. "Automatic sense disambiguation: How to tell a pine cone from an ice cream cone" in *Proceedings of the 1986 SIGDOC Conference*, pp. 24-26, New York. Association fro Computing Machinery.

[12]Dagan, Ido and Alon Itai. 1994. "Word Sense Disambiguation using a second language monolingual corpus," *Computational Linguistics* 20:563-596.

[13]Kilgrriff, Adam and Joseph Rosenzweig. 2000. "English SENSEVAL: Report and Results," in *Proceedings of 2nd International Conference on Language Resources & Evaluation.*

[14]Ng, Hwee Tou and Hian Beng Lee. 1997. Defence Science Organization Corpus of Sense-Tagged English. http://ldc.upenn.edu/LDC97T12.htm