Data-Driven Approach for Bridging the Cognitive Gap in Image Retrieval

Xin-Jing Wang[*†] , Wei-Ying Ma[†] , Xing Li[*]
Microsoft Research Asia[†]
Tsinghua University, China[*]

Abstract

Bridging the cognitive gap in image retrieval has been an active research direction in recent years. Existing solutions typically require a large volume of training data that could be difficult to obtain in practice. In this paper, we propose a data-driven approach that uses Web images and their surrounding textual annotations as the source of training data to bridge the cognitive gap. We construct an image thesaurus that contains a set of codewords, each representing a semantically related subspace in the feature space. We also explore the use of query expansion based on the constructed image thesaurus for improving image retrieval performance.

1. Introduction

One recent research focus for content-based image retrieval is on how to bridge the gap between low-level visual features and high-level user concepts. Works along this direction include image auto-annotation [4], annotation propagation [3], region-based methods [9] and learning-based methods [8]. A key challenge in bridging the cognitive gap is to get enough training data to learn the mapping functions from low-level feature spaces to high-level semantics. Today some commercial search engines claim to index 500M images on the Web. As Web images are typically surrounded by abundant textual annotations, they can be considered as labeled data set. In this paper we propose to use Web images as training data to bridge the cognitive gap. Although people may argue that the annotations for Web images are noisy and may not necessarily reflect the concepts in the images, we hope that through a data driven approach, useful knowledge can be extracted from this freely available data set.

Many previous research works have discussed the possible usage of such annotations [5, 7]. For example, [5] tries to organize pictures in a semantic structure by learning a joint probability distribution for words and art picture elements which makes use of statistical natural language processing and WordNet [1]. In [7], theory of "visual semantics" provides useful insight into some of the challenges of integrating text indexing with image understanding algorithms.

In this paper, our key idea is to construct an image thesaurus to present the knowledge extracted from the Web. The image thesaurus contains two parts. One is a codebook that is trained to partition the feature space into sub-spaces, each corresponding to a semantically related concept. The other is a correlation matrix that indicates how two given concepts coexist in a same image. With this correlation matrix, we could also perform query expansion to improve image retrieval performance.

2. Using the Web to Build Image Thesaurus

The annotations for Web images come from many sources such as surrounding text, file name, alterative tag, etc. If we could extract the right keywords and associate them with the corresponding regions in the images, we will be able to construct an image thesaurus that can serve as a vehicle to bridge the gap between low-level features and high-level semantics for image retrieval. The various key technologies for building such an image thesaurus are discussed in the following.

2.1. Key Term Extraction

We use a vision-based web page analysis technique [2] to extract image surrounding texts. The HTML tag of each term is used to assign different weight to the term appeared in the texts. The basic idea is to give a lower weight to a term that occurs more frequently in a less important tag (similar to term frequency (TF) in information retrieval). A list of candidate terms with their weights in a descending order is constructed as shown in Figure 1.

Figure 1. Terms extracted for a Web image and its attention map

0-7803-8603-5/04/$20.00 ©2004 IEEE.

Coyote, prairie wolf, brush wolf, Canis latrans
=> wolf
=> canine, canid
=> carnivore
=> placental, placental mammal
=> mammal

Figure 2. Hypernyms for "coyote" from WordNet

To filter out noisy terms, we use WordNet [1] to keep nouns and save the output of WordNet with sense 1. Furthermore, we can obtain the hypernym of a word (IS_KIND_OF) to build a hierarchical codebook for image retrieval. A part of the hypernym tree of the word "coyote" is shown in Figure 2. We also keep the synonyms of each term for more advanced matching.

We also employ some heuristic rules to weight the terms differently. For example, using hypernym tree structure, the term that is more specific, e.g. "coyote," is given a higher weight than the term that is more collective, e.g. "mammal". The term with highest weight is selected as the key term for the image.

2.2. Key Image Region Extraction

The next step is to associate the key terms with their corresponding regions in the images so that we can relate high-level semantics to low-level features to bridge the cognitive gap. To solve this problem, we first segment each image into homogeneous regions using JSEG algorithm [10]. Since the resulted regions are not yet at semantic or object level, we further use the image attention model [12] to help identify the most important region in an image. We order the regions based on their attention values. An example of image attention map is shown in Figure 1. As can be seen, it helps us identify the "coyote" region from the "grass" background. The most salient region is selected as key region to associate with the key term. By this way, we obtain a large collection of key regions and the associated key terms that are very likely to be the semantic annotation of these regions.

2.3. Image Thesaurus Construction

The constructed image thesaurus is shown in Figure 3. It contains two parts: the codebook and the correlation matrix. The codebook contains codewords as the leaf nodes. Some of them have semantic meanings and are interrelated in a hierarchical manner. These codewords, denoted as *semantic-level codewords*, are trained using key regions as their semantic meanings (i.e. associated key terms) are known and their hierarchical relationship can be obtained using WordNet. The codewords in the flat structure are trained using the rest of regions that have no mapped key terms. As there is no semantic meaning associated with these codewords, we call them *low-level codewords*.

For every image region in our database, we extract a set of low-level color and texture features to represent it. The image regions with the same or similar key term (based on synonym) are grouped together, and the centroid of their feature vectors is used to present the codeword (semantic-level). The regions without semantics are clustered using K-means algorithm, and similarly we use the centroid of the cluster in the feature space to present the corresponding codeword (low-level).

The next step is to learn the correlation matrix. There are three kinds of correlation here: 1) the correlation between semantic-level and low-level codewords, 2) the correlation between low-level codewords, and 3) the correlation between semantic-level codewords. The correlation is calculated based on how frequently two regions coexist in a same image. A conditional probability is used to measure how likely a codeword would appear in an image given the existence of another codeword:

$$p(c_j \mid c_i) = \frac{p(c_j, c_i)}{p(c_i)} = \frac{\sum_{I_k \in I} f(c_j, c_i \mid I_k)}{\sum_{I_k \in I} f(c_i \mid I_k)}$$

$$f(c_i \mid I_k) = \begin{cases} 1 & c_i \in I_k \\ 0 & c_i \notin I_k \end{cases} \qquad f(c_j, c_i \mid I_k) = \begin{cases} 1 & c_i, c_j \in I_k \\ 0 & c_i, c_j \notin I_k \end{cases}$$

where c_j and c_i denote the j^{th} and the i^{th} codewords, and I_k denotes the k^{th} image in image set I.

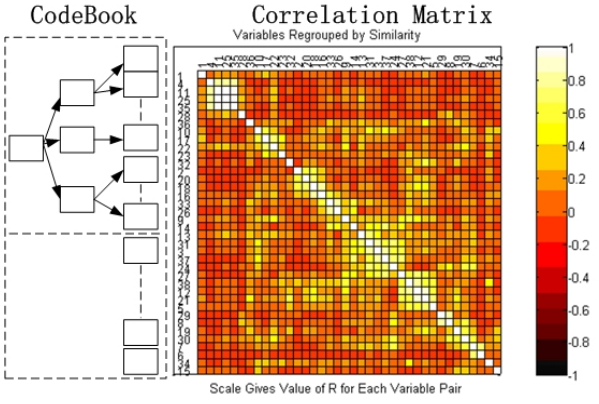

Figure 3. The image thesaurus constructed from the Web image data

3. Image Retrieval by Query Expansion

To illustrate how our constructed image thesaurus can help improve the performance of image retrieval, we pick query expansion which is often used in text retrieval as an example here. There are many other

possible uses of the thesaurus that are not covered in this paper.

3.1. Representing an Image Using Codewords

For each image in our database, we first perform image segmentation and extract feature vectors from all the image regions. Each region is then mapped to a codeword which has minimum distance to it in the feature space. After this stage, each image in our database is represented by a set of codewords.

3.2. Query Expansion

With the thesaurus we constructed, we can support both query-by-example and query-by-keyword for image retrieval. In the case of query-by-example, we first perform the same analysis as Section 3.1 on the query image, and then select a key region from the image to expand the query. We augment the query by including highly correlated codewords based on the correlation matrix. We use the EMD [11] to compute the similarity between the query example and images in our database.

In the case of query-by-keyword, if the keyword (e.g. "wolf") maps to a semantic-level codeword at the leaf node, then the query will contain the mapped semantic-level codeword and a set of highly correlated codewords based on the correlation matrix. If the keyword is a concept (e.g. "mammal") that maps to an immediate node in the semantic hierarchy, then the query will contain the semantic-level codewords that are children of that immediate node. Note that these codewords are used as "OR" queries to retrieve images. Similarly, we can also expand each of these codewords by adding those highly correlated ones.

The similarity measure used in query-by-keyword search is the *Jaccard coefficient* [6]. Let A denote the set of codewords of image I_i in the database and B the set of codewords of a query Q_j which is the single query or one of the "OR" queries. The similarity measure is defined as below:

$$Sim(I_i, Q_j) = \frac{\|A \cap B\|}{\|A \cup B\|} = \frac{\|A \cap B\|}{\|A\| + \|B\| - \|A \cap B\|}$$

where $\|A \cap B\|$ is the number of common codewords in A and B, and $\|A \cup B\|$ is the total number of different codewords in A and B. The similarity between an image and the query is equal to

$$Sim(I_i, Q) = \max_{j \in (1, \|Q\|)} Sim(I_i, Q_j)$$

where Q represents the set of "OR" queries. In single query case, $\|Q\| = 1$.

4. Experiments

We crawled 17,123 images from the Web with 10,051 images successfully identified their key terms. These images cover animals, human beings, scenes, advertise posters, books, and sweaters, etc. The visual feature extracted from each image is a combination of color moments, correlogram and wavelet texture features which result in 171 dimensions. From these images, we constructed a codebook with 4750 semantic-level codewords and 500 low-level codewords that are clustered by K-means algorithm.

4.1. Performance of Key Term Extraction

To evaluate the performance of our key term extraction for Web images, we randomly selected 20 query words to search images in our database. The retrieval result (precision@10) is given in Figure 4. As can be seen, the performance is generally satisfactory.

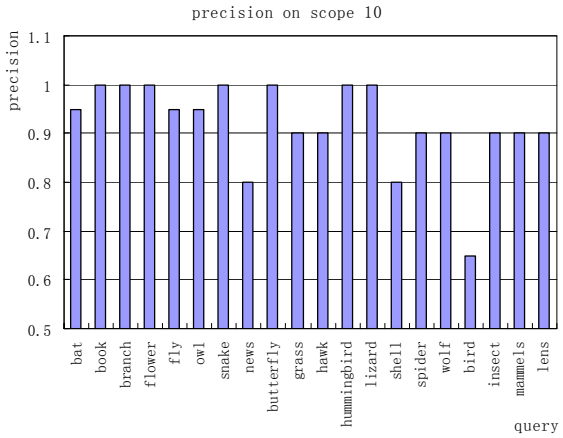

Figure 4. Precision@10 for image retrieval based on key term extraction.

4.2. Image Retrieval

To evaluate how the learned image thesaurus can improve image retrieval performance, we select 10,000 images from Corel Stock Photo Library as our testing data set. These images do not overlap with our training images obtained from the Web, but they cover similar high-level concepts as the training images with hundreds of outliers. After image segmentation and feature extraction, these 10,000 Corel images are indexed using our image thesaurus with each image region represented by a codeword.

In the case of query-by-keyword, the key term submitted by a user is first matched to a semantic-level codeword, and the feature of the codeword and those features of other correlated codewords are then used to form a content-based query to search images in the database. Figure 5 shows the result of query "wolf".

The images in red box are correct hits. Note that this example shows the capability of indexing images using high-level concept. Figure 6 shows the result of query "bird" which is an example when the query maps to an immediate node in the semantic hierarchy. In this case, all semantic-level codewords at the leaf nodes whose father is "bird" are used to form the query set.

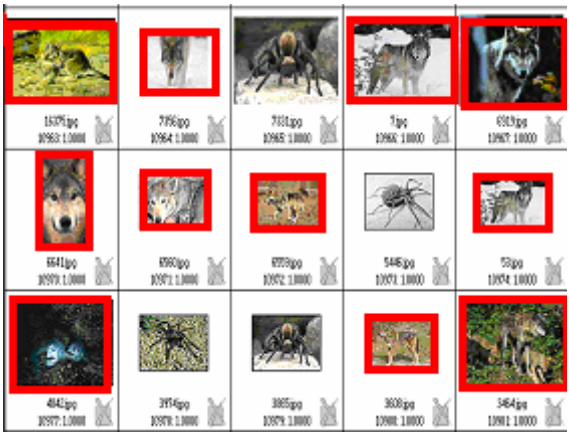

Figure 5. Retrieval result of query "wolf"

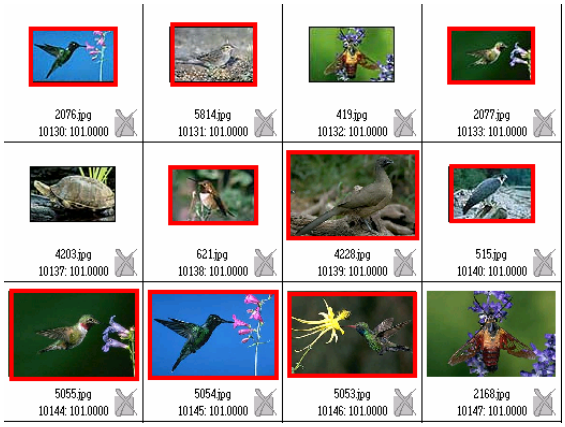

Figure 6. Retrieval result of query "bird"

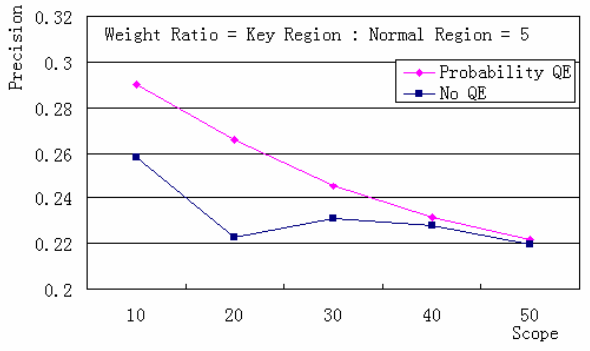

Figure 7. Performance Evaluation for Query Expansion

We randomly selected 100 queries related to the testing data set to evaluate the use of query expansion for image retrieval. Traditional precision-scope measure is used for performance evaluation. Figure 7 shows the comparison of the base line with our query expansion method.

6. Summary

In this paper, we presented an idea of using Web images as training data to create an image thesaurus which helps solve the problem of cognitive gap in image retrieval. The query expansion was also introduced to take advantage of this correlation information in the thesaurus to further improve image retrieval performance. The hyperlinks between Web images are valuable information to use for learning image thesaurus. We believe that by leveraging link information and combining it with WordNet, we can further improve the performance of this work. We plan to investigate this direction in our future works.

7. References

[1] C. Fellbaum, WordNet: An electronical lexical database, MIT Press, Cambridge, Mass., 1998

[2] D. Cai, S. Yu, J.R. Wen, and W.-Y. Ma, "VIPS: a vision-based page segmentation algorithm", Microsoft Technical Report, MSR-TR-2003-79, 2003

[3] H.J. Zhang and Z. Su, "Improving CBIR by Semantic Propagation and Cross-Mode Query Expansion", Multi-Media Content Based Indexing and Retrieval, 2001

[4] K. Barnard , P. Duygulu, D. Forsyth, N. Freitas, D.M. Blei and M. Jordan, "Matching Words and Pictures", Journal of MLR 3, 2003, 1107-1135

[5] K. Barnard, P. Duygulu, and D. Forsyth, "Clustering Art", CVPR 2001, pp. II:434-439.

[6] P.Sneath, and R.Sokal, "Numerical Taxonomy: the Principles and Practice of Numerical Classification", San Francisco: W.H. Freeman, 1973. 573p

[7] R.K. Srihari, "Use of Multimedia Input in Automated Image Annotation and Content-Based Retrieval", Presented at SPIE'95, San Jose, CA, Feb. 1995.

[8] S. Tong, E. Chang, "Support Vector Machine Active Learning For Image Retrieval", In Proc. ACM Multimedia, Ontario, Canada, 2001.

[9] W.Y. Ma and B. S. Manjunath, "Netra: A toolbox for navigating large image databases", in IEEE ICIP, 1997.

[10] Y. Deng, and B.S. Manjunath, "Unsupervised Segmentation of Color-Texture Regions in Images and Video", IEEE Trans. on PAMI, 23(8), 2001, 800-810

[11] Y. Rubner, L.J. Guibas, and C. Tomasi, "The Earth Mover's Distance, Multi-Dimensional Scaling, and Color-based Image Retrieval," Proceedings of the ARPA Image Understanding Workshop, New Orleans, LA, May 1997, pp. 661-668

[12] Y.F. Ma, and H.J. Zhang, "Contrast-based Image Attention Analysis by Using Fuzzy Growing", ACM Multimedia, 2003.