EXPLORING STATISTICAL CORRELATIONS FOR IMAGE RETRIEVAL

Xin-Jing Wang [1], *Wei-Ying Ma* [2], *Xing Li* [3]

[1] CERNET Center, Room 305, Tsinghua University, Beijing 100084, China

Phone: (86-10) 64281296

Email: wxj01@mails.tsinghua.edu.cn

[2] Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, China

Phone: (86-10) 62617711-3175

Fax: (86-10) 6255-5531

Email: wyma@microsoft.com

[3] CERNET Center, Room 224, Tsinghua University, Beijing 100084, China

Phone: (86-10) 62785983

Fax: (86-10) 62785933

Email: xing@cernet.edu.cn

ABSTRACT

Bridging the cognitive gap in image retrieval has been an active research direction in recent years, of which a key challenge is to get enough training data to learn the mapping functions from low-level feature spaces to high-level semantics. In this paper, image regions are classified into two types: key regions representing the main semantic contents and environmental regions representing the contexts. We attempt to leverage the correlations between types of regions to improve the performance of image retrieval. A *Context Expansion* approach is explored to take advantages of such correlations by expanding the key regions of the queries using highly correlated environmental regions according to an image thesaurus. The thesaurus serves as both a mapping function between image low-level features and concepts and a store of the statistical correlations between different concepts. It is constructed through a data-driven approach which uses Web data (images, their surrounding textual annotations) as training data source to learn the region concepts and to explore the statistical correlations. Experimental results on a database of 10,000 general-purpose images show the effectiveness of our proposed approach in both improving search precision (i.e. filter irrelevant images) and recall (i.e. retrieval relevant images whose context may be varied). Several major factors which have impact on the performance of our approach are also studied.

Key words: query expansion, image thesaurus, content-based image retrieval, region-based image retrieval

1. INTRODUCTION

One recent research focus for Content-Based Image Retrieval (CBIR) is to enable retrieval at semantic or concept level. Some related works along this direction include retrieving image at a finer granularity, i.e. region-based methods[6][7][9][8][16][17][23], user's relevance feedback[6][7][14][22][17], images/regions auto-annotation[2][8][9], and learning-based methods[14].

However, one drawback of the previous works is the ignorance of the associations between image regions provided by images themselves. In fact, each general-purpose image is normally constructed by a dominating object (i.e. concept) and its environment information, and different concepts will have their specific environments. For example, Figure 1 shows two groups of example images selected from the Coral Database. In the first row, i.e. tiger images, the concepts are the tigers and the environments are the forests. And in the second row, they are penguins and snow. Assume that each image is composed of key regions which represent the main semantic content (e.g. the tigers and penguins) and environmental regions which represent the context (e.g. the forests and snow). It can be seen from Figure 1 that the environmental regions share some similarity within the same concept, but are largely varied when across different concepts. It is also the common knowledge that the tiger rarely appears in sea (blue region) as penguin does while the penguin seldom shows up in forest (green region) as tiger does. This indicates that the statistical correlations between image regions contain useful information to assist image retrieval.

Figure 2 shows a 38x38 correlation matrix learned from 10 categories of manually labeled Corel images mapping to 38 distinct region patterns, by counting how frequently two region patterns appear together in a same image. The brighter an element is in the correlation matrix, the more likely the corresponding two region patterns co-exist in a same image. Clearly from Figure 2 we can see some strong correlation among these region patterns.

Motivated by the above observation, we propose in this paper a *Context Expansion* approach to help an image retrieval system to filter irrelevant images and retrieve more relevant images whose contexts may be varied. It is realized in such a way: for each query image, we first identify its key region and then augment the query by including its highly correlated region patterns based on a pre-learned correlation matrix. Then we use this augmented query to search the image database.

Note that in our approach, the additional regions used to expand the key region are learnt from a large collection of images' "contexts", i.e. the environmental regions in those images. In some sense, we try to discover the underlying "rules" of image construction. For example, when a "tiger" region appears in an image, it implicitly imposes a conditional probability model to confine what other environmental regions may appear in that image. Then, we use these rules (i.e. conditional probability models) to perform context expansion to improve the precision and recall of image retrieval. This is fundamentally different from previous query expansion approaches used in image retrieval where the expansion is either based on keywords [4][22] (therefore follows the traditional query expansion technique in text retrieval) or based on relevance feedback which uses the (pseudo-)relevant images to modify/expand the features of the query[10].

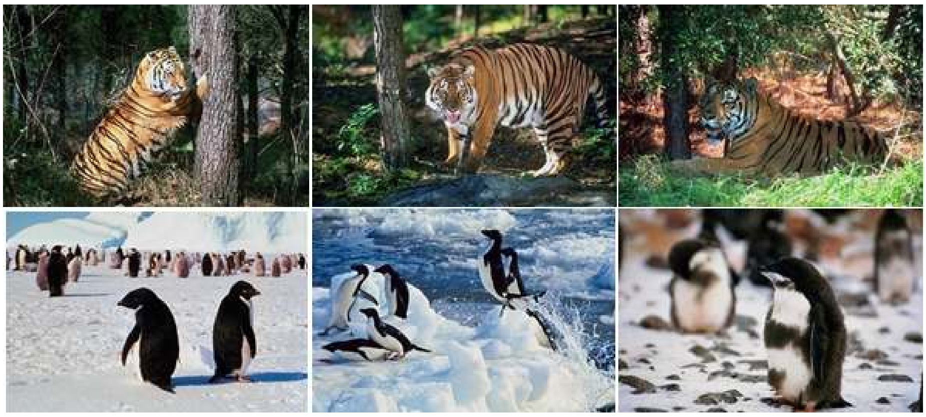


Figure 1 Region Correlations in Images

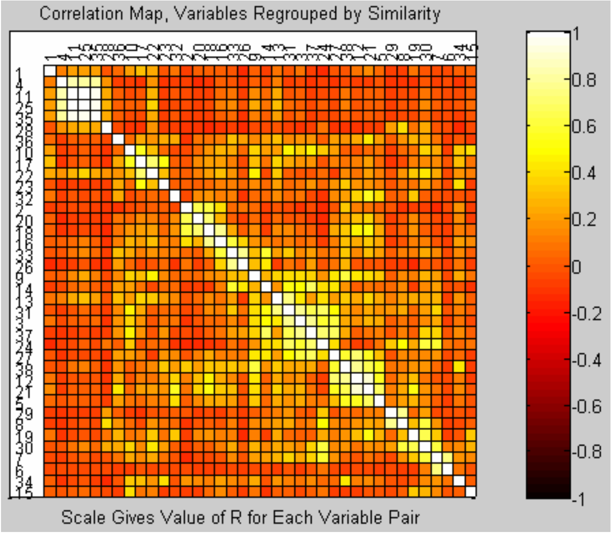


Figure 2. Correlation matrix of 38 region patterns obtained from the training images from the Corel Database

However, there are two key technical challenges which have a great impact on our proposed approach: 1) how to identify the key region in an image, which is needed during the learning phase in which the correlation matrix is constructed, as well as during the retrieval phase in which the key region of the query example needs to be identified; 2) how to precisely categorize homogeneous regions and learn the correlation matrix based on their iconic representation.

As to the first problem, we adopt the image Attention Model [21] which is used to select the region that attracts most of the user's attention. We will discuss it in details in Section3.

As to the second problem, it is risky to group image regions according to only their low-level features due to the cognitive gap. This falls into the old problem --- the lack of training data. As Web images are typically surrounded by abundant textual annotations and also structuralized by Web links, they can be considered as labeled dataset. In this paper, we use Web images as training data to quantize homogeneous image regions into image codewords and learn their inter-correlation matrix. The constructed region codebook and their correlation matrix make up of a so-called Image Thesaurus in our approach. 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 dataset.

It is worthwhile to highlight several aspects of the proposed approach here:

1. We attempt to investigate the latent object correlations in images and make use of such information to help image retrieval.
2. We provide a data-driven approach which uses the Web as training data to bridge the cognitive gap between image regions' low-level features and human's concepts, and hence secure the region categorization. This semi-supervised approach helps the construction of a more precise correlation matrix.

We organize our paper as follows. Section 2 briefly reviews the related works. Section 3 presents the construction of Image Thesaurus leveraging Web data and Section 4 details our Context Expansion approach. We show the experimental results in Section 5 and discussed some related problems in Section 6. We conclude our approach in Section 6.

2. RELATED WORKS

The idea of query expansion (QE) has been used in some previous research works in image retrieval [4][5][10][22]. In [5], the authors attempt to automatically annotate and retrieve images by applying QE in its relevance model based on a set of training images. They represent an image as a set of blobs resulted from clustering on image features as well as a set of keywords. They assume that for each image there exists a relevance model containing all possible blobs that could appear in this image and all the keywords that could appear in the caption of the image. The probability of drawing a keyword or a blob from the model of an image is given by a linear combination of the probability of this keyword/blob appearing in this image and that in the training set. That is to say, this keyword/blob is expanded by those keywords/blobs that co-occur with it in the selected training dataset. And the linear combination procedure is fundamentally a smoothing scheme, which is different from our approach.

In [4], the authors proposed a cross-modality query expansion approach to further improve image retrieval performance. When a query is given in form of keywords, it is expanded by a set of keywords selected from a semantic keyword network constructed through keyword propagation, and the low-level features of the corresponding images are then incorporated into that of the query. In [22], a keyword similarity matrix is learned by a statistical algorithm through users' relevance feedback, and a soft query expansion approach based on this similarity matrix is adopted to infer keywords which is somewhat related to the user's interest. For example, when "Ford" and "Toyota" is used as query, the system will automatically infer "car" through query expansion. Both these two approaches adopt text-based QE.

In [10], the authors presented two QE methods: Similar Expansion (SE) and Distance Expansion (DE) to reformulate queries. SE approach selects those images that are most relevant to the query and adds their features to those of the queries. DE approach selects not only the most relevant images as SE approach does, but also those image that are less similar (but also relevant) in case that they may give the user opportunities of adding new information. Similar to [4][22], it obtains expansion terms through user's relevance feedback and based their retrieval by global image.

Many previous research works also discussed the possible usage of Web annotations [9][13]. For example, [9] tries to organize pictures in a semantic structure by learning a joint probability distribution for keywords and art picture elements which makes use of statistical natural language processing and WordNet [1]. In [13], theory of "visual semantics" provides useful insight into some of the challenges of integrating text indexing with image understanding algorithms.

3. DATA-DRIVEN APPROACH 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.

3.1 Key Term Extraction

An effective web-page segmentation technique called VIPS (VIsion-based Page Segmentation) [3] is used to extract images' surrounding texts from the web-pages containing those images.

VIPS extracts the semantic structure of a web-page based on its visual presentation. The semantic structure extracted is represented as a tree. Each node in the tree corresponds to a *block* and each block will be assigned a value called DOC (*Degree of Coherence*) to indicate how coherent of the content in the block is based on visual perception.

Based on the terms output by VIPS algorithm, we first filter out the stop words and the words inexistent in WordNet, and keep only nouns. Then the remained terms are assigned different weights according to the following strategy:

1) *The more important a term's HTML tag is, the higher is its weight*. Currently we defined 12 HTML tags. Each tag is assigned a certain weight (see Table 1. the tags are placed in descending order by their importance from left to right). If one term appears in many HTML tags, its final weight will be the arithmetic mean of all the tags' weights. Hence the more frequently a term occurs in less important tags, the lower is its weight. We denote such a weight as W_{tag} .

2) *The more common a term is used as a noun, the higher is its weight*. WordNet [1] defines for each noun a property of "familiarity" which indicates how common this term is. For example, the term "wolf" scores 5 and "canine" scores 2. We denote the familiarity weight as W_{fami} .

3) *The more superior is the category of the term, the higher is its weight.* Currently we defined 5 categories which have different priorities (see Table 2. the category names are in priority descending order). We use WordNet [1] hypernym tree (i.e. IS_KIND_OF, Figure 4 shows an example) to classify each term into a category and use the category's priority as its category weight, denoted as W_{cat} .

Table 1. HTML Tags in Descending Order of Their Importance

ANCHOR, CAPTION, TITLE, ALT, META, URL, H1_H2, H3_H6, STRONG, LARGE, MIDDLE, SMALL

Table 2. Term Category Names in Descending Order of Their Priorities

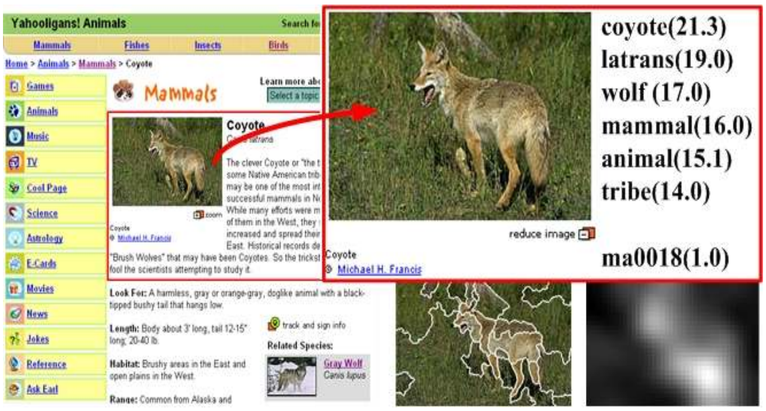
ANIMAL, PLANT, HUMAN, ARTIFICIAL, OTHERS

4) *The more specific a term is, the higher is its weight.* We use WordNet [1] hypernym tree to calculate the speciality of a term, i.e. the higher a term's hypernym tree is, the more specific this term is. For example, the hypernym tree of "coyote" is higher than that of "mammal". This means we prefer proper noun to collective noun. We denote the term level in a hypernym tree as n_{levl} and the value of n_{levl} increases as the noun becomes more specific (e.g. the term "entity" in Figure 4 has $n_{levl} = 1$ and the term "coyote" has $n_{levl} = 12$). We define a weight W_{levl} which is the same for all levels.

Assume t_i is a candidate term, and $Score_i$ is its final score. The value of $Score_i$ is given by:

$$Score_i = \begin{cases} W_{tag} * W_{fami} * (W_{cat} + n_{levl} * W_{levl}) & ,if < cat > \neq "OTHERS" \\ W_{tag} * W_{fami} * W_{cat} & , otherwise \end{cases} \quad (1)$$

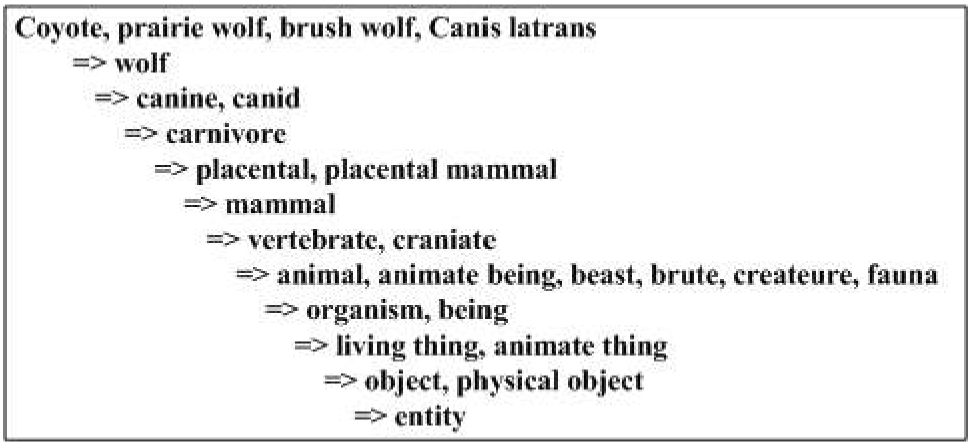
Equation (1) means that when the term belongs to the category "OTHERS", we do not consider its level weight (as detailed in the 4th strategy above). Images with all of their term weights less than a certain threshold (currently 1.0) will be filtered out from the training data. We sort the rest terms in score descending order and assume the top ranked one to be the key term. If more than one term ranks the highest, we simply remove the corresponding image from the training dataset. An example of sorted candidate terms is shown in the top-right corner of Figure 3.



coyote(21.3)
latrans(19.0)
wolf (17.0)
mammal(16.0)
animal(15.1)
tribe(14.0)

ma0018(1.0)

Figure 3. Terms extracted for a Web image and its attention map



Coyote, prairie wolf, brush wolf, Canis latrans
=> wolf
=> canine, canid
=> carnivore
=> placental, placental mammal
=> mammal
=> vertebrate, craniate
=> animal, animate being, beast, brute, createure, fauna
=> organism, being
=> living thing, animate thing
=> object, physical object
=> entity

Figure 4. Hypernym Tree for "coyote" from WordNet

3.2 Key Image Region Extraction

We adopt the attention model technique proposed in [21] to identify the key region of an image. The attention model detects the attention area in images. It first generates a saliency map based on local contrast analysis and then applies a *fuzzy growing* algorithm to extract attended areas or objects from the saliency map. Figure 5 shows two sample images and their saliency maps. It can be seen that the coyote and tiger are separated from the background of "grass".

In our approach, we assume that the key region in an image is the region that is the most "salient" to human eyes. We first segment each image into homogeneous regions using JSEG algorithm [19]. Since the resulted regions are not yet at semantic or object level, we further use the saliency map to modify this segmentation result.

Because each pixel's salience value output by the saliency map is represented by a float number normalized to (0, 1), we define the salience of a region r_j (resulted from JSEG algorithm) to be the average salience value of pixels enclosed in this region, denoted by $av(r_j)$, and merge all the regions whose salience exceeds a certain threshold λ, which results in the key region $r^{(d)}$:

$$r^{(d)} = \{r_j \mid av(r_j) > \lambda\} \tag{2}$$

λ can either be a fixed value or dynamically adapted for different images.

The resulted key region is also the one having the largest salience value. We keep the rest of the regions as the original ones output by JSEG segmentation algorithm, which make up of the "context" of the key region. In this way, each image is separated into two kinds of regions: the key region (only one) which represent the main semantic content and the environmental regions (normally more than one) which represent the context.

We then associate the key terms with their corresponding key regions. As a result, 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. And the image codebook and correlation matrix can then be learned from these training data, which produces the Image Thesaurus.

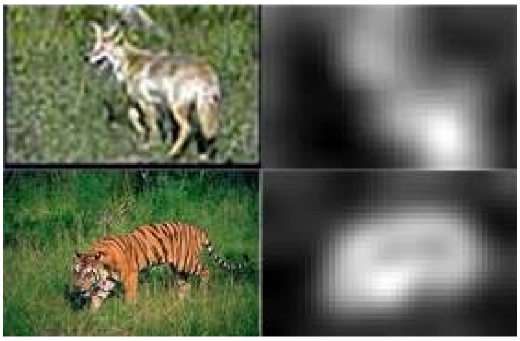


Figure 5. Two samples on saliency map obtained by [21]

3.3 Image Codebook Generation

Given the annotated regions learned in Section 3.1 and 3.2, we want to organize the concepts into a tree-structure. The reason is that hierarchical structure can better reflect the relationships between concepts which coincide with human concept space, and hence enable the query-by-keyword search in various scales (e.g. both “sparrow” and “bird” are easily supported).

One method to generate the codebook hierarchically is to leverage WordNet [18]. As shown in the left part of Figure 7. The codebook contains codewords as the leaf nodes. There are two types of codewords: the semantic-level codewords and the low-level codewords. The semantic-level codewords have meaningful concepts learnt during key term extraction. They are the centroids of key regions’ clusters, each corresponding to an individual concept. We integrate all the identified concepts (keywords) into a single tree according to the hypernym trees given by WordNet, and associate the leaf-nodes with the low-level features of cluster centroids. These leaf nodes are the semantic-level codewords. Note that they have semantic meanings. Their fathers are keywords without low-level visual features.

The low-level codewords have no semantic meanings and are learnt from the environmental regions. They are the cluster centroids output by K-means clustering algorithm. The pseudo-codes of our K-means clustering algorithm are shown in Figure 6. We apply the clustering procedure for several iterations (decided by the parameter #neps) and select the one which has the smallest sum of intra-cluster distances as the final clustering result. In each iteration we adopt a strict criterion that if no samples change their cluster ids or the number of

```
Input:    #neps --- the number of k-means iterations
          #k    --- the number of clusters
Output:   centroids of the clusters

for i = 1 to #neps
    1.  randomly select #k samples as the initial cluster centroids;
    2.  while #changed_samples > 0 && #cluster_iterations < MAX_ITER_THRESHOLD
        a)  for the rest of samples in the dataset, assign the samples to its nearest cluster.
            The distance metric is Euclidean distance;
        b)  update cluster centroids
        c)  calculate the number of samples who changed their cluster ids.
        d)  # cluster_iterations = # cluster_iterations + 1
    3.  calculate the sum of intra-cluster distances #dist_sum. The intra-cluster distance of a
        cluster is the sum of distance of any two samples inside that cluster;
    4.  if i = 1
        a)  save #dist_sum to #smallest_dist_sum, record current centroids
        b)  i=i+1;
        else if #dist_sum < #smallest_dist_sum
        a)  delete the old centroids and record current centroids
        b)  i=i+1;
        end
end
```

Figure 6. Pseudo-codes for k-means Clustering Algorithm Used

clustering iterations (note that it is not #neps but #cluster_iteration) exceeds a threshold, we stop the clustering approach. The reason that we set up a max iteration threshold is to avoid the possible oscillation that a few samples will change their cluster_ids across certain clusters. Note that ideally, if the Web data set used to train the thesaurus is large enough, the low-level codewords will shrink and all environmental regions can be mapped to semantic codewords because each of them will be assigned a concept.

The structure of the learnt codebook is shown as the left part of Figure 7, which contains hierarchical semantic-level codewords and flat-structured low-level codewords.

3.4 Learning Image Region Correlation Matrix

Based on the codebook, a correlation matrix is learnt which measures the co-occurrence probabilities of any two codewords. It is used to determine which codewords will be selected to expand a query.

There are three kinds of correlations 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.

We use conditional probability to measure how likely a codeword would appear in an image given the existence of another (other) codeword(s). Let c_j denote the j^{th} codeword and $\Theta = \{c_i \,|\, 1 \le i \le N\}$ denote a set of codewords, where c_i denote the i^{th} codeword and N is the total number of image codewords learnt. Let I_k denotes the k^{th} image in image set I .

$$p(c_j \,|\, \Theta) = \frac{p(c_j, \Theta)}{p(\Theta)} = \frac{\sum_{I_k \in I} f(c_j, \Theta \,|\, I_k)}{\sum_{I_k \in I} f(\Theta \,|\, I_k)} \tag{3}$$

where

$$f\left(\Theta \,|\, I_k\right) = \begin{cases} 1, & \Theta \subseteq I_k \\ 0, & else \end{cases} \tag{4}$$

$$f\left(c_j, \Theta \,|\, I_k\right) = \begin{cases} 1, & \{\Theta, c_j\} \subseteq I_k \\ 0, & else \end{cases} \tag{5}$$

The function f is meant to reduce the effect of over-segmentation because current image segmentation algorithms often break an object into multiple regions.

We explore two kinds of correlations in this paper: first-order correlation and second-order correlation. For first-order correlation, $|\Theta| = 1$, which means Θ contains only one codeword. For second-order correlation, $|\Theta| = 2$, i.e. Θ is the collection of two codewords.

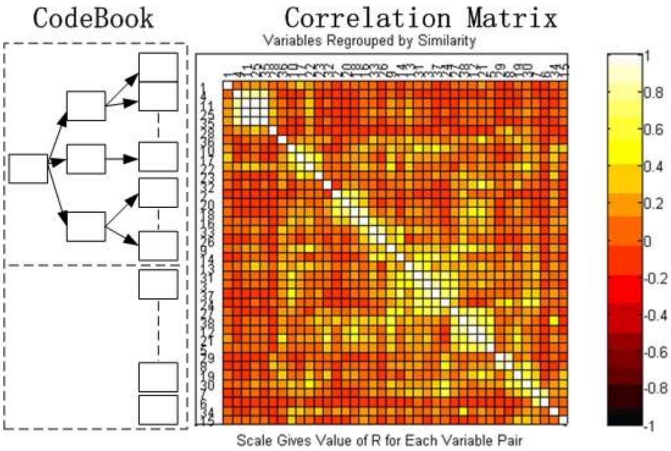

CodeBook
Correlation Matrix
Variables Regrouped by Similarity
Scale Gives Value of R for Each Variable Pair
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 7. The image thesaurus constructed from the Web image data

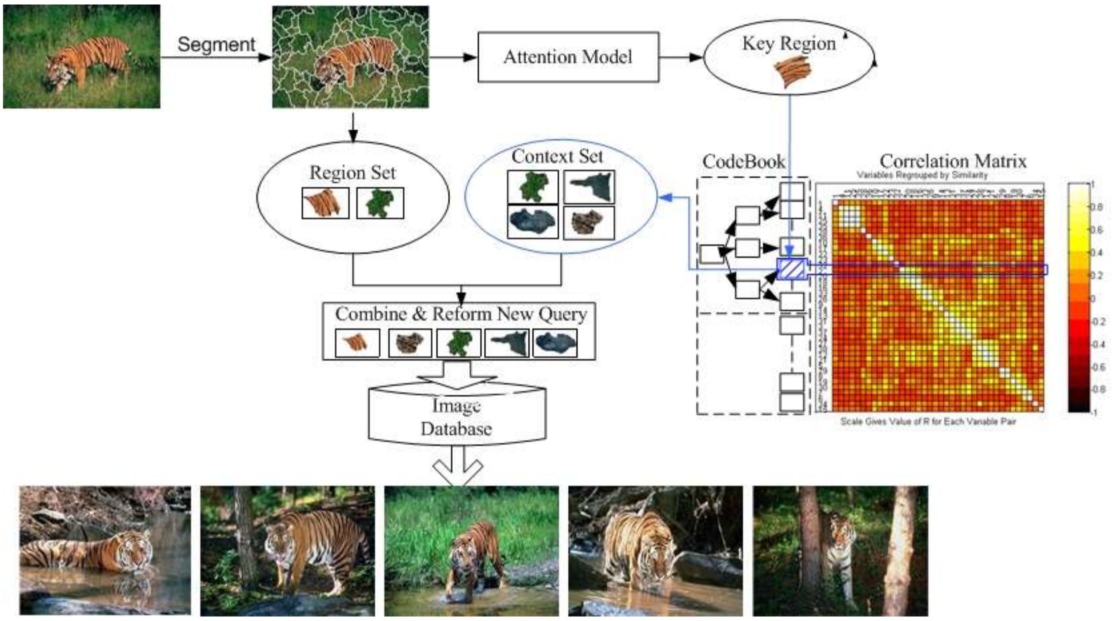

Segment
Attention Model
Key Region
Region Set
Context Set
CodeBook
Correlation Matrix
Variables Regrouped by Similarity
Combine & Reform New Query
Image
Database
Scale Gives Value of R for Each Variable Pair

Figure 8. The Context Expansion approach for region-based image retrieval

3.5 The Learned Image Thesaurus

The constructed image thesaurus is shown in Figure 7. It consists of the codebook and its associated correlation matrix. It has two functions: 1) mapping a query to semantic codeword(s) in the codebook 2) using the destined codeword(s) as an index (indices) to obtain a group of top-ranked correlated codewords from the correlation matrix. The extracted group of correlated codewords, when augmented to the original query, forms the new query.

4 IMAGE RETRIEVAL BY CONTEXT EXPANSION

The framework of our context expansion approach has three major components here: 1) a pre-constructed image thesaurus which converts each region in an image into an iconic representation (i.e. codeword) and measures the co-occurrence probabilities between codewords 2) an image attention model which identifies the key region in an image as presented in Section 3.2, and 3) the context expansion approach which makes use of the above two models to reformulate the query for retrieval.

As shown in Figure 8, when a query is submitted, it is first segmented into a set of homogeneous regions, from which the image attention model is adopted to find the key region. Then the key region is mapped to a codeword in the image thesaurus as an index to extract the set of codewords to which it highly correlates. These extracted codewords, as the blue lines shown in Figure 8, form the context set to expand the query codeword.

The original region set, combined with the extracted context set, forms the new query which is submitted to the retrieval system. Because the query is enriched by more context information, the images of tigers in various natural environments besides grass can be retrieved (see Figure 8). On the other hand, the correlation information indicates the context region patterns that the tiger is seldom associated with, which helps to filter irrelevant images and hence improves the recall performance.

Obviously, the proposed framework can support both query-by-example and query-by-keyword image retrieval schemes. In the case of query-by-keyword, if the keyword (e.g. “wolf”) maps to a semantic-level codeword (i.e. at the leaf node), 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 all the semantic-level codewords that are children of that immediate node. Note that these codewords are used as "OR" queries to retrieve images. That is, for each semantic-level codeword selected, we expand it by those highly correlated ones and thus construct a certain query. In such a way, we obtain a group of queries and each of them is used to retrieve the image database respectively. Then the sets of retrieved images are merged to a single result pool and images are re-ranked according to their similarities.

It is possible that the query keyword will not match any nodes in the semantic hierarchy. In this case, the synonyms of this query keyword given by WordNet will be used as queries one-by-one. Because we use the Web to train the thesaurus, when the training dataset is large enough, it is possible that any query keyword will match to a semantic-level codeword.

In the case of query-by-example, we first segment the query image using JSEG algorithm and modify the segmentation results as discussed in Section 3.2. Then visual features are extracted from all the resulted image regions. The key region is compared to all the codewords and is mapped to the one which is of minimum distance in the feature space. Based on the correlation matrix, we augment the query by including the feature vectors of highly correlated codewords indexed by the one associated to the key region.

4.1 The Context Expansion Approach

We summarize our Context Expansion approach in the query-by-example case as below:

Let $I = \{I_k \mid 1 \le k \le M\}$ be the image dataset, where M is the number of images. Let $C = \left\{c_i \mid 1 \le i \le N\right\}$ be the codebook and N is the number of codewords. Without loss of generality, the region-to-codeword mapping is defined as

$$\varphi : I \rightarrow C \ s.t. \ \varphi(r_j) = \arg\max_i p(c_i \mid r_j) \quad (6)$$

Where $r_j \in I$ denote a region in an image, $1 \le i \le N$.

Assume $c_i = \varphi(r_j)$. Let E_{c_i} be the set of codewords correlated to c_i, i.e. the possible context that c_i can coexist with. The equation to find E_{c_i} is defined as below:

$$E_{c_i} = \{c_k \mid p(c_k \mid c_i) > \theta,\ 1 < k < N\} \quad (7)$$

where θ is a threshold for controlling the scope of expansion.

We propose two kinds of expansion approach: the first-order expansion and the second-order expansion.

1) *First-Order Expansion*

Step 1: Context Set Extraction

1. Segment the query image into homogeneous regions

2. Identify the key region $r^{(d)}$ using the equation (2)

3. Map $r^{(d)}$ to a codeword c^* in the codebook by the equation (6)

4. Use c^* as an index to extract the context set E_{c^*} by the equation (7).

Step 2: Query Reformulation and Query-Expanded Retrieval

1. Combine the original query region set with the context set E_{c^*} in Step 1 to form the new query.

2. Submit the new query to the image database.

Note that we only use the key region as an index to select a set of context regions that are highly correlated with it. The reason is that if we also expand for those environmental regions, noises may be included.

2) *Second-Order Expansion*

In this approach, only the Step 1 in 1) is changed. Here the two regions with the largest saliency value (i.e. the key region and a second important region) are used together to index the second-order correlation matrix. Assume the codewords mapped by the two regions are c^* and c^{**}, the expansion term set E_c is given by

$$E_c = \left\{ c_k \mid p(c_k \mid c^*, c^{**}) > \theta,\ 1 \le k \le N \right\} \tag{8}$$

where $p(c_k \mid c^*, c^{**})$ is given by equation (3)-(5) with $\Theta = \{c^*, c^{**}\}$.

4.2 Similarity Measure

The similarity measure used in query-by-keyword search is the Jaccard coefficient [12]. 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:

$$Sim(I_i, Q_j) = \frac{\|A \cap B\|}{\|A \cup B\|} = \frac{\|A \cap B\|}{\|A\| + \|B\| - \|A \cap B\|} \tag{9}$$

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

$$Sim(I_i, Q) = \max_{j \in (1, \|Q\|)} Sim(I_i, Q_j) \tag{10}$$

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

We use the EMD [20] distance to compute the similarities in the query-by-example approach. The reason that we do not quantize the visual features of regions to codewords and then use the similarity measure of the query-by-keyword case is to keep the information provided by the query image because the quantization will cause energy loss.

5 EXPERIMENTS

To learn the image thesaurus, 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 three color moments, 36-bin color correlogram and three-level wavelet texture features which result in 171 dimensions. From these images, we constructed a codebook with 829 semantic-level codewords and 1,000 low-level codewords.

Two performance measures, precision and recall, are applied. Scope specifies the number of images returned to the user. Precision is defined as the number of retrieved relevant objects over the value of scope. Recall is defined as the number of retrieved relevant objects over the total number of relevant objects.

5.1 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.

Figure 9 shows the retrieval precision when scope is 10 (precision@10). From this figure we can see that the performance is satisfying. Those queries which perform worse (precision < 0.8) are either collective nouns (e.g. bird) or ambiguous (e.g. shell, news). The fact that collective nouns do a worse job than proper nouns proves the effectiveness of our approach: that we delete those key regions whose associated key terms do not map to leaf nodes in the hierarchical codebook. Such key regions are generally noisy.

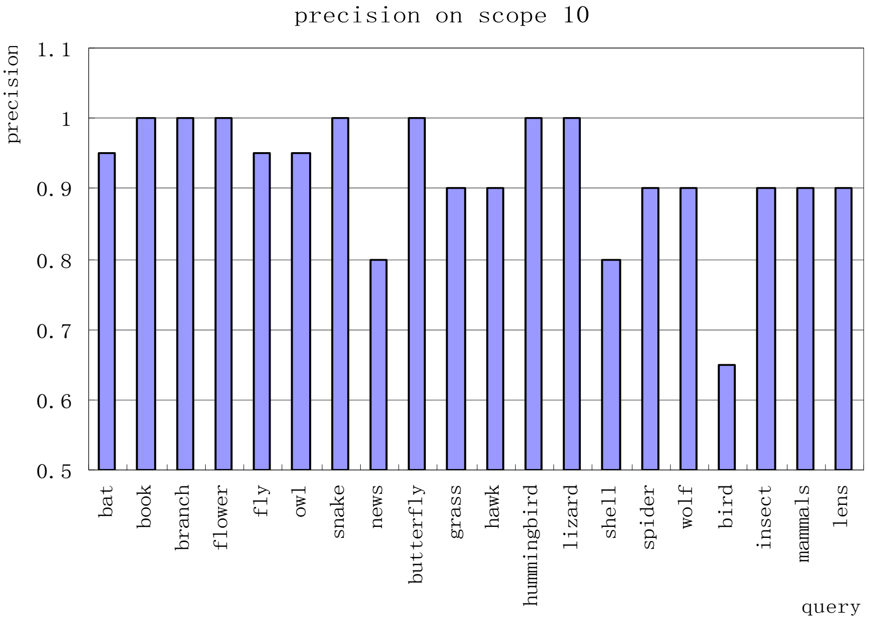


Figure 9 Precision@10 for image retrieval based on key term extraction.

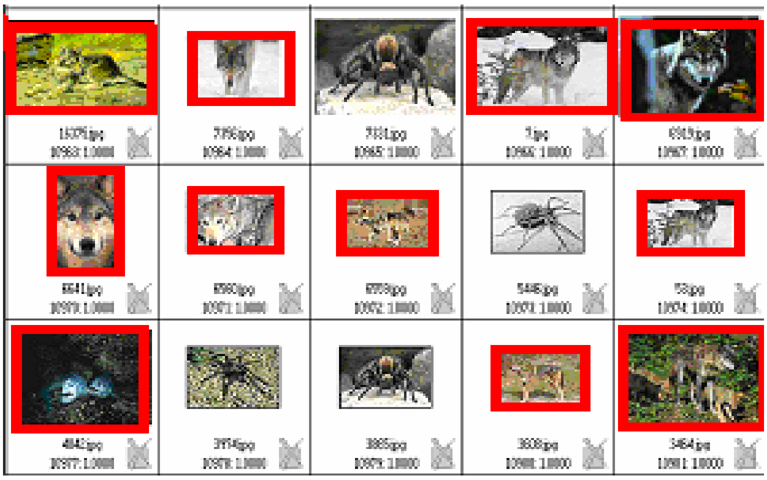


Figure 10 Retrieval result of query "wolf"

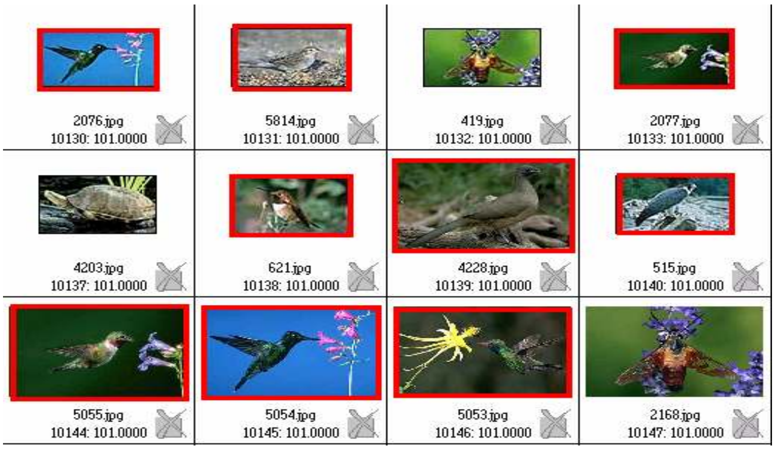


Figure 11 Retrieval result of query "bird"

5.2 Experiments on Query-By-Keyword Retrieval

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.

These experiments are performed in order to examine the effectiveness of the learnt thesaurus in supporting this query scheme. Figure 10 shows the result of query “wolf”. The images in red box are correct hits. Note that this example shows the capability of retrieving images by high-level concept. Figure 11 shows the result of query “bird” which corresponds to the situation when the query does not map to a leaf node but an immediate node in the semantic hierarchy. In this case, all semantic-level codewords whose father is “bird” are used to form the query set.

5.3 Experiments on Query-By-Example Retrieval

We use the Corel Stock Photo Library as our testing image database. 10,000 images (containing 206,115 regions) from 90 categories of the Corel database, either natural or artificial, are used for performance evaluation. These images do not overlap with our training images obtained from the Web, but they cover similar high-level concepts 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.

The baseline method we used is the traditional region-based image retrieval approach using EMD [20] distance measure. Five groups of queries, each containing 100 images, are randomly selected.

5.3.1 First-Order Context Expansion

Figure 12 shows the performance of our method vs. the baseline method (precision/recall at the scope of 10). The blue bars represent the performance of the baseline method and the red bars represent our method. The maximum expansion length is 2 and correlation probability threshold $\theta = 0.001$.

It can be seen that on these five query sets, both precision and recall are greatly improved by our context expansion approach ($p < 0.05$ on T-TEST) except query set 3.

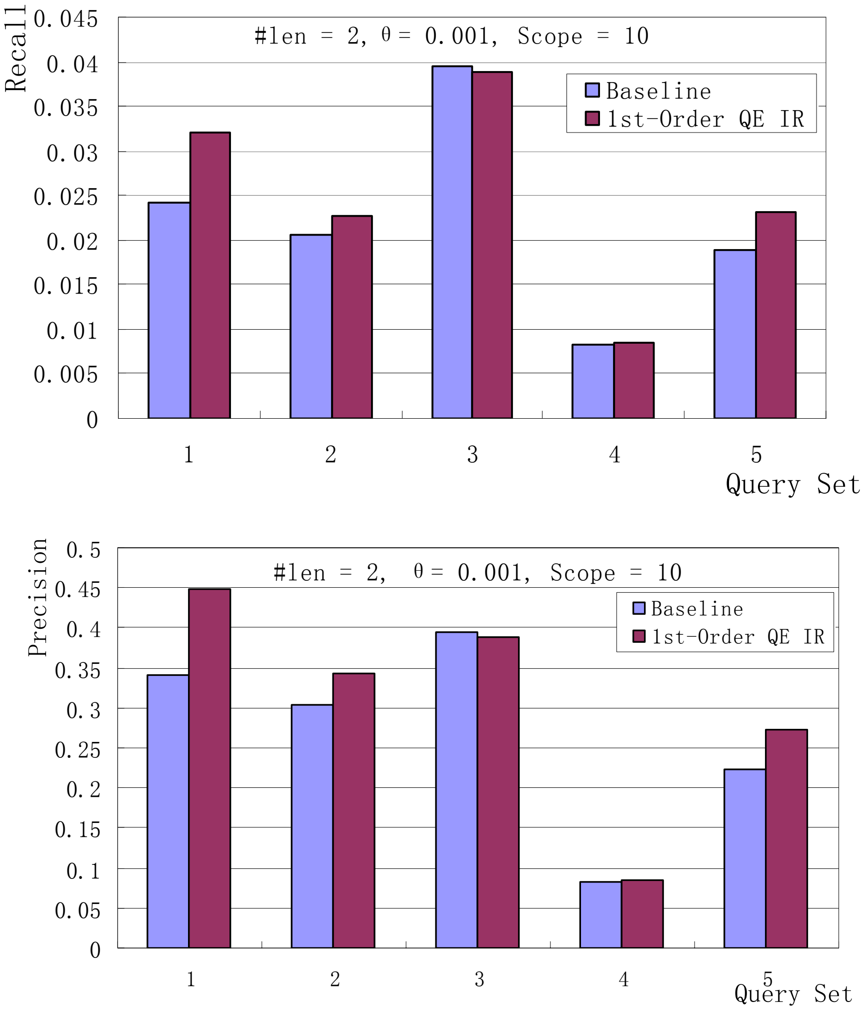


Figure 12. Average precision and recall for the 5 query sets, each containing 100 queries

5.3.2 Second-Order Context Expansion

Now we explore the effect of second-order context expansion on image retrieval. In Figure 13, it can be seen that although the second-order expansion outperforms the baseline method, it still performs worse than the first-order expansion. A possible reason is that the second-order correlation matrix is very sparse (the ratio of the number of non-zero items to the total number of items on this query set is 47.49%), hence only a very small subset of queries are indeed expanded. On the other hand, the performance of the second-order expansion is also affected by the size of the codebook.

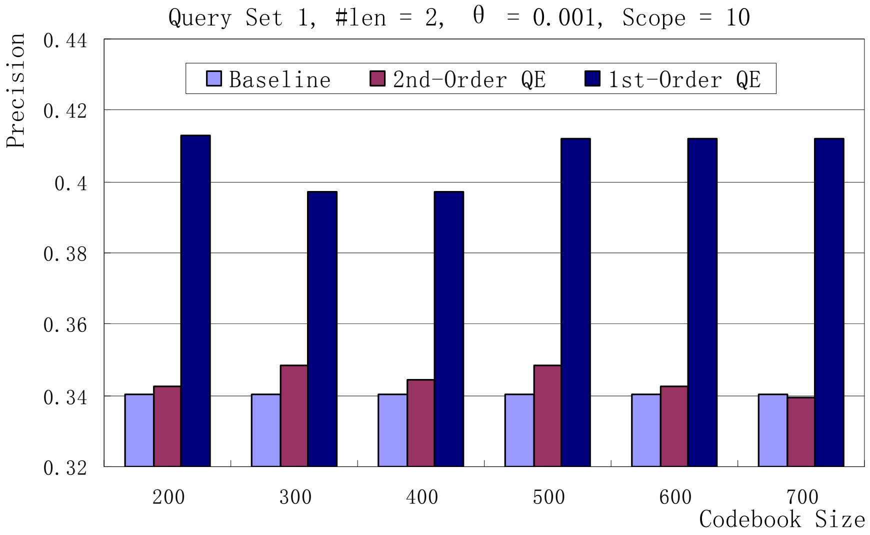


Figure 13. Comparison of the baseline method, first-order expansion and second-order expansion

5.4 Parameter Selection

The parameters need to be evaluated are:

1) the (low-level) codebook size k (see Section 3.3 kmeans clustering approach)

2) the maximum expansion length *len*

3) the correlation probability threshold θ

len and θ together determine the final expansion length for a query image. That is, $\left\|E_{c_i}\right\| \le len$, and the selected codewords have correlations larger than θ with the key region of the query.

5.4.1 Effect of Codebook Size

Now we investigate why for query set 3 the experiment shown in Figure 12 did not outperform the baseline method.

Figure 14 shows the average region number of each query set. From set 1 to 5, the average number is 22.91, 20.52, 28.09, 18.36, and 21.72, respectively. It can be seen that the 4th query set contains the fewest average number of regions while the query set 3 has the largest. And the average region number in the query set 1, 2 and 5 are very close (22.91, 20.52, 21.72).

We suspect that the ratio of average query region number to codebook size may affect retrieval performance. Figure 15 shows the curve of retrieval precision vs. codebook size (from 200 to 2000) on query set 1. As can be seen, the performance reaches its peak when codebook size is 1000. We tuned the codebook size on the query set 3 and 4, and found that on query set 3, when the codebook size is 1600, our method achieves best precision performance (precision 41%, recall 4.1%), which is better than the baseline method (precision 39.5%, recall 3.95%). On query set 4, the optimal codebook size is 900.

From the analysis above, we found that all the query sets achieve their best performance (precision) when the ratio of average query region number to codebook size is around 2%.

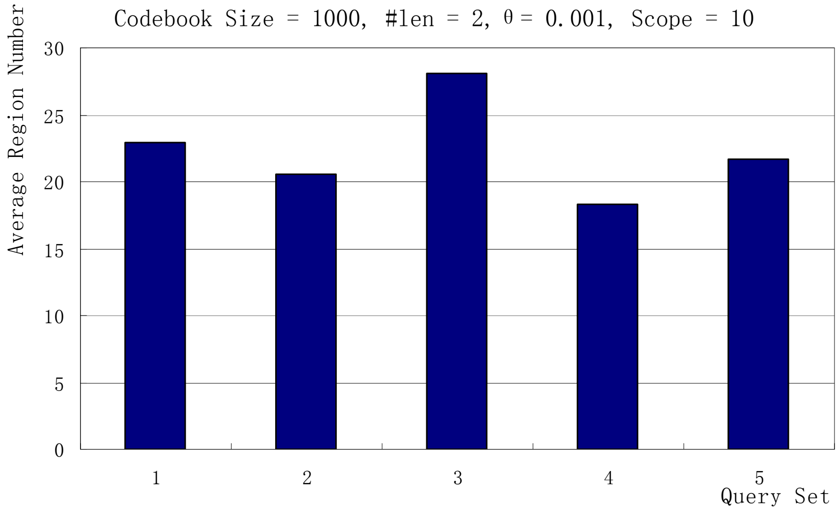


Figure 14. Average region number of the six query set.

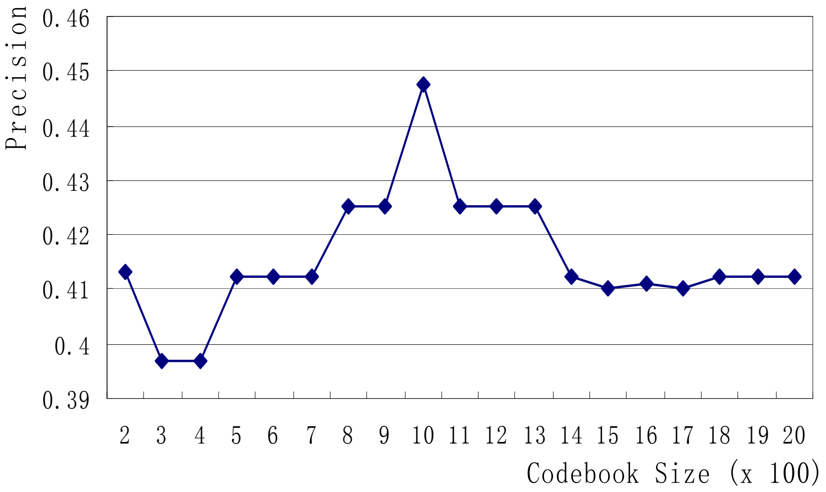


Figure 15. Effect of Codebook Size, *len*=2, *θ* =0.001

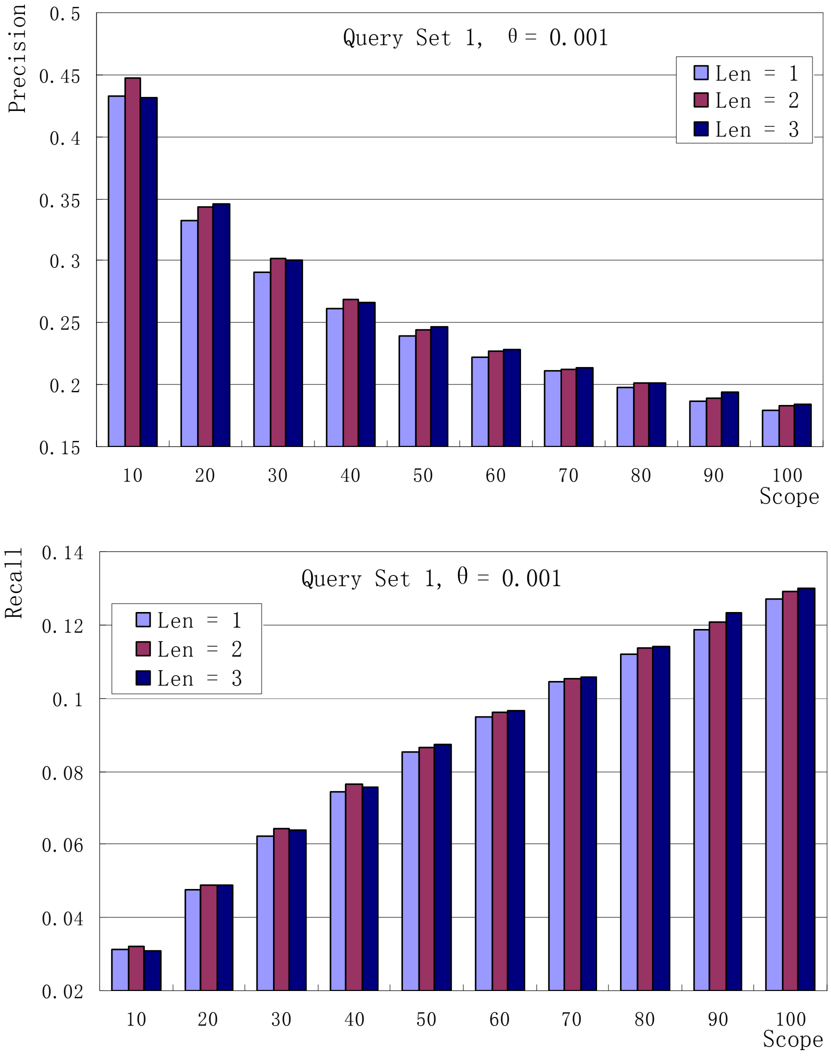


Figure 16. Effect of Expansion length. The best performance is obtained when len = 2, query set 1, codebook size = 1000.

5.4.2 Effect of Expansion Length

Another factor which may influence the overall performance is the choice of expansion length. By expansion length, we mean how many low-level codewords are selected from the correlation matrix to expand the query. Expansion length is also an intrinsic problem of query expansion technology. Figure 16 shows the curves of precision and recall of our method on query set 1 vs. scope. Note that the expansion length in the figure means the maximum expansion length *len*. The actual expansion length $\|E_{c_i}\|$ is determined by *len* and θ simultaneously.

According to the experimental results, it can be seen that the best performance is achieved when each query image is expanded by at most two codewords at the scope of 10, but larger expansion length (e.g. three) is preferred when the number of retrieved images (i.e. scope) increases.

6 DISCUSSIONS

1) In our proposed approach, we assume that each image has only one semantically important region, and the attention model used in our approach tends to model an attention area as one continuous region hence is suitable for our assumption. Although our current training dataset is consistent with this assumption, for future large-scale image training database, we should not ignore the existence of multi-concept images. In this case, we can take advantages of discrete salient point identification models [11][15] and extract those salient points which fall into the key regions. We will discuss this in our future work.

2) Currently we adopt a heuristic key-term extraction scheme although we can take

advantages of more advanced techniques (e.g. linguistic models in natural language processing area). This is because the key term extraction is a component for our framework but not the main goal. Although a more complex and accurate model can improve our performance, we prefer to propose the idea of leveraging the abundant and valuable information embedded in web-pages.

3) In the case of retrieval-by-example, the extra computational overhead results from a) key region extraction; b) mapping to a semantic-level codeword and c) context expansion. In fact, the time expense on extracting key region can be ignored (the attention model [21] processes images in millisecond level). So does the context expansion approach because we use a hash map to index all the codewords and their correlated ones. The time cost for mapping the key region of query image to a semantic-level codeword is proportional to the size of the semantic-level codewords. Because today's computer hardware has become so fast that for a database of 10,000 images, one retrieval iteration costs only a few seconds, and since our codebook is much smaller, this step does not bring large time expense on our method. In short, the online computational overhead is nearly no difference to traditional content-based image retrieval methods. In the case of query-by-keyword case, because the hierarchical codebook is also indexed by a hash map, the computational overhead can also be ignored.

4) The thesaurus is fundamentally a compressed storage of image concepts --- the codewords are cluster centroids and only the leaf nodes of the hierarchical codebook are associated with low-level features. Hence the storage requirements are also acceptable.

5) Relevance feedback approach can be easily integrated into our framework.

7 CONCLUSION

In this paper, we explore the latent statistical correlations embedded in images and expand the user submitted queries leveraging such information to improve retrieval performance. Web images are used as training data for both bridging the cognitive gap and learning the statistical correlation matrix between concepts. Experimental results show the effectiveness of our context expansion approach.

As shown in this paper, the quality of image thesaurus has much impact on the overall performance of our context expansion approach. In fact, the hyperlinks between Web images are valuable information to be used 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 in our future works.

8 REFERENCES

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

[2]. E. Chang, K. Goh, G. Sychay, and G. Wu, "CBSA: Content-based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines", *IEEE Transactions on CSVT Special Issue on Conceptual and Dynamical Aspects of Multimedia Content Description*, Volume 13, Number 1, January 2003, pp.26-38

[3]. 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

[4]. H.J. Zhang and Z. Su, "Improving CBIR by Semantic Propagation and Cross-Mode Query

Expansion", *Multi-Media Content Based Indexing and Retrieval*, 2001

[5]. J. Jeon, V. Lavrenko and R. Manmatha, "Automatic Image Annotation and Retrieval using Cross-Media Relevance Models", In *26th Annual International ACM SIGIR Conference*, July 28-August 1, 2003, Toronto, Canada

[6]. F. Jing, M.J. Li, H.J. Zhang, and B. Zhang, "Support Vector Machines for Region-Based Image Retrieval", *In. IEEE International Conference on Multimedia & Expo*, July 6-9, 2003, Baltimore, Maryland

[7]. F. Jing, M.J. Li, H.J. Zhang, and B. Zhang, "An Efficient and Effective Region-based Image Retrieval Framework", *to appear in IEEE Transaction on Image Processing*

[8]. K. Barnard , P. Duygulu, D. Forsyth, N. Freitas, D.M. Blei and M. Jordan, "Matching Words and Pictures", *Journal of Machine Learning Research*, 2003, vol 3, pp 1107-1135

[9]. K. Barnard, P. Duygulu, and D. Forsyth, "Clustering Art", *Computer Vision and Pattern Recognition*, 2001, pp. II:434-439.

[10]. K. Porkaewand, and S. Mehrotra, "Query Reformulation for Content Based Multimedia Retrieval in MARS", *Technical Report TR-MARS-99-05*, Univ. of California at Irvine, 1999

[11]. K. Mikolajczyk, and C. Schmid, "Scale and Affine Invariant Interest Point detectors", *International Journal of Computer Vision*, 60(1), 2004

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

[13]. R.K. Srihari, "Use of Multimedia Input in Automated Image Annotation and Content-Based Retrieval", *Storage and Retrieval for Image and Video Databases*, 1995, pp. 249-260.

[14]. S. Tong, E. Chang, "Support Vector Machine Active Learning For Image Retrieval", *In ACM*

International Conference on Multimedia, October, 2001, Ontario, Canada.

[15]. T. Kadir, “Scale, Saliency and Scene Description”, *Ph.D thesis*, Oxford University, 2002

[16]. W.Y. Ma and B. S. Manjunath, “Netra: A toolbox for navigating large image databases”, *In International Conference on Image Processing*, October 26-29, 1997, Washington DC, USA.

[17]. M.E.J. Wood, N.W Campbell., and B.T. Thomas, “Iterative Refinement by Relevance Feedback in Content-Based Digital Image Retrieval”. *In ACM International Conference on Multimedia*, 1998, Bristol, UK

[18]. X.J. Wang, W.Y. Ma, and X. Li, “Data-Driven Approach for Bridging the Cognitive Gap in Image Retrieval”, *IEEE International Conference on Multimedia and Expo*, June 27-30, 2004, Taipei, Taiwan

[19]. Y. Deng, and B.S. Manjunath, “Unsupervised Segmentation of Color-Texture Regions in Images and Video”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, 23(8): 800-810

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

[21]. Y.F. Ma, and H.J. Zhang, “Contrast-based Image Attention Analysis by Using Fuzzy Growing”, *In ACM International Conference on Multimedia*, November 2003, Berkeley, CA USA

[22]. X.S. Zhou, T.S. Huang, “Unifying Keywords and Visual Contents in Image Retrieval”, *IEEE Multimedia*, 2002, 9(2): 23-33

[23]. L. Zhu, A.B. Rao and A.D. Zhang, “Advanced Feature Extraction for Keyblock-Based Image

Retrieval”, *Information Systems*, December 2002, 27(8):537 - 557