Web Image Retrieval via Learning Semantics of Query Image

Chuanghua Gui, Jing Liu, Changsheng Xu, Hanqing Lu

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Science, Beijing 100190, China
{chgui, jliu, csxu, luhq}@nlpr.ia.ac.cn

ABSTRACT

The performance of traditional image retrieval approaches remains unsatisfactory, as they are restricted by the well-known semantic gap and the diversity of textual semantics. To tackle these problems, we propose an improved image retrieval framework when querying with an image. The framework considers not only the discriminative power of various visual properties but also the semantic representation of the query image. Given a query image, we first perform CBIR to obtain some visually similar image sets corresponding to different visual properties separately. Then, a semantic representation to the query image is learnt from each image set. The semantic consistence among the textual indexes of each image set is measured in order to judge the confidence of various visual properties and the obtained semantic representation in search. Obtaining these items, both visually and semantically relevant images are returned to the user by a combined similarity measure. Experiments on a large-scale web images demonstrate the effectiveness and potential of the proposed framework.

Index Terms—web image retrieval, semantics learning, feature selection.

1. INTRODUCTION

With the advance of digital devices and Internet techniques, the number of images has exploded rapidly. Given the expanding image data, the capability to support efficient and effective image retrieval has become increasingly important and necessary.

There are two types of image retrieval approaches. One is the content-based image retrieval (CBIR), as a means of searching visually similar images given a query image. As we know, the semantic gap between the low-level visual description and the high-level semantics has become a major obstacle to CBIR. From the example shown in Fig.1 (a), most of the results are not relevant to "apple" when only an "apple" query image is submitted. Some work introduces relevance feedback from the users to get better results [1]. However, as images are described only with visual features, it is hard to ensure the relevance from semantic perspective. That is, it cannot fundamentally solve the semantic gap issue.

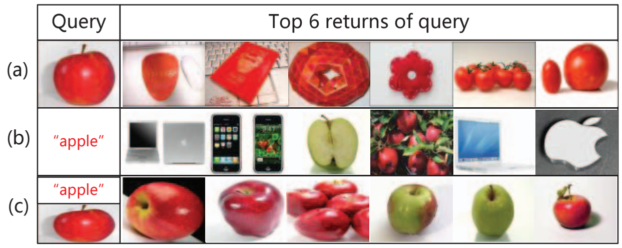

Fig.1 Top 6 results of three different retrieval methods: (a) CBIR, (b) TBIR, (c) Combined search.

The text-based image retrieval (TBIR) is the other approach. The TBIR requires annotating each web image and searches semantically relevant images given a text query. Due to the query polysemy, the results always contain multiple topics and they are mixed together. As shown in Fig.1(b), when querying with a keyword of "apple", the result images include images of "apple notebook", "apple fruit", "apple iPhone" and "apple's logo". Based on this view, IGroup [2] attempts to cluster the resulted images according to their semantics. It first identified several key phrases related to a given query, and assigned all the resulted images to the corresponding phrases. Ding et al. [3] further improved IGroup by clustering the key phrases into semantic clusters. However, both methods tend to make users puzzled because too many phrases are given, which make the results really diverse.

As mentioned above, the semantic gap and the diversity of textual semantics influence the performance of traditional CBIR and TBIR. To tackle these problems, an ideal method is to query with a keyword and an image together as shown in Fig. 1(c). However, it is rigorous and unpractical for users to input an image together with a keyword when they are performing search. In this paper, we propose an improved image retrieval framework via learning semantics of query image. The learned semantics combined with the visual features enrich the traditional representation of the query image and bring more special and relevant image results in search. Besides, considering that different visual features have varying discriminative power under a certain semantic context, we apply a statistical scheme to decide different confidence of the visual features.

978-1-4244-4291-1/09/$25.00 ©2009 IEEE

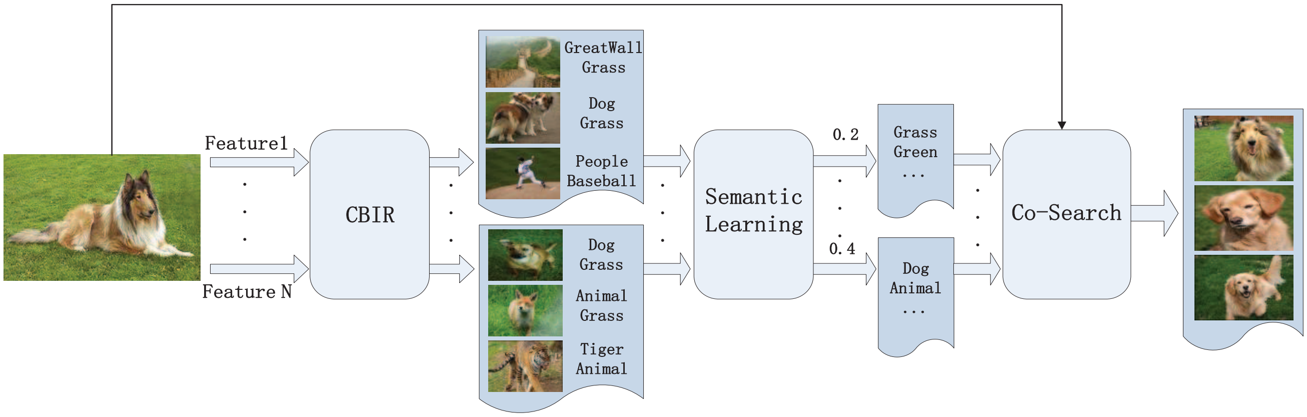

Fig.2 Illustration of our framework.

The rest of the paper is organized as follows. The framework is introduced in Section 2. The details of our method are presented in Section 3, 4, 5 respectively. The experimental results are reported in Section 6. The paper is concluded in Section 7.

2. OVERVIEW

In this paper, we unite the confident visual property and the descriptive semantics to search relevant images on the web. The framework of the proposed solution contains three main stages: the CBIR stage, the semantic learning stage and the co-search stage, which are illustrated in Fig. 2.

In this section, we will overview the proposed framework. First, we perform CBIR to obtain some visually similar image sets corresponding to different visual features respectively. Second, we explore each image set to learn the semantic description of the query image and decide its confidence by the semantic consistence within the textual indexes of the image set. Simultaneously, the discriminative power of various visual features is measured with the consistence. More semantic consistence indicates more confidence subject to the visual feature of the set and accordingly more weights should be given to the corresponding semantic description. Finally, we present a co-search process by designing a vision-and-semantics combined formulation to find more relevant images to the query image. The detailed introduction of the three sequential stages will be given in the following sections.

3. CBIR

As visual features are of high dimension generally, the similarity-oriented search based on visual features is always a bottleneck on the search efficiency for the large-scale image database. To solve this problem, we adopt Locality Sensitive Hashing (LSH) [4] to speed up the process.

LSH is an approximation method, which addresses the similarity match problem in sub-linear time. LSH uses a hash function h to divide the images in a database into bins. All the images with the same output value for the hash function are placed in a single bin. Given a query image, the hash function is applied to it and is mapped into a bin depending on the output value. Only the images in the same bin are retrieved as the results of the query and thereby the performance is improved.

Here, the hash function h is defined as

$$h_{m,n}(V) = \left\lfloor \frac{mV+n}{W} \right\rfloor \qquad (1)$$

where V is a d-dimensional original feature vector and W defines the quantization of the features, m is a d-dimensional random vector with entries chosen independently from a Gaussian distribution and n is a real number chosen uniformly from the range of [0, W].

As different features have varying discriminative power to describe an image. It is unclear which descriptors are more appropriate and where users take more attention. For example, when a user inputs a scenery image with simple background, a global feature such as color histogram is adequate for CBIR. But it is not always sufficient when the user wants to find an object in images with various backgrounds, such as a tiger on the grass or a cat in a room. In this case, the local feature maybe a better choice. Therefore, we use different features to perform CBIR separately in order to find the best representation to describe the image. Obtaining the resulted image sets corresponding to different types of features, we learn the semantic representations of the query image respectively. The learning process will be introduced in Section 4.

4. SEMANTIC LEARNING

When a user searches images with a query image, he/she expects to find the similar ones on both visual appearance and semantics. To achieve this, we try to learn the semantics of the query image from the resulted images by CBIR.

There are a lot of algorithms applied to extract the semantic representation of an image set, such as DCMRM [5] and BGRM [6], which not only extracted candidate annotations from its surrounding text and other textual information, but also expanded and refined them by exploring the word correlation. However, it took much time on learning, which was not suitable for on-line image retrieval. From this view, we use a new scheme modified from the page search results clustering method (PSRC) [7].

Given the ranked list of the results (containing images and their textual information) returned by CBIR using the feature f_i, we first extract all candidates from the texture information, and then calculate several properties for each candidate such as phrase frequencies, document frequencies, and more. A regression learning model from training data is applied to combine these properties into a salience score. The candidates are ranked by the salience scores, and the top-ranked candidates can be constructed into a vector as the textual description of the result image set (R_i) corresponding to the feature f_i. Naturally, the semantic representation (q_i) of the query image to the set is obtained.

Besides, the semantic consistence of each result image set is also calculated to measure the confidence of different visual features and the semantic representation to the query image. The higher the semantic consistence score is, the more confident they are to describe the subject of the query image with this feature and therefore the corresponding semantic representation should be given more weights. For clarity, we denote the semantic consistence of the result set (R_i) as c_i, and detail its calculation as follows.

First, we extract a collection of words from the textual information of each image in the set and calculate the similarity between two word collections for any image pair (I_a, I_b) in R_i as:

$$g(W_a, W_b) = \frac{1}{N_a N_b} \sum_{j=1}^{N_a} \sum_{k=1}^{N_b} exp[-\gamma \cdot NGD(w_j, w_k)] \quad (2)$$

where γ is an adjustable factor, W_a and W_b are the word collections of the images I_a and I_b, N_a and N_b are their sizes, and w_j and w_k are words in the two collections. In practice, we can also average maximal K similarities instead of all word-word similarities (as in Eq.2) to reduce the influence of noise. In addition, $NGD(w_j, \mathrm{w}_k)$ is the Normalized Google Distance [8] defined as:

$$\mathrm{NGD}(\mathrm{w}_\mathrm{j}, \mathrm{w}_\mathrm{k}) = \frac{\max\{\log \mathrm{f}(\mathrm{w}_\mathrm{j}), \log \mathrm{f}(\mathrm{w}_\mathrm{k})\} - \log \mathrm{f}(\mathrm{w}_\mathrm{j}, \mathrm{w}_k)}{\log \mathrm{G} - \min\{\log \mathrm{f}(\mathrm{w}_\mathrm{j}), \log \mathrm{f}(\mathrm{w}_\mathrm{k})\}} \quad (3)$$

where G is the total number of web pages indexed by Google, $f(\mathrm{w}_j)$ is the count of pages where word w_j appears, and $f(\mathrm{w}_j, w_k)$ is the count of pages where both w_j and w_k appear.

Then, we repeat the above calculation to get the similarities for every image pair in the result set.

At last, we denote the similarities of each image and the other images as a vector and the variance over all these vectors can be used to measure the semantic consistence of each image set, which is given as:

$$c_i = {}^{1}/_{\sigma_n} \quad (4)$$

where σ_n is the standard deviation of the vectors.

Just as mentioned above, for the image sets with different visual representations, the semantic consistencies are considered as the measures that are positively relevant to the confidence of their semantic descriptions and the corresponding visual features.

5. CO-SEARCH

To get relevant images both on visual and semantic levels, a linear fusion is applied to combine the ranked list of TBIR and CBIR, which is formulated as:

$$S_{final} = S_{TBIR} + \alpha * S_{CBIR} \quad (5)$$

where α is a parameter to leverage the roles of the semantic similarity and the visual similarity. That is, bigger α gives more confidence on the visual similarity while less on the semantic relevance, and vice versa.

For the process of TBIR, we rank the images with the similarities between the query keywords and the textual descriptions of these images. As mentioned in Section 4, the semantic representation is considered as the textual query to perform TBIR for each type of features. Then, the S_{TBIR} is formulated as a weighted ranking score according to the different confidence from each type of visual features, which is defined as:

$$S_{TBIR} = \sum c_i S_{q_i} \quad (6)$$

where c_i is the semantic consistence corresponding to the feature f_i, and S_{q_i} is the score computed by the ranking function BM25 [9].

Similarly, the S_{CBIR} is formulated as a weighted ranking score through performing CBIR with different visual features and its definition is given as:

$$S_{CBIR} = \sum c_i S_{\mathrm{f_i}} \quad (7)$$

where c_i is the semantic consistence and $S_{\mathrm{f_i}}$ is the similarity of the object image and the query image under feature f_i.

6. EXPERIMENT

All the data used in our experiments are crawled by searching on Google and Flickr. We select 1000 popular keywords as queries. For each query, 1000 top-ranked images are crawled and their corresponding web pages are also downloaded. With an HTML parser which depends on DOM-tree structure, the textual information of each image, which includes the words in title, URL, ALT tag, anchor text and surrounding text, is extracted for the image indexing on semantic level. For each image, three types of visual features are extracted, including 144-dimensional Color Correlogram, 24-dimensional Polynomial Wavelet Tree (PWT) and 36-dimensional Color Histogram.

To evaluate the performance of our system, the mean average precision (MAP) is employed, which is also widely used by the image retrieval community. For each query, we first compute the average precision (AP) and average them to obtain MAP. AP is a common metric in information retrieval that measures precision at all depths of a search process and averages all measurements up to a given depth. Given a query and k relevant results, let $rank_i$ be the rank of the *i-th* retrieved relevant result, then average precision is defined as follows:

$$AP = \frac{1}{k} \sum_{i=1}^{k} \frac{i}{rank_{\ i}} \quad (7)$$

6.1 Parameter setting in Co-Search

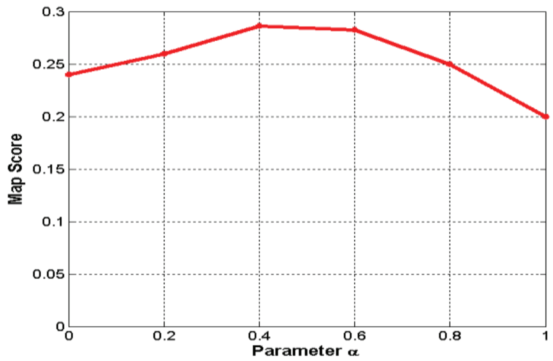

Fig.3 The effect of the parameter α.

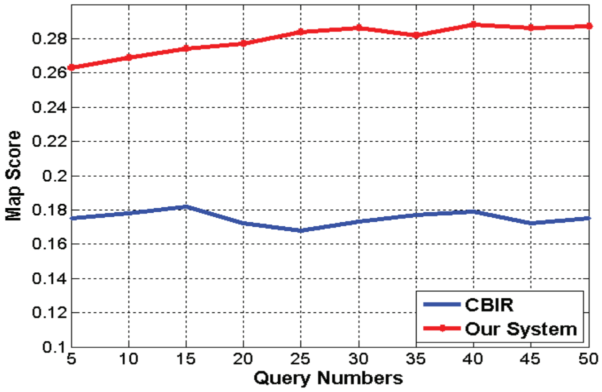

Fig.4 Search performance comparison.

When we combine CBIR and TBIR to retrieve, the parameter α dominates the weights of their scores and finally influences the precision of results. Fifty query images are randomly selected to test the effect of different parameter α. From Fig.3, we can see that 0.4 is the best choice for the parameter α. Too small α or too large α will lead to lower visual or semantic similarity between the results and the query image.

6.2 Performance of our system

Ten participants are asked to evaluate the performance of CBIR and our system with arbitrary queries they liked. The parameter α is set to 0.4 as default. The experimental results shown in Fig.4 demonstrate that our system performs far better than CBIR. This is because we make a great effort on learning the semantics of the query image, which is closer to the needs of users. Fig.5 presents some search examples by CBIR and our system respectively.

7. CONCLUSIONS

In this paper, a novel image retrieval method via learning semantics of query image is described. The method has the following advantages. First, we perform the CBIR process with consideration of different visual features. Second, we present a scheme to learn the semantic representations of the query image, which respectively correspond to the result sets of CBIR using different visual features, and their confidences are evaluated by the semantic consistence within the set. Third, the confident semantic representation and visual features are associated to return the image results with great visual and semantic similarities. Finally, the reasonable and comprehensive evaluations are performed to demonstrate the effectiveness of the framework.

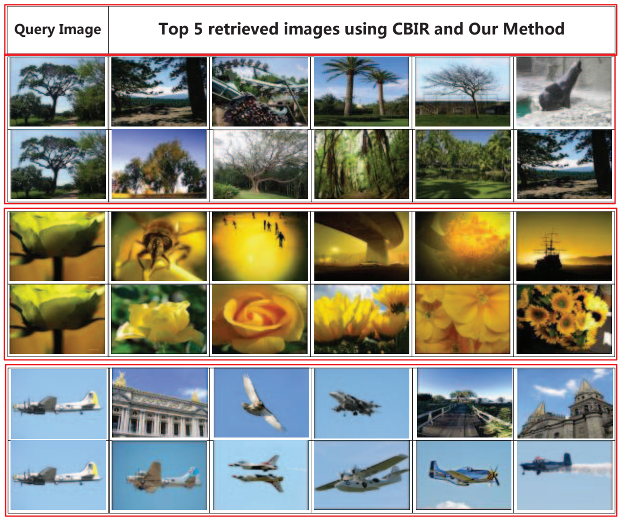

Fig. 5 The search result list by traditional CBIR (given in the 1-st, 3-rd and 5-th rows) and our method (given in the 2-nd, 4-th and 6-th rows) respectively.

8. ACKNOWLEDGEMENTS

The research was supported by National 863 Project (2006AA01Z315), and National Natural Science Foundation of China (60675003, 60833006).

9. REFERENCES

[1] Rui, Y., Huang, T.S., Ortega, M. and Mehrotra. *Relevance feedback: A power tool in interactive content-based image retrieval.* IEEE Trans. on Circuits and Systems for Video Technology, 1998.

[2] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, W.Y. Ma. *IGroup: Web Image Search Results Clustering.* ACM Multimedia, 2006.

[3] Haoyang Ding, Jing Liu, Hanqing Lu. *Hierarchical Clustering-Based Navigation of Image Search Results.* ACM Multimedia, 2008.

[4] Mayur Datar, Nicole Immorlica, Piotr Indyk, Vahab S. Mirrokni. *Locality-sensitive hashing scheme based on p-stable distributions*, Proceedings of the twentieth annual symposium on Computational geometry, 2004.

[5] J. Liu, B. Wang, M. Li, Z. Li, W.-Y. Ma, H. Lu and S. Ma. *Dual cross-media relevance model for image annotation*. ACM Multimedia, 2007.

[6] X. Rui, M. Li, Z. Li, W. Ma, N. Yu. *Bipartite Graph Reinforcement Model for Web Image Annotation*. ACM Multimedia, 2007.

[7] H.J Zeng, Q.C He, Z. Chen, W.Y. Ma, J. Ma. *Learning to Cluster Web Search Results*. ACM SIGIR, 2004.

[8] R. Cilibrasi and P. M. B. Vitanyi. *The Google similarity distance*. IEEE Transactions on Knowledge and Data Engineering, 2007.

[9] Robertson SE, Walker S, Beaulieu MM, Gatford M, and Payne A. *Okapi at TREC-4*. The 4th Text Retrieval Conference, 1996.