Image Annotation Using Search and Mining Technologies[1]

Xin-Jing Wang, Lei Zhang, Feng Jing, Wei-Ying Ma
Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, China
xjwang@msrchina.research.microsoft.com, {leizhang, fengjing, wyma}@microsoft.com

ABSTRACT
In this paper, we present a novel solution to the image annotation problem which annotates images using search and data mining technologies. An accurate keyword is required to initialize this process, and then leveraging a large-scale image database, it 1) searches for semantically and visually similar images, 2) and mines annotations from them. A notable advantage of this approach is that it enables unlimited vocabulary, while it is not possible for all existing approaches. Experimental results on real web images show the effectiveness and efficiency of the proposed algorithm.

Categories and Subject Descriptors
I.4.8 [Image Processing and Computer Vision]: Scene Analysis – object recognition. H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval –search process.

General Terms: Algorithms, Performance.

Keywords: Image annotation, search result clustering, hash indexing.

1. INTRODUCTION
Image annotation nowadays is still far from practical and satisfactory given so many computer vision and machine learning approaches. Possible reasons are: 1) it is still unclear how to model the semantic concepts effectively and efficiently; 2) the lack of training data to bridge effectively the semantic gap.

With the explosive development of the Web, it has become a huge resource of all kinds of data and has brought about possible solutions to many problems that were believed to be "unsolvable." In this paper, we leverage the huge number of images existing on the Web and propose a novel idea for image auto-annotation. The key idea is to find a group of similar images both semantically and visually, extract key phrases from their textual descriptions, and select the highest-scored ones to annotate the query image.

To by-pass the semantic gap, an accurate keyword is assumed initially associated with the query image, and we call it query keyword. This requirement is not as lacking in subtlety as it may first seem, e.g., in desktop photo search, location or event names are usually provided as folder names. Or in Web image search, a surrounding keyword can be chosen as the query.

A notable advantage is that the proposed approach is entirely unsupervised. No supervised learning approach is required to train a prediction model as a traditional approach does. And as a direct result, this method has no limitations on vocabulary, making it fundamentally different from the previous works.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

2. ANNOTATING IMAGE BY SEARCH AND MINING
The entire process is as this: given one query image and one query keyword, a text-based search is first employed to retrieve a group of semantically similar images. Then content-based image retrieval is adopted based on the selected images to pick up those visually similar ones, and rank them accordingly. To speed up this step, a hash coding-based algorithm is adopted to map the high-dimensional visual features to hash codes. Then key phrases are mined from the textual annotations of the top N ranked images using a clustering approach. Finally, after removing the duplicates, the rest phrases are output as the predicted annotations.

Below we detail three key techniques of this approach.

2.1 Hash Coding Algorithm
We modified the algorithm proposed by Wang et al. [2] to encode the image visual features to hash codes. Firstly, images are divided into even blocks and 36-bin color Correlograms [1] to represent each block. Then the features are transformed by a PCA mapping matrix learned beforehand, and quantized into 32-dimension hash codes. The quantization strategy is that if a feature component is larger than the mean of this vector, it is quantized to 1, otherwise to 0.

2.2 Distance Measures
Three distance measures are proposed and compared.

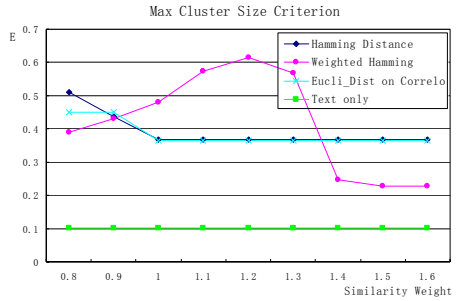
1) **Hamming distance**. It measures the number of different bits of two hash codes.

2) **Weighted Hamming distance**. Since the higher bits of the hash codes contain majority energy of an image, difference in higher bits should be larger-weighted. We evenly separates the 32-bit hash codes into 8 bins, and weights the corresponding Hamming distance by $2^{8-i}, 1 \le i \le 8$ for the i-th bin.

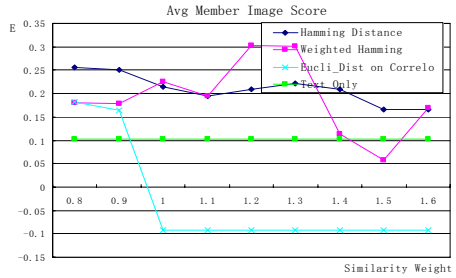
3) **Euclidean distance on color Correlograms**. We use this measure as a baseline to assess the effectiveness of the hash code based methods.

2.3 Mining Annotations by Clustering
The Search Result Clustering (SRC) algorithm [3] is used to cluster the retrieved semantically and visually similar images according to their titles, URLs and surrounding texts. Since SRC algorithm can generate clusters with highly readable names, we

[1] The work was done when Xin-Jing Wang was an intern in Microsoft Research Asia. Now she is with IBM China Research Lab in Beijing and her contact email is wangxinj@cn.ibm.com.



(a) Precision w.r.t. maximum cluster size criterion



(b) Precision w.r.t. average member image score criterion

Figure 1. *E* vs. similarity weight.

use these names as our candidate phrases, and rank them according to the following two criteria:

Maximum cluster size criterion. The score of a cluster is equal to the number of its member images. It assumes that the best key phrases are the dominant concepts of the member images.

Average member image score criterion. The cluster score is given by the average similarity of its member images to the query image.

At last, we remove duplicates from the top ranked phrases and output the rest ones as annotations of the query image.

3. EXPERIMENTS

2.4 million high quality photos with rich descriptions from online photo forums are extracted. Though the descriptions are noisy, they cover to a certain degree the concepts of the corresponding photos. The query dataset is 30 Google images from 15 categories (Apple, Beach, Beijing, Bird, Butterfly, Clouds, Clownfish, Japan, Liberty, Lighthouse, Louvre, Paris, Sunset, Tiger, Tree) that are randomly selected.

An evaluation criterion is proposed (see Eq.1), which differentiates "perfect" annotations (e.g. "Eiffel tower") from just "correct" ones (e.g. "France" for an Eiffel tower image).

$$E = (p + 0.5 \times r - w)/n \quad (1)$$

n denotes the number of annotations predicted. p, r, w are the number of "perfect", "correct", and "wrong" annotations respectively.

Figure 1 shows the curves of E of the three distance measure vs. the similarity weight. This weight weights the average similarity of images retrieved and gives a threshold for image filtering. The remained images are clustered for annotation mining. The green square curves in Figure 1 represent the text-based method as a baseline method that no visual features are available.

Figure 2. Examples of the Outputs

Figure 1 (a) shows that the weighted Hamming distance measure performs the best with maximum cluster size criterion. The reason is that it captures the important features of an image and weights them high. Interestingly, Euclidean distance measure performs nearly the same of the Hamming distance measure. It means that the information-loss due to PCA can be ignored on this dataset.

Figure 1 (b) shows that maximum average member image score criterion performs generally worse than maximum cluster size criterion. A possible reason is that SRC is a text-based clustering algorithm; hence images in a cluster may not be visually similar. Note that the system performance jumps when the threshold is too large so that images retrieved are too few to ensure good clustering performance.

The approach efficiency was also tested on a Dual Intel Pentium 4 Xeon hyper-threaded CPU and 2G memory computer. Images retrieved are 24,000 on average. The time cost is 0.034, 0.072, and 0.122 seconds for the three measures respectively (image ranking procedure is included).

Figure 2 shows a few examples of the query images and their predicted annotations. The boldfaced words are query keywords.

4. CONCLUSIONS

In this paper, we proposed a novel approach which reformulates the image auto-annotation problem as searching for semantically and visually similar images on the Web and mining annotations from their descriptions. To make it an online system, a hash coding algorithm is adopted to speed up the content-based search. Experiments conducted on 2.4 million photo forum images proved the effectiveness and efficiency of this proposed approach.

5. REFERENCES

[1] Huang, J., Kumar, S. R., Mitra, M., Zhu, W.-J., and Zabih, R. Image Indexing Using Color Correlograms. IEEE Conf. on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, (1997)

[2] Wang, B., Li, Z.W., and Li, M.J. Efficient Duplicate Image Detection Algorithm for Web Images and Large-scale Database. In Technical Report of Microsoft Research, 2005

[3] Zeng, H.J., He, Q.C., Chen, Z., and Ma, W.-Y. Learning To Cluster Web Search Results. In Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, Sheffield, United Kingdom, (July 2004). 210-217