# Leveraging Auxiliary Text Terms for Automatic Image Annotation [*]

Ning Zhou[†], Yi Shen[†], Jinye Peng[‡], Xiaoyi Feng[‡] Jianping Fan[†]
[†]Dept of Computer Science, UNC-Charlotte, Charlotte, NC 28223, USA
[‡]School of Electronics and Information, Northwestern Polytechnical University, Xi'an, CHINA
{nzhou, yshen9, jfan}@uncc.edu, {jinyepeng, fengxiao}@nwpu.edu.cn

## ABSTRACT

This paper proposes a novel algorithm to annotate web images by automatically aligning the images with their most relevant auxiliary text terms. First, a DOM-based web page segmentation is performed to extract images and their most relevant auxiliary text blocks. Second, automatic image clustering is used to partition web images into a set of groups according to their visual similarity contexts, which significantly reduces the uncertainty on the relatedness between images and their auxiliary terms. The semantics of the visually-similar images in the same cluster are then described by the same ranked list of terms which frequently co-occur in their text blocks. Finally, a relevance re-ranking process is performed over a term correlation network to further refine the ranked term list. Our experiments on a large-scale database of web pages have provided very positive results.

## Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content analysis and indexing

## General Terms

Algorithms, Performance, Experimentation

## Keywords

Automatic image annotation, image-text alignment, relevance re-ranking

## 1. INTRODUCTION

As the number of digital photos is growing exponentially on the Internet, there is an urgent need to develop new algorithms to annotate images with semantically meaningful labels by automatically aligning the web images with their most relevant auxiliary text terms extracted from associated texts. For each web page, it consists of two key components: web images and auxiliary text document. The auxiliary text document often contain a rich vocabulary of text terms. Some of them are used to describe the semantics of the images, but most of them are used for other web content description. It is unreasonable to use all these auxiliary text terms to interpret the web image semantics as most of them are weakly-related or even irrelevant with the image semantics.

In this paper, we seek to develop an automatic algorithm to achieve more precise alignment between the web images and their auxiliary text terms. There have been many related works on content-based image annotation where models are often learned from a training image database with labels (e.g., [2]). However, the success of applying such supervised annotation models on large-scale real-world web image data is very limited. We therefore propose an unsupervised algorithm to leverage auxiliary text term extracted from associated text to annotate web images. Experiments are conducted on an image-text data set produced from large-scale web pages.

## 2. IMAGE AND TEXT TERM ALIGNMENT FRAMEWORK

In this paper, an automatic algorithm is developed for achieving more precise alignment between the web images and their auxiliary text terms which consists of the following key components as illustrated in Fig. 1: (1) two filters are designed to extract the informative images from web pages by filtering out the low-quality images according to their sizes and aspect ratios; (2) a Document Object Model (DOM)-based web page segmentation algorithm is used to partition web pages into a set of image-text pairs where each informative image is associated with the most relevant surrounding text blocks; (3) an automatic image clustering is performed to group the web images into a set of image clusters according to their visual similarity contexts and the semantics of the visually similar web images in the same cluster are then effectively described by the same set of auxiliary text terms; and (4) an automatic alignment algorithm is developed to identify the relatedness between web images and their most relevant auxiliary text terms.

## 3. WEB PAGE SEGMENTATION AND INFORMATIVE IMAGE EXTRACTION

Informative images are extracted by filtering out those image whose aspect ratios are lager than 5 or smaller than 0.2 and those images whose widths or heights are less than 60 pixel. In addition, a DOM-based method is adopted to

[*]This work is partly supported by NSFC-61075014 and NSFC-60875016, by the Program for New Century Excellent Talents in University under Grant NCET-07-0693, NCET-08-0458 and NCET-10-0071 and the Research Fund for the Doctoral Program of Higher Education of China (Grant No.20096102110025).

*WWW 2011,* March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

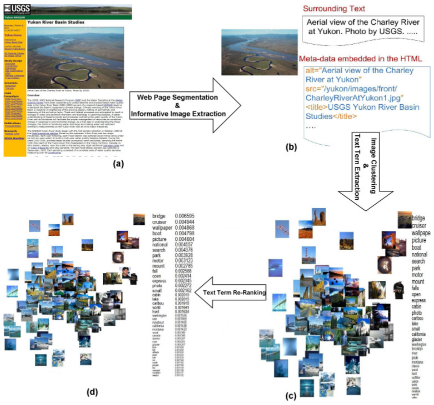

Figure 1: **The illustration of the key components of our image-text alignment scheme: (a) web page; (b) image-text pair; (c) image cluster and ranked auxiliary terms; (d) image cluster and re-ranked auxiliary text terms.**

extract the most relevant text block(s) for each informative image. Specifically, a region growing algorithm is employed to extract an informative web image's most relevant text block(s), where the corresponding image node in the DOM-tree is set as the start point, and a growing search is performed until it reaches any text node. The inner texts embedded in the text node(s) which have been touched by the region growing search are extracted as the text block(s). We also extract meta-data embedded in the HTML tags as side information, which often strongly reflects the semantics of an image. In this work, alternate texts, image titles, image file names, and web page titles, are used as the meta-data.

## 4. IMAGE CLUSTERING AND TEXT TERM EXTRACTION

Given two web images, the diverse visual similarity contexts between them are characterized by using a mixture of base image kernels strategy. To achieve more effective image clustering, an image similarity graph is first constructed, where each node denotes one particular web image and an edge between two nodes is used to characterize the pairwise visual similarity context. Automatic image clustering is then achieved by passing messages between the nodes through affinity propagation [3] which determines the number of clusters automatically. To extract semantically meaningful terms for image semantic interpretation, a standard list of the stop words is used to remove high-frequency words, such as "the", "to" and "also". Given a cluster, NLTK [1] tool kit is used to extract nouns from the text blocks as text terms which are initially ranked by the term relevance scores, computed as

$$\rho(C,t) = \frac{\sum_{x\in\Theta(t)} P(x,t)}{\sum_{x\in\Theta}\sum_{w\in\mathcal{W}} P(x,w)}, \qquad (1)$$

where $x$ is one particular web image in the cluster $C$.

## 5. TEXT TERM RE-RANKING

A term correlation network is automatically generated to characterize inter-term cross-modal similarity contexts and provides a good environment to refine the relevance scores.

Table 1: **Average precision of the alignment algorithm with or without clustering and random walk being employed.**

| | P@20 | P@40 | P@60 |
|---|---|---|---|
| Without Clustering | 0.6516 | 0.5827 | 0.5060 |
| Without RandomWalk | 0.7538 | 0.6935 | 0.6458 |
| Integration | 0.8086 | 0.7373 | 0.6900 |

In the term correlation network, each node represents a term and an edge indicates the pairwise term correlation. The inter-term cross-modal similarity are characterized by inter-term co-occurrences and inter-term semantic similarity contexts derived from WordNet. In order to leverage the advantage of the term correlation network to achieve more precise alignment between the web images and their auxiliary text terms, a random walk process is performed for automatic relevance score refinement. Given the term correlation network with $n$ most significant text terms, we use $\rho_k(t)$ to denote the relevance score for the text term $t$ at the $k$th iteration. The relevance scores for all these text terms in our term correlation network at the $k$th iteration will form a column vector $\overrightarrow{\rho(t)} \equiv [\rho_k(t)]_{n\times 1}$. We further define $\Phi$ as an $n \times n$ transition matrix, its element $\phi_{ij}$ is used to define the probability of the transition from the term $i$ to its inter-related term $j$. $\phi_{ij}$ is defined as

$$\phi_{ij} = \frac{\phi(i,j)}{\sum_k \phi(i,k)}, \qquad (2)$$

where $\phi(i,j)$ is the pairwise inter-term cross-modal similarity context between term $i$ and $j$.

## 6. ALGORITHM EVALUATION

Experiments are conducted on an image-text parallel set comprising around 500,000 web pages and 5,000,000 informative images. The effectiveness of the proposed algorithm is assessed from the view of image retrieval. We index each image with the ranked list of terms and build up an image retrieval system. Given a query term, the system returns all the images that are annotated with the term and ranks them according to the relevance scores. The precision of the top $k$ returned images by each term, denoted as "P@$k$", is computed and the average precision over 61 terms is reported. Specifically, we have compared the alignment algorithm under three different scenarios: (a) image clustering is not performed for reducing the uncertainty of the relatedness between images and their auxiliary terms; (b) random walk is not performed for term relevance re-ranking; (c) both image clustering and random walk are performed to achieve more precise alignment. As shown in Table 1, it is seen that incorporating image clustering for uncertainty reduction and performing random walk for relevance re-ranking can significantly improve the precision of automatic text-image alignment.

## 7. REFERENCES

[1] S. Bird. Nltk: The natural language toolkit. In *ACL*, 2006.

[2] S. Feng, V. Lavrenko, and R. Manmatha. Multiple bernoulli relevance models for image and video annotation. In *Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, volume 2, pages 1002–1009, 2004.

[3] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.