Word2Image: Towards Visual Interpretation of Words

Haojie Li, Jinhui Tang, Guangda Li, Tat-Seng Chua

School of Computing, National University of Singapore

{lihj, tangjh, chuats}@comp.nus.edu.sg, g0701808@nus.edu.sg

ABSTRACT

Besides the traditional textual semantic description to convey the meanings of a certain concept or word, visual illustration is a complementary, yet important and more intuitive way to interpret the word. Thus the technique that converts word to image is desirable though it is very difficult. Since a word usually has different semantic aspects, we need several correct and semantic-rich images to represent the word. In this paper, we explore how to leverage the web image collections to fulfill such task and develop a novel multimedia application system, *Word2Image*. Various techniques, including the correlation analysis, semantic and visual clustering are adapted into our system to produce sets of high quality, precise, diverse and representative images to visually translate a given word. The objective and subjective evaluations show the feasibility and effectiveness of our approach.

Categories and Subject Descriptors

H.4 **[Information Systems Applications]:**Multimedia application

General Terms

Algorithms, Design, Experimentation

1. INTRODUCTION

A picture is worth a thousand words. This proverb reveals the importance of visualization in explaining a concept or word. Suppose someone is introducing the concept of an animal, such as elephant, to a child. If he/she can provide some pictures of elephant to this child in the description, it will be more readily for the child to comprehend the meanings. Since manually finding such appropriate images is time consuming, we aims to develop a system that can automatic generate sets of images to visually interpret a given word. Automatically linking word to image is very helpful for people to rapidly and conveniently acquire knowledge, but it also involves some challenging issues. First, the correctness of linked images is critical; otherwise unrelated images will lead to misunderstanding. Second, since most word has different semantic aspects, the result images should be diverse enough to represent these aspects. Third, representative images should be selected from the image sets to reduce redundancy, that is to say, we should present the compact and visual appealing results to the users. Therefore, an automated word to image translation system should satisfy four requirements. They are: precision, diversity, representativeness of result images, and the friendliness and appealing of interface.

In recent years, the digital image collections on the web have grown rapidly, and many image search engines including content-based and keyword-based have been developed to help people to access these resources. In the keyword-based engines, such as *Google* [1] and *AltaVista* [2], when user types a keyword (or concept), the systems will return a large number of related images. Obviously, directly applying the search results is not an appropriate strategy for our application since the result list is not well organized and contains many irrelevant images. Moreover, since the images are crawled from all kinds of webpages, their quality is not ensured.

In this paper, we develop a system named *Word2Image*, in which we use the community-contributed photo website *Flickr* [3] as image resource, and employ various techniques to produce sets of high quality, precise, diverse and representative images to visually translate a given word. *Flickr* is a growing photo sharing website. As at 13th November 2007, it contained over 2 billion photos, and the contributors continuously upload 3-5 million premium photos daily. Each photo is loosely annotated with tags or metadata by the owner and other users, which is a great advantage over other image collections, making it possible to perform the semantic analysis needed in *Word2Image*.

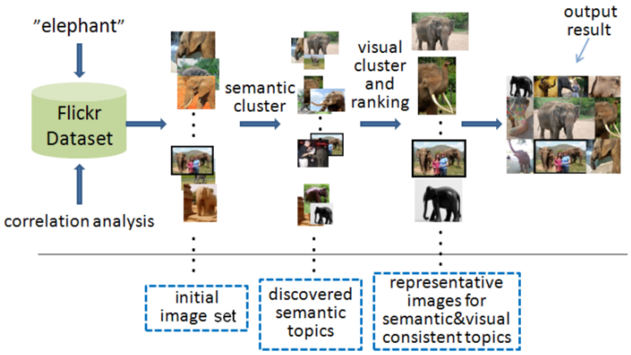
"elephant"
Flickr Dataset
correlation analysis
semantic cluster
visual cluster and ranking
output result
initial image set
discovered semantic topics
representative images for semantic&visual consistent topics

Figure 1. The systematic flowchart of *Word2Image*

As depicted in Figure 1, *Word2Image* works as follows. User inputs a word, here "elephant" is demonstrated as an example. First a large number of (say, 1000) diverse and precise images are retrieved from *Flickr* using heuristic rules and correlation analysis. Second, text-based clustering on images' tags is performed to generate image categories, which correspond to different semantic aspects of the given word. Each category is further clustered into visual consistent sub-categories and the representative images are selected. Finally, the collage technique is applied to generate a visually appealing presentation to user. To make it more comprehensible, each representative image is associated with several top keywords extracted from the cluster it belongs to.

In the literatures, our work is related to the image search result clustering, text-to-picture synthesis and applications on

MM'08, October 27–November 1, 2008, Vancouver, BC, Canada.

community-contributed web resources. In the multimedia research community, image clustering has attracted a lot of attention as it is a critical technology to help users digest large image collections. Cai *et al*. [4] clustered web image search results using visual, textual and link analysis to discover the underlying topics. Gao *et al.* [5] simultaneously used the low-level visual features and surrounding texts in one framework based on tripartite graph model. IGroup system [6] first identified some query-related semantic clusters based on web search result analysis. They then used the cluster names to retrieve images and organized the resulting images into a cluster structure with semantic level for user. These three works are closely related to ours. However, they are all designed for clustering the image search results, addressing the diversity of results while our purpose is to give a visual explanation of word. Again, the precisions in their systems were not satisfying and large amount of junk images may mislead the understanding of the word. Zhu *et al.* [7] proposed a text-to-picture system that attempts to visually translate unrestricted natural language text by synthesizing a picture based on both the image parts and extracted keyphrases. Compared to this work, our system uses the collage of sets of high quality real images to interpret a word but not just using one picture. Thus it can better represent the diverse semantic aspects of that word. Recently, Kennedy *et al.* [8] proposed to use *Flickr* to generate diverse and representative image search results for landmarks. They used visual clustering to find a landmark's diverse views and the results were encouraging. However their work was limited to landmarks, and the semantic diversity was not considered. Our work can be seen as a more general case.

2. GENERATING DIVERSE AND PRECISE IMAGE SET

As we have pointed out before, precision and diversity of images are two key requirements for the visual translation task. According to precision, we want the images to be correct; according to diversity, we want the images to be able to represent the different semantic aspects of the word. Generally speaking, the state-of-the-art image search and processing techniques have much difficulty in meeting such requirements. In this paper, we use *Flickr*, where the images are accompanied with some useful semantic cues. Such information includes: 1) image title along with several to dozens of tags added by the owner used to describe the content of image; 2) metadata, such as the photo's date and location, name of owner, etc. All these data and the images can be conveniently downloaded using *Flickr API* [3].

Heuristic for diversity. It is difficult to directly define diversity. However, we can expect that images come from different groups, different users, even different time and locations will show enough variations in both semantic and visual levels. Therefore, in *Word2Image*, we uniformly sample images from different groups and users to ensure the diversity of result image sets.

Correlation analysis for precision. The performance of today's keyword-based image search engines is not high enough to support our application. Even for the manually labeled image collections such as *Flickr*, the tag-based search results also contain many irrelevant images due to the noise in user provided tags. To filter out the wrong images, we conduct correlation analysis using *Flickr's* Related Tags.

Flickr's Related Tags is "a list of tags 'related' to the given tag, based on clustered usage analysis" [3]. For example, the top-10 related tags for "elephant" are "zoo, animal, Africa, animals, safari, London, wildlife, Kenya, nature, Tanzania". It can be seen that these words are either semantics related to the query or have high co-ocurrence with the query. We can deduce that if an image's title or tags contain the words in its related tags, it will be more likely to be relevant to the query. This motivates our criterion for filtering out unrelated images.

For a given word *w*, the related tags RT_w are firstly retrieved. Then for a retrieved image J, the correlation score of J with *w* is computed as:

$$CorrScore(J,w) = \#\{w' \mid w' \in RT_w \;\&\; w' \in (Tag_J \bigcup Title_J)\} \qquad (1)$$

where Tag_J and $Title_J$ are the tags and title of *w* respectively. $\#\{*\}$ is the cardinality of set $\{*\}$. If CorrScore(J,*w*) is above a threshold Th, image J is accepted as relevant.

3. DISCOVER DIVERSITY USING SEMANTIC CLUSTERING

When we learn a word or concept (take "elephant" as the example), some related concepts ("zoo", "animal") are helpful for capturing the meaning of the target. Also we may care about the unique characteristics ("trunk", "tusk") or sub-concepts ("African elephant") of the target, that is, the semantic diversity of a word. At the same time, the generated image set is diverse enough to include most of the topics of the word. Therefore, the visual translation system needs to discover these topics and group the images into their corresponding categories. This function is performed by the semantic clustering component.

Text clustering is a well-studied issue in text mining research community [9]. But the existing methods cannot be applied in our system directly because each image has a varying number of tags ranging from a few to several dozens, which is too sparse as compared to documents. Moreover, different keywords in the images' tags and titles contribute differently to the discovery of topics. For example, "trunk" will dominate over "water" in finding interested topics for "elephant". In *Word2Image*, we first compute the saliency of each keyword in the set of tags and titles and only top-M keywords are kept and used to represent each image with an M-dimensional vector. Then the agglomerative algorithm [10] is used to separate the images into different clusters. A cluster merging process is followed to combine the small clusters.

3.1 Computing the Saliency Score of Keywords

Saliency is used to measure a keyword's importance in discovering the distinct topics of a given word (denoted as *w* from now on). Here keyword refers to the words in the tag set. There are many factors that influence the saliency of a keyword. We consider four properties in our work currently. Before computing, each keyword is replaced with its stem using Porter algorithm [11].

Keyword Frequency/Inverce Document Frequency

This is similar to the traditional meaning of Term Frequency / Inverce Document Frequency (TFIDF). Intuitively, more frequent keyword will be more salient; however, keyword with higher document frequency (DF) will be too general and less informative. The keywords with too high and too low DF are further filtered out. The TFIDF for keyword K is computed as:

$$TFIDF(K) = freq(K)\log\frac{N}{I(K)} \qquad (2)$$

where freq(K) is the frequency of K in keyword set of *w*. N is the total number of images and I(K) is the number of images whose tags contain K.

Hyponymy and meronym with *w*

The hyponyms of a word reveal some of its important semantic aspects. For example, the hyponyms of “athlete” include “acrobat”, “baseball player”, “tennis player”, “runner” and so on. Obviously, these concepts should be selected as distinct topics of “athlete”. So, if a keyword is among the hyponyms of a target word, it will have higher saliency. This is similar with the meronyms of a word. For example, “tusk” and “trunk” are meronyms of “elephant”, while they are also two important aspects of “elephant”. Here HM(K) is defined to indicate whether the keyword K is the hyponyms or meronyms of *w:*

$$HM(K)=\begin{cases}1, & K\in(Hyponym(w)\bigcup Meronym(w))\\ 0, & otherwise\end{cases} \quad (3)$$

where Hyponym(w) and Mernym(w) are the hyponyms and meronyms of *w* and are obtained from WordNet [12].

Hyponymy between keywords

Some keywords may have hyponyms inside the keyword set. Such keywords should have less saliency score than their hyponyms since they are corresponding to relatively general topics. We define HH (K) to denote the number of hyponyms of K within the keyword set KS:

$$HH(K)=\#\{w'\,|\,w'\in Hyponym(K)\,\&\,(w'\in KS)\} \quad (4)$$

Related Tags

Here we prefer to the keywords in the related tags of *w*, since that they are selected based on the global statistics of *Flickr* dataset, thus they tend to be unbiased.

$$RT(K)=\begin{cases}1, & K\in RT_w\\ 0, & otherwise\end{cases} \quad (5)$$

Finally, the saliency score of K is calculated by combining the above four measurements with a simple fusion rule as follows.

$$Saliency(K)=TFIDF(K)\cdot(1+HM(K))\cdot\frac{1}{1+HH(K)}\cdot(1+RT(K)) \quad (6)$$

3.2 Text-based Clustering of Images

Given the saliency score of each keyword, the top-M keywords, KEYWORD={K_1,K_2,…,K_M}, are kept and each image’s text feature is represented using KEYWORD with a M-dimensional vector V_J=(v_1,v_2,…,v_M), where v_i is defined as:

$$v_i=\begin{cases}Saliency(K_i), & K_i \text{ is in image } J\text{'s tags or title}\\ 0, & otherwise\end{cases} \quad (7)$$

As expected, the topic-related keywords for *w* are ranked at top positions. For example, the top-10 salient keywords for “elephant” are “African, tusk, wildlife, trunk, safari, zoo, Thailand, animal, nature, India”. Evidently, these keywords are more distinctive and informative than the top-10 Related Tags (see Section 2) in discovering interesting topics.

Using the keyword vectors, we apply the agglomerative algorithm to hierarchically cluster the image set into different groups. Here the stopping criterion for clustering is controlled by the inconsistent coefficients [10]. Generally, it is difficult to determine the coefficients to get reasonable clusters. In our work since a cluster merging process is followed, we simply select a value, say 0.8 to make the resulting clusters more semantic consistent.

3.3 Cluster Merging

We merge the potentially similar clusters to reduce duplicated clusters and form larger cluster for later visual clustering. Specifically, each cluster is represented with top-k (k=6 in our experiments) salient keywords and if the number of overlapped keywords between two clusters exceeds a certain threshold, they will be merged into one cluster.

After merging, we obtain some interested clusters for the given word. Take “elephant” as the example again, the resulting clusters include topics like “India- wildlife- pachyderm- temple”, “animal-art- sculpture-Asia”, “zoo-London- trunk- tusk”, etc.

4. GENERATING REPRESENTATIVE IMAGES USING VISUAL CLUSTERING

In this Section, we apply visual clustering on each semantically consistent cluster obtained from Section 3 to divide them into visually coherent sub-clusters, and then select representative images for each cluster.

K-means is used here to perform clustering in the visual feature (grid color moments) space and the number of clusters is determined such that the average number of images in each resulting cluster is about 20, similar to what was done in [8].

After the 2-step clustering, we obtain clusters that are consistent in both the semantic and visual spaces. All these clusters then compete for the chance of being selected to be shown to the user. Here we use the following criteria to compute a cluster’s ranking score:

a) the sum of saliency score of keywords in the cluster;
b) the number of images in the cluster; and
c)the semantic and visual coherence of the cluster. This is measured with the ratio of inter-cluster distance (the average visual and semantic distance between images within the cluster and outside the cluster) to intra cluster distance (the average distance between images in the cluster).

Within each cluster, the images are also ranked according to theirs representativeness. The representativeness score of image is based on the intra cluster distance: the lower intra cluster distance, the higher the representativeness score.

5. PRESENTATION OF RESULTS

Besides the quality of results, the presentation of results is also important for the system to be accepted by the users. An ideal presentation should allow user to rapidly and conveniently digest the visual translation results. In *Word2Image*, we adopt the collage technique [13] to construct a compact and visually appealing image collage from the top representative images of the top-K clusters. The images are resized according to the respective cluster’s ranking score. To make the representative image more easily understandable, a large version, i.e. the original sized image will be shown when the user places the mouse over it, and the top-4 keywords will be displayed to depict its content.

6. EXPERIMENTS

To validate the effectiveness of *Word2Image*, we submit 25 concepts (such as *elephant, camel, buildings, athlete, pyramid, holidays, temple, flower, bridge* and so on) to the system and evaluate the results using two types of evaluating methods:

objective evaluation and subjective evaluation. The objective evaluation addresses the precision; while subjective evaluation is based on user study, focusing on diversity and representativeness.

6.1 Precision

This evaluation is used to validate the effectiveness of correlation analysis in improving the accuracy of retrieval and generating representative images. Two methods: tag based (used as baseline) and tag+correlation based methods are tested. Three metrics: the precision for image retrieval (P-IR) of 1000 images, the precision at generating top-10 (P@10) and top-20 (P@20) representative images are calculated for comparison. The performance in term of average precision is tabulated in Table 1. The results clearly show that the correlation analysis is helpful in improving all the 3 precision measures, which is the basic requirement of *Word2Image*.

Table 1. Precision for two methods

Precision Method	P-IR	P@10	P@20
Tag based	58%	82%	79%
Tag+Correlation based	83%	93%	90%

6.2 User Study

The second experiment highlights the system's usability and performances on discovering diversity and representativeness. 21 student volunteers are invited to take part in the evaluation. Among the volunteers, there are 7 primary school students, 7 middle school students and 7 graduate students. The volunteers are required to submit the 25 concepts to the system and explore the top-10 resulting representative images for each concept. They are then asked to fill in an assessment form with 4 questions as shown in Table 2. Each question requires a numerical answer based on the scale of: 1-strongly disagree, 2-disagree, 3-neutral, 4-agree, 5-strongly agree.

Table 2. Survey results from students on *Word2Image*

Score Assessment Questions	1	2	3	4	5
1) Do you think this system is useful in explaining the meaning of a word?	0	0	0	13	8
2) The coverage of the discovered topics. (The topics are explored with the superposed keywords shown on the representative images when mouse is placed over them)	0	0	1	16	4
3) The representativeness of the representative images	0	0	5	14	2
4) Overall satisfaction with the system	0	0	3	16	2

The survey results are tabulated in Table 2. From the answers to question 1, we can see that such visual translation system is highly desirable for all the three types of users. The answers to question 2 reveal that *Word2Image* can successfully find most of the interesting topics. This is attributed to the salient keyword detection and category clustering process. Results for two more examples: holidays and athlete, also support the answers. For "holidays", *Word2Image* discovers topics such as "Winter-december-happy-xmas", "Beach-sea-ocean-sun", "Disneyland-disney-california-travel", and "Vacation-travel-hotel-happy", etc. For "athlete", the topics like "Run-marathon-race-track", "Run-swim-ironman-bike", "Soccer-girl-ball-woman", "Basketball-ball-people-high", etc, have been extracted. The answers to question 3 are not so good as compared to others because we currently use simple visual features (color moments) to generate the representative images. We expect the use of more complex, specifically the object-level, features may alleviate this problem. The resulting collages for "holidays" and "athlete" are shown in Figure 2.

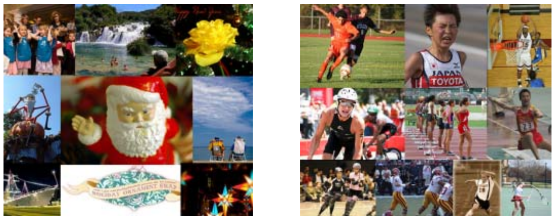

Figure 2. Results for "holidays" (left) and "athlete" (right)

7. CONCLUSION & FUTURE WORK

In this paper, we have introduced a novel multimedia application, *Word2Image*, which attempts to leverage the web image collection to translate a word into its visual counterpart with sets of high quality, precise, diverse and representative images. The preliminary experimental results have demonstrated its usability and effectiveness. This is a step towards our ultimate goal, to build a large scale multimedia dictionary, where multi-modality information including image, video, audio and text are integrated to explain the concepts. In the future, we will investigate more effective visual features for clustering, how to extract other modality cues and how to combine them.

8. REFERENCES

[1] Google image search engine, http://images.google.com/

[2] AltaVista image search, http://www.altavista.com/image/

[3] Flickr, http://www.flickr.com/

[4] D. Cai, X. He, Z. Li, W.Y. Ma, and J.R. Wen, "Hierarchical clustering of www image search results using visual, textual and link information", ACM MM 2004

[5] Bin Gao, Tie-Yan Liu, Xin Zheng, Qian-Sheng Cheng, Wei-Ying Ma, "Web image clustering by consistent utilization of visual features and surrounding texts", ACM MM, 2005

[6] Feng Jing, Changhu Wang, et al, "IGroup: web image search results clustering", ACM MM, 2006

[7] X Zhu, AB Goldberg, et al, "A Text-to-Picture Synthesis System for Augmenting Communication", AAAI 2007

[8] Lyndon S. Kennedy, Mor Naaman: "Generating diverse and representative image search results for landmarks", WWW 2008

[9] Berry, Michael W, "Survey of Text Mining I: Clustering, Classification, and Retrieval", Springer-Verlag, New York, 2003

[10] A. Jain and R. Dube, "Algorithms for Clustering Data", Prentice-Hall, Englewood Cliffs, NJ, 1988

[11] M.F. Porter, "An algorithm for suffix stripping", Program, 14(3), pp 130−137, 1980

[12] Christiane Fellbaum, "WordNet: An Electronic Lexical Database", MIT Press, 1999

[13] X.-L. Liu, T. Mei, X.-S. Hua, et al, "Video Collage", ACM MM 2007