COMMUNITY-GENERATED ONLINE IMAGE DICTORNARY

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

School of Computing
National University of Singapore
Singapore
E-mail: g0701808@nus.edu.sg, {lihj, tangjh, chuats}@comp.nus.edu.sg

ABSTRACT

Online image dictionary has become more and more popular in concepts cognition. However, for existing online systems, only very few images are manually picked to demonstrate the concepts. Currently, there is very little research found on automatically choosing large scale online images with the help of semantic analysis. In this paper, we propose a novel framework to utilize community-generated online multimedia content to visually illustrate certain concepts. Our proposed framework adapts various techniques, including the correlation analysis, semantic and visual clustering to produce sets of high quality, precise, diverse and representative images to visually translate a given concept. To make the best use of our results, a user interface is deployed, which displays the representative images according the latent semantic coherence. The objective and subjective evaluations show the feasibility and effectiveness of our approach.

Keywords: multimedia understanding, *Flickr*, user interface

1. INTRODUCTION

A picture dictionary is a dictionary containing word entries that are mostly accompanied by photos or drawings illustrate to what the words mean. 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. Traditional picture dictionaries are usually used for young children to learn foreign language or to get acquaintance with some knowledge. On the other hand, when one knows or has an idea of what something looks like, what usually lacks is the correct terms to describe it. The benefit of using pictures is that the visualization of concepts or words is very easy. Recently, Due to the rapid growth of internet users and the digital image collections on the web, several online picture dictionaries have been successfully developed, such as Visual Dictionary [1], and Visual Dictionary Online [2]. For example, when a query "elephant" is given to Merriam-Webster Visual Dictionary Online, several images are return to demonstrate the concepts. Returned results are shown in Figure 1. Although web image dictionary is very convenient to use, it doesn't overcome the disadvantages of traditional printed picture dictionary. Firstly, web visual dictionaries only present manually gathered image illustrations, which is impossible for user to grasp the diversity of the knowledge just through the set of strictly picked images. Secondly, not all concepts have the corresponding images to illustrate the meaning of the word. This is partly due to the reason aforementioned, that manually finding such appropriate images is very time consuming. Last but not least, the other disadvantage is that the pictures they usually chosen are not real word images, which is not a perfect way to illustrate the concepts in real scenario.

To overcome the disadvantage of the existing web image or visual dictionary, we aim to propose a new framework that can automatically generate the sets of images to visually interpret a given word. Automatically linking images to words is very helpful for people to rapidly and conveniently acquire knowledge, but it also involves several challenges. First, the correctness of linked images is critical; otherwise unrelated images will lead to misunderstanding. Second, since most words have different semantic aspects, the resulting set of images should be diverse enough to represent these aspects. Third, the representative images should be selected from the image sets to reduce redundancy, that is, we should present the compact and visual appealing results to the users. Hence, an automated word to image translation system should satisfy four requirements. They are: precision, diversity, representativeness of resulting images, and the friendliness and appealing of interface.

Fig. 1: The Result using Merriam Webster Visual Dictionary Online When the Query is Elephant

On the other hand, in recent years, the digital image collections on the web have grown rapidly, and many image search engines including the content-based and keyword-based systems have been developed to help users

to access these resources. For the keyword-based image search engines, such as *Google* [3] and *AltaVista* [4], 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 web pages, their quality is not ensured. Because of the advancements in video and image capturing, and the increasing data transmission rate, high quality of digital media is currently stored, shared, accessed, and distributed across in the internet. On a daily basis, there have been millions of new digital media uploaded and accessed on the online public media sharing websites. For example, *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 associated with metadata, in the form of tags annotated by the owners, as well as notes, comments, and even geographical location information. Such metadata provides valuable benefit for potential multimedia applications.

Based on the above reasons, we are using the community-contributed photo website *Flickr* [5] as image resource, and propose a novel framework, to visually interpret a given word. Our works are as follows. Firstly, there is no application developed before which tries to utilize large scale community generated multimedia resource to facilitate the development of traditional image dictionary. Secondly, motivated by the current natural language processing methods, various techniques are employed to produce sets of high quality, precise, diverse and representative images to demonstrate the concepts. Lastly, to make the best use of our results, a user interface is deployed. Differing from existing web visual dictionaries interface, which only show a limited number of thumbnails, the proposed interface display the representative images according to the latent semantic coherence within a certain concept.

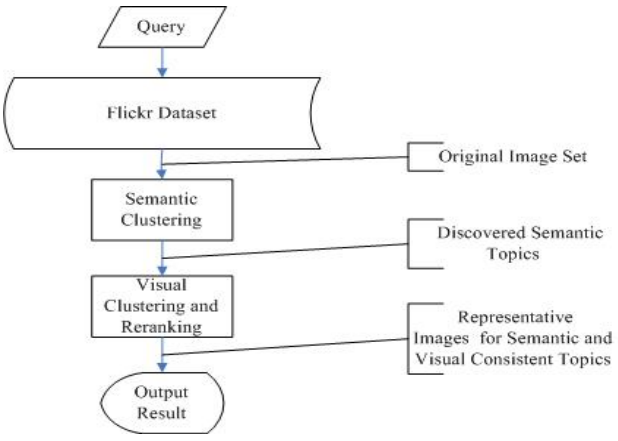

Fig. 2: Flowchart of Proposed Framework

The organization of the paper is as follows. Section 2 presents related work. Section 3 and 4 describe the overall framework and the design of user interface. In Section 5, objective and subjective study results are presented and analyzed. Conclusions are drawn in Section 6.

2. RELATED WORK

In the literatures, our work is related to the clustering of image search result, text-to-picture synthesis and applications on 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. [6] clustered web image search results using visual, textual and link analysis to discover the underlying topics. Gao et al. [7] simultaneously used the low-level visual features and surrounding texts in one framework based on tripartite graph model. IGroup system [8] 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. Although these three works are closely related to our work, they are all designed for clustering the image search results, and addressing the diversity of results. On the other hand our purpose is to give a visual explanation of words. Again, the precisions in their systems were not satisfying and a large amount of junk images may mislead the understanding of the words. Zhu et al. [9] 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 key phrases. 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. [10] 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.

3. DISCOVERING SEMANTIC DIVERSITY

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: a) image title along with several to dozens of tags added by the owner used to describe the content of image; b) 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* [5].

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 our proposed framework 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" [5]. 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 semantically 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 unrelated images.

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

$$CorrScore(J,W) = \#\{w' | w' \in RT_w \& \, w' \in (Tag_w \cup Title_w)\} \quad (1)$$

where Tag_w and $Title_w$ 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.

3.1 Saliency Words discovery

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 [11]. 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 our framework, 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 [12] is used to separate the images into different clusters. A cluster merging process is followed to combine the small clusters.

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 [13].

3.1.1 Keyword Frequency/Inverse Document Frequency

This is similar to the traditional weighing scheme of Term Frequency/ Inverse Document Frequency (TFIDF) [14]. 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:

$$TFJIF(K) = \sum_{j=1}^{N} freq(K,j) * log \frac{N}{I(K)} \quad (2)$$

where freq (K,j) is the frequency of K in j^{th} image's tag and title. N is the number of images and I(K) is the number of images whose tags contain K.

3.1.2 Hyponymy and Meronym

In linguistics, a hyponym is a word or phrase whose semantic range is included within that of another word. 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:

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

where Hyponym(w) and Mernym(w) are the hyponyms and meronyms of w and are obtained from WordNet [15].

3.1.3 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:

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

3.1.4 Related Tags

Here we prefer to the keywords in the related tags of w, since 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.

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

3.2 Text-based Clustering of Images

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

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

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". The comparison between the top-10 related tags in *Flickr* and our algorithm are given in Table 1. Evidently, the keywords are generated by our algorithm more distinctive and informative than the top-10 Related Tags from *Flickr*'s(see Section 2.1) in discovering interesting topics.

Table 1: The Comparison with Original *Flickr* Related Tags

Flickr Related Tags	zoo, animal, Africa, animals, safari, London, wildlife, Kenya, nature, Tanzania
Text-based Clustering Results	African, tusk, wildlife, trunk, safari, zoo, Thailand, animal, nature, India

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 [12].

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 semantically consistent. 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, the resulting clusters include topics like "India- wildlife- pachyderm- temple", "animal- art-sculpture- Asia", "zoo- London- trunk- tusk", etc.

3.3 Visual-based Clustering of Images

Next, we apply visual clustering on each semantically consistent cluster obtained from Section 3.2 to divide them into visually coherent sub-clusters, and then select representative images for each cluster. K-means is used here to perform the clustering in the visual(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 [11]. 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 the intra cluster distance, the higher the representativeness score.

4. USER INTERFACE

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 our framework, we adopt the collage technique [16] to construct a compact and visually appealing image collage from the top representative images of the top-K clusters. The UI basically consists of two functions. The first function is semantic resizing. The images are resized

according to the respective cluster's ranking score. The browsing system is extended with collage. The second function is Zoom function. 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.

Fig. 3: Results for "holidays"

The resulting collages for "Pyramid" in Figure 3. "France- Paris- museum- louvre", "Africa- Egypt- cairo- desert", "Mexico- yucatan- maya- temple" and "history- architecture- giza- spohinx" are discovered.

5. EXPERIMENT

To validate the effectiveness of our proposed framework, we manually chose 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.

5.1 Performance of Queries

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 Figure 4. The results clearly show that the correlation analysis is helpful in improving all the 3 precision measures, which is the basic requirement of our proposed framework.

5.2 Subjective Evaluation

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 these 25 concepts to the system and explore the top-10 resulting representative images for each concept. After each task, 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.

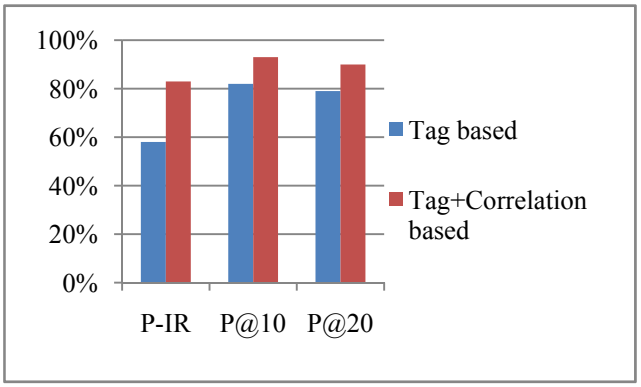

Fig. 4: Precision Value Comparison of Different Metrics

The survey results are tabulated in Table 5. 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 our framework can successfully find most of the interesting topics. This is attributed to the salient keyword detection and category clustering process.

Table 2: User Study

Score / 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

Results for two more examples: athlete and holiday also support the answers. For "athlete", the topics like "Run- marathon- race- rack", "Run- swim- ironman- bike", "Soccer- girl- ball- woman", "Basketball- ball- people- high", etc, have been extracted. For "holidays", topics such as "Winter- december- happy- xmas", "Beach- sea- ocean- sun", "Disneyland- disney- california- travel", and "Vacation- travel- hotel- happy", are discovered. The answers to question 3 are not so good as compared to

others because we currently use simple visual features to generate the representative images. We expect the use of more complex, specifically the object-level, features may alleviate this problem.

Fig. 4: Results for "athlete" (left) and "holiday" (right)

6. CONCLUSION AND FUTURE WORK

Billions of images, shared on websites bring profound social impact to the human society, and at the same time pose a new challenge: how to effectively make use of these multimedia data other than just searching and manipulating them. In this paper, we have introduced a novel framework, 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 adapting various techniques, including the correlation analysis, semantic and visual clustering methods. To make the best use of our results, a user interface is deployed. Different from existing web visual dictionaries interface, which only show a limited number of thumbnails, the proposed interface displays the representative images according to the latent semantic coherence. 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.

6. REFERENCES

[1] Visual Dictionary, http://www.infovisual.info/

[2] Visual Dictionary Online, http://visual.merriamwebst-er.com/

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

[4] AltaVista image search engine, http://www.altavista.c-om /image/

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

[6] 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

[7] 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

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

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

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

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

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

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

[14] Salton, G. and M. J. McGill, "Introduction to modern information retrieval", McGraw-Hill, 1983

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

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