# A Game Based Approach to Assign Geographical Relevance to Web Images

Yuki Arase†, Xing Xie‡, Manni Duan*, Takahiro Hara†, Shojiro Nishio†

† Dept. of Multimedia Engineering Grad. Sch. of Information Science and Tech., Osaka University, 1-5 Yamadaoka, Suita, Osaka 565-0871, Japan

{arase.yuki, hara, nishio}@ist.osaka-u.ac.jp

‡ Microsoft Research Asia

4F, Sigma Building, No.49, Zhichun Road, Haidian District, Beijing 100190, P. R. China

xingx@microsoft.com

* Dept. of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, 230027, P. R. China

mnduan@mail.ustc.edu.cn

## ABSTRACT

Geographical context is very important for images. Millions of images on the Web have been already assigned latitude and longitude information. Due to the rapid proliferation of such images with geographical context, it is still difficult to effectively search and browse them, since we do not have ways to decide their relevance. In this paper, we focus on the geographical relevance of images, which is defined as to what extent the main objects in an image match landmarks at the location where the image was taken. Recently, researchers have proposed to use game based approaches to label large scale data such as Web images. However, previous works have not shown the quality of collected game logs in detail and how the logs can improve existing applications. To answer these questions, we design and implement a Web-based and multi-player game to collect human knowledge while people are enjoying the game. Then we thoroughly analyze the game logs obtained during a three week study with 147 participants and propose methods to determine the image geographical relevance. In addition, we conduct an experiment to compare our methods with a commercial search engine. Experimental results show that our methods dramatically improve image search relevance. Furthermore, we show that we can derive geographically relevant objects and their salient portion in images, which is valuable for a number of applications such as image location recognition.

## Categories and Subject Descriptors

H1.2 [**MODELS AND PRINCIPLES**]: User/Machine Systems – human factors; H5.3 [**HCI**]: Web-based interaction.

## General Terms

Algorithms, Design, Human Factors

## Keywords

Geographical relevance, image search, image annotation, human computation



*WWW 2009*, April 20–24, 2009, Madrid, Spain.

ACM 978-1-60558-487-4/09/04.

## 1. INTRODUCTION

With the explosive growth of images on the Web, image search has become an important service for search engines, as all major engines, like Live Search, Yahoo, and Google provide it. They usually utilize the file name, <alt> attribute, caption, anchor text, Web links, and the surrounding text of images to rank them instead of analyzing the image content, since there is a big gap between low-level image features and their semantic meanings. In addition, the computational cost of image content analysis is still too high for practical use. Thus, owing to a lack of the information of the image itself, it is difficult to decide the image's relevance to a search query.

(a) Highly relevant (b)(c) Lowly relevant

**Figure 1. Example of geographical relevance**

According to [24], 11.7% of Web search engine queries have associated locations. The authors of [24] also mentioned that more queries will have a geographical intention, that is, queries like "map", "weather", and "restaurant" have such an intention, although there is no location term in the query. Regarding images, geographical context is also very important. In theory, every personal photo is associated with a location since it should be taken somewhere. It would be very helpful to organize and visualize photo collections according to their spatial distribution. Flickr [9], a popular photo-sharing Web site, enables users to tag their images with latitude and longitude and presents them on a map. The number of such geo-tagged images in Flickr has recently exceeded 100 million. Google Earth [11] also enables users to upload their images with latitude and longitude information, and present them on the surface of the earth. Unfortunately, the fast growing pace of such geo-tagged images has exceeded the technology for searching and browsing them, as we have to comb through search results containing a considerable amount of irrelevant images.

In this paper, we focus on images' *geographical relevance*. In the information retrieval (IR) domain, *relevance* usually means to what extent the topic of a result matches the topic of a query.

Following this definition of relevance, we define geographical relevance for images, as to what extent objects of an image match a given location. Figure 1 shows an example, where an image that contains a visible landmark at its center can be regarded as relevant (Figure 1 (a)), while an image that does not contain the landmark can be regarded as irrelevant (Figure 1 (b)). Additionally, we can also define the probability of the geographical relevance; the former image can have a higher probability of geographical relevance than an image with a high level of occlusion and that doesn't focus on the landmark (Figure 1 (c)). Geographical relevance enables search engines to rank typical landmark images higher, and rank the images containing the landmark but being unrepresentative lower. However, it is also possible to adjust the diversity of the results by changing the weight of relevance, allowing images with low relevance to show up in the search results. In addition, geographical relevance can present images on maps in an organized way, which currently shows a large number of images in a more haphazard manner. It enables better summarization of these images and provides an ability to find more relevant images easily.

Although quite a few computer vision techniques have been proposed for image location recognition [8][13][26], most of them are based on feature matching of images, and none of them considered the concept of relevance. Determining image geographical relevance is challenging for computers because the criteria of relevance is too abstract. However, there is one approach which can solve this problem very easily – *asking humans*. Luis von Ahn proposed a novel concept called *human computation* [1] that makes use of normal people's brains to solve computationally difficult tasks. The typical example is the ESP game [4], which allows people to label images while having fun. Being inspired by the concept of human computation, we designed a Web-based and multi-player game to collect image geographical relevance from humans. Though the ESP game is a milestone of this area, it still has some drawbacks. The game logic is a bit simple, so people might not be loyal because they can prove little about their ability. Moreover, it has not been apparent how the game based approaches could collect useful human knowledge. When designing our game, we paid attention to attract people to play it for a longer period, and more importantly, we provide an in-depth study on game logs and how they can improve existing applications.

The contributions of this paper include:

- We define the concept of the geographical relevance of images, and design a game that collects human knowledge to assign geographical relevance to images while people are enjoying the game.
- We thoroughly analyze the game logs obtained by a three week study. We find locations' characteristics and their similarity based on human perception. Then we propose methods to determine the geographical relevance of images. Furthermore, we also derive the relevant regions and the salient portion of the images based on the game logs.
- We apply image geographical relevance to image search to refine the ranking of results and conduct an experiment to examine its effectiveness.

The rest of the paper is organized as follows. In Section 2, we introduce related works and discuss their difference from our approach. In Section 3, we explain the design of our game. In Section 4, we analyze the game logs obtained by a three week study in detail. In Section 5, we discuss applications on which our game logs provide an advantage. Especially, we propose methods to refine image search results and conduct an experiment to verify their effectiveness. Finally, we conclude our paper in Section 6.

## 2. RELATED WORK

The main research efforts related to our work are computer vision approaches for image location recognition, as well as automatic image quality estimation and image annotation. Another closely related field is games, especially geographically based and human computation based games.

Image location recognition has been studied for decades in the computer vision field. Zhang et al. [26] proposed a method to identify a location from an image in an urban environment. It finds the closest reference images from the given database of images assigned with GPS data, and then estimates the location of a query image by triangulation, using the camera angles of the reference images. Hays et al. [13] collected six million GPS-tagged images from the Web and used them as training data. They estimate a query image's location by low-level feature matching. The estimated location is a probability distribution over the earth's surface. Epshtein et al. [8] propose a method to automatically classify images at a location according to their focusing objects. They define *Geo-relevance* that represents an object's relevance to a location, while our geographical relevance represents an image's relevance to the location. These image location recognition methods aim to decide the location of a query image. However, we assume images' locations are given and aim to decide their geographical relevancies.

As for automatic image quality estimation, there are two approaches, as one is computer vision based and the other is Web based. Ke et al. [17] proposed a method to assess the quality of a photo using high level features. Their method can distinguish photos taken by professional photographers and being worth hanging on the wall, from ones taken by ordinary people and enough for photo albums. Jing et al. [16] applied the PageRank algorithm on an inferred visual similarity graph to identify an authority image that has common features with other images. As for the Web based approach, PicASHOW [19], WPicASHOW [23] and ImageSeer [14] make use of actual link structures of Web pages, in which images are contained to decide authority images. Image quality there is orthogonal against our geographical relevance. Image quality corresponds to a static aspect of relevance, while our geographical relevance corresponds to a dynamic aspect of relevance. Image quality can be combined with geographical relevance for better image retrieval.

As for the automatic image annotation method, the main approach uses machine learning techniques to learn the probabilities between images and keywords [6][18]. They rely on supervised training to learn prediction models, and thus, the annotations are limited to the ones included in the training set. They estimate the probability of the relevance between annotations and images, which are based on the models and supervised training. However, our geographical relevance is directly based on human perception. A notable work in this field is the AnnoSearch [25] system, since it combines Web search and data mining techniques and enables annotation with unlimited vocabulary. A problem with these automatic image annotation methods is that their accuracy gets lower when the size of their vocabulary is large. To achieve accurate and scalable image annotation, Luis von Ahn proposed the ESP game [4], as we discuss later. These methods aim to

annotate images, which is essential for image search, while our method aims to assess their geographical relevance.

Next, we focus on games. Regarding the geographically based games, there are some commercial game services [10][12]. In these games, players can set actual location based "missions" with each other. In the GPS Mission [12], players register a location and set a small quiz relating to the location on the game Web site. Registered locations are uploaded on a common map, and other players can challenge the quiz. Players use GPS enabled cellular phones and actually visit the quiz point to solve the quiz. Geocaching [10] is closer to the practical world; as with the GPS Mission, players can register their own mission through the Web site. However, the mission here is to find an actual treasure (usually small gifts, such as a key holder) hidden by a mission creator. Other players can know the latitude and longitude of the location, and visit there to find the treasure. These games do not focus on research, however, given that the missions on these games are actively created (missions are varied through more than 35 countries for the GPS Mission, and the number of missions exceeds 650,000 for the Geochaching), we can find geographically-based games are getting increased attention.

In addition to the above works, human computation is an especially important research field for us. Though there are many works following this concept [3][5][7][22] , the pioneer and the most related application to us is the ESP game [4]. In the ESP game, randomly paired players are shown the same image. They score points when they enter the same description about the image. The matched descriptions can be regarded as representative labels for the image. Additionally, Peekaboom [2] is a subsequent application of the ESP game. It uses labeled images by the ESP game and tries to detect the object regions of the labels in images. Human computation is quite an effective way to collect human knowledge to solve computationally difficult problems. It is especially suitable to solve recognition problems based on human perspective. Being inspired by this concept, we designed a Web-based and multi-player game to collect human knowledge to determine the geographical relevance of images. However, our game is different from the ESP game on the following points:

- Topic dependency: Our game focuses on the geographical relevance of images, while the ESP game does not limit its domain. By focusing on the specific domain, we can obtain its rich knowledge and design our game so that players have fun and learn additional knowledge from the game.
- Detailed data analysis: Most importantly, our main contributions are game log analysis and proposal of methods to assign geographical relevance to images rather than only implementing and testing a game. Moreover, we derive a relevant region and its salient portion from an image.

# 3. GAME DESIGN

Our purpose with regards to the game is to collect knowledge from players to determine image geographical relevance. The distribution of geographical relevance is not uniform within an image. Therefore, we collect human knowledge on 1) which image is relevant for a landmark as a whole 2) which image region is relevant for the landmark. The former is effective for the refinement of image search results, while the latter can help image location recognition techniques.

## 3.1 Basic Game Play

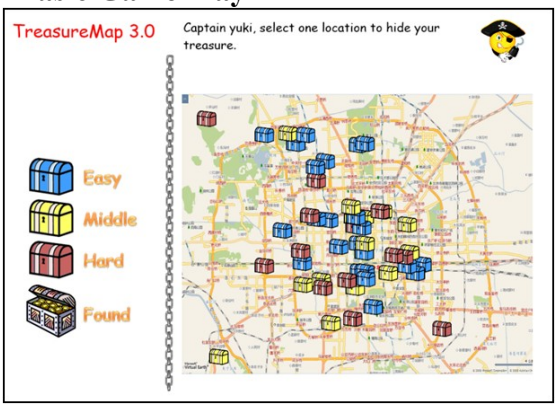


**Figure 2. Captain can hide a treasure to a location**

Our game is based on a motif of a pirate hunting treasure. Pirates are divided into Captain and Sailor, indicating different roles in the game. At first, the Sailor can decide a city to play from a world map. Then, the Captain selects a location from a map to hide a treasure (see Figure 2). When selecting the location, six candidate images are displayed to the Captain, and he can select one to play; when the Captain selects the image, the time countdown starts. The Captain can open image regions by clicking the image to give hints to the Sailor (see Figure 3). Each click reveals a 30 pixel square region, and decreases the score players can get. The Sailor can only see the regions opened by the Captain, and the Sailor has to infer where the treasure is (see Figure 3).

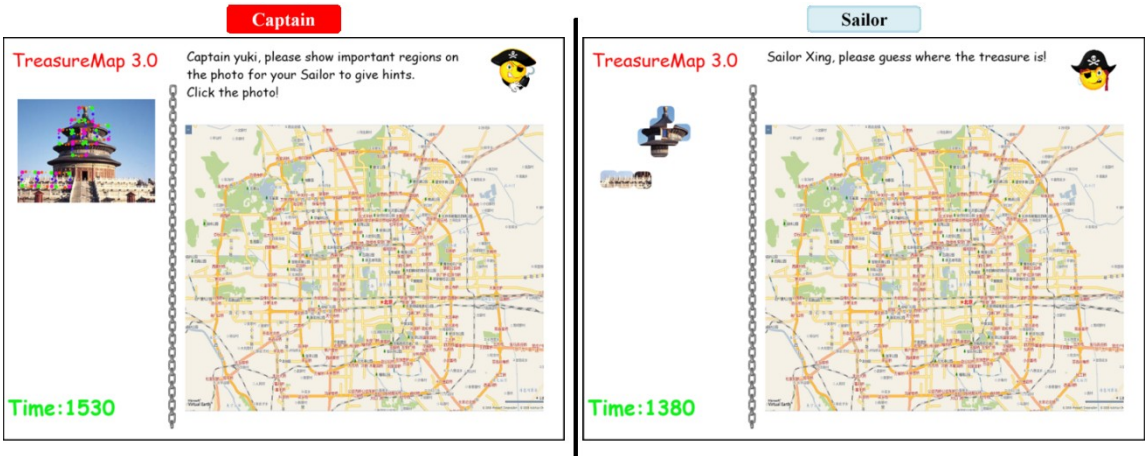


**Figure 3. Captain opens image regions to Sailor, and Sailor infers the location**

When the Sailor comes up with an answer, he can click the location on the map, and then both the Captain and Sailor get the same score. Here, the Captain and Sailor cannot know who is the partner nor communicate with each other. Therefore, the only way for the Captain to tell the location of the treasure to the Sailor is by opening the image region.

We designed the game to use a map instead of directly inputting landmark names, since text input is too simple as an interaction and cause problems in the form of spelling errors, language dependency and ambiguity of expression. For example, "Great wall", "Great wall of China", and "Chinese wall" are all correct. Furthermore, map based interaction is more challenging and players can prove their knowledge on locations to others, which can be an attraction to drive players.

## 3.2 Score

In the game, the Captain and Sailor receive the same score. The score is calculated based on the distance from the Sailor's guess and the correct location of the landmark (the unit is [pix]), the time used ([sec]), the number of clicks to open image regions, and the difficulty level of the location according to equation (1).

$$score = \frac{default\ score - \alpha \cdot time - \beta \cdot \text{num. of } click}{(1 + \gamma \cdot distance)^2} \quad (1)$$

Here, we set $\alpha = 1, \beta = 30, \gamma = 0.01$ to keep valance of each term. Regarding the *default score*, it is determined by the level of the location, as a difficult location has a higher score than those of easy levels. In the current implementation, the level of each location is manually assigned. However, it can be determined automatically based on the aggregated players' scores.

Players can get a high score if the Sailor clicks a closer point to the correct location on the map with a small number of opened regions in a shorter time. Therefore, the strategy to maximize the score is that the Captain selects a relevant image and opens a relevant portion of a landmark so that the Sailor can infer the location as quickly as possible. Thus, given a landmark and images, the game logs reveal relevant images and their relevant portions for the landmark.

## 3.3 Game Logs

We save a game log for each image selection. As Table 1 shows, we save the city selected by the Sailor, the location and image selected by the Captain, the location of Sailor's guess, and the play time. In addition, as Table 2 shows, we save all of the central coordinates of the opened image regions and their time by the Captain. Based on these game logs, we can know all actions of the Captain and Sailor on the image.

**Table 1. Elements saved in a game log**

| Name | Description | Example |
|---|---|---|
| Selected city | Selected city name | Beijing |
| Selected location | Selected location name | Summer Palace |
| Selected image | Selected image ID | 5514 |
| Guessed location | Latitude and longitude of Sailor's guessed point | (40.022, 116.200) |
| Start time | The date and time the countdown started. | 2008/09/08 22:02:33 |
| End time | The date and time Sailor clicked a location on the map. | 2008/09/08 22:02:41 |
| Actions | Information of Captain's clicks as Table 2 shows. | |

**Table 2. Each click is saved as Captain's action**

| Name | Description | Example |
|---|---|---|
| Clicked point | X,Y-coordinate of Captain's click on the image | (138, 87) |
| Clicked time | The date and time of Captain's click | 2008/09/08 22:02:37 |

## 3.4 Implementation

We implemented the architecture of the game under the server/client model. The client application is accessible as a Microsoft Silverlight application through Web browsers, while the server is written by Microsoft Visual C# .Net. The client handles the whole user interaction, and sends the information of players' actions to the server. The server processes the whole game logic, such as matching players, database accesses, and saving game logs into a corresponding database.

# 4. GAME LOG ANALYSYS

We released our game prototype for internal study in Microsoft Research Asia in the summer of 2008. In the prototype, players can select Beijing and Shanghai to play, which contain 50 and 30 locations to hide a treasure, respectively. The total number of images was 1641.

All images were collected using Live Image Search [20] by querying the name of landmark, such as "National Stadium". These images were stored in our game database in advance.

The test duration was three weeks. Since the study was conducted internally, we tried to keep participants' anonymity to guarantee the fairness of the study. We sent a call for participation to all intern students in MSRA by e-mail. As a result, we received 147 participants and 2761 game logs. Participants were all undergraduates or graduate students, and have a computer and information science background. We provided small gifts to active players.

In the following sections, we analyze the game logs obtained by this study from location, image, and image region based aspects. For the analysis, we use game logs of locations which were played at least 10 times (67 out of 80 locations were played at least 10 times). Finally, we describe the results of the questionnaire survey we conducted after the study.

## 4.1 Location Characteristics

To analyze the characteristics of each location, we calculate the average discrepancy distance between each player's guessed point and the correct location. The discrepancy distance is the distance between a guessed point and the correct location, which is calculated based on the latitude and longitude of these points. Then we averaged the sum of all discrepancy distances for each location by the number of guesses made for the location. Here, we excluded game logs which failed to make an answer.

Table 3 and Table 4 show the top 5 and worst 5 locations with regarding to the average discrepancy distance. The commonality of the top 5 ranked locations is that they have something characteristic in their locations, such as picture of Mao Tse-tung on Tiananmen. As we examined the players' guessed points for Tiananmen, most players' guesses concentrated near the correct location, which indicates that players could clearly recognize Tiananmen from the opened image regions by the Captain. Another commonality is that these locations are easier to find on the map. For example, Nanpu Bridge is on a big river in Shanghai and clearly illustrated on the map, and thus, players could easily find it.

**Table 3. Top 5 locations with regards to average distance**

| Rank | Location name | City | Avg. disc. dist. [m] |
|---|---|---|---|
| 1 | People's Square | Shanghai | 463 |
| 2 | Nanpu Bridge | Shanghai | 940 |
| 3 | Longhua Temple | Shanghai | 975 |
| 4 | Tiananmen | Beijing | 1083 |
| 5 | Imperial Palace | Beijing | 1163 |

**Table 4. Worst 5 locations with regards to average distance**

| Rank | Location name | City | Avg. disc. dist. [m] |
|---|---|---|---|
| 67 | Wofo Temple | Beijing | 8852 |
| 66 | Jingshan Park | Beijing | 7904 |
| 65 | Taoranting Park | Beijing | 7060 |
| 64 | Summer Palace | Beijing | 6681 |
| 63 | Grand View Garden | Beijing | 6670 |

The commonality of the worst 5 locations is that they lack remarkable characteristics. As you can see, all locations ranked in here were open-plan locations like a park and garden. Figure 4 plots the guesses on the location of the Grand View Garden (a large garden modeled on a Chinese classic novel called "Dream of the Red Chamber"), where each red point represents a guess and the blue arrow represents the collect location. The scattered guessed points indicate that players could not recognize the location from opened image regions by the Captain. The typical example here is Summer Palace, which is one of the most famous locations in Beijing, is ranked quite low. The Summer Palace is indeed famous, but is famous for facing a big lake, and thus, most images are shots of lake and some Chinese style buildings looking similar with ones in other locations. Therefore, it is difficult to recognize the Summer Palace from these images. When plotting the players' guesses, wrong guesses concentrated near the other lakes in Beijing, which indicates players misunderstood the location because of the similarity with other locations.

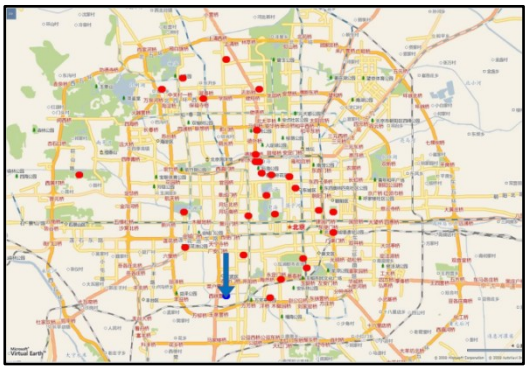

**Figure 4. Guessed points plot for Grand View Garden**

These locations are challenging to automatically decide image relevance since there are no characteristic objects, that is, there are less features to match with. Therefore, our approach based on human perception works effectively.

Furthermore, we can derive similar landmarks based on the plot of players' guesses. Figure 5 plots the guessed points for Zhengyangmen (a former front gate of the ancient Beijing) and Deshengmen (a former gate having been a part of Beijing's northern city wall) Archery Tower. The bottom blue arrow indicates Zhengyangmen's location and cyan points plot players' guesses when they played with Zhengyangmen, while the above green arrow indicates the location of Deshengmen Archery Tower and magenta points plot players' guesses when they played with Deshengmen Archery Tower.

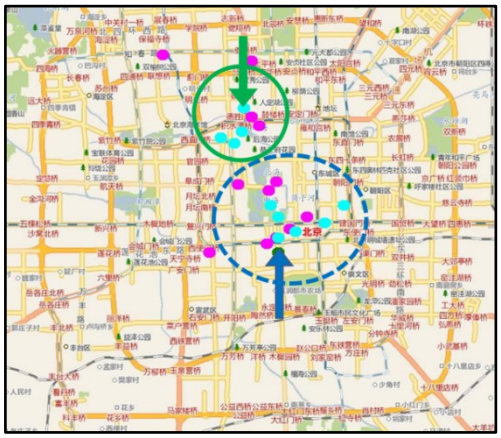

**Figure 5. Two classes of plots; one is for correct location and the other is for location of similar appearance**

(a) Zhengyangmen (b) Deshengmen Archery Tower

**Figure 6. Zhengyangmen and Deshengmen Archery Tower**

As we can see, there are two groups of plots, as the one group (indicated by the dotted blue circle) concentrates on the location of Zhengyangmen, and the other group (indicated by the solid green circle) concentrates on the location of Deshengmen Archery Tower, where guesses of the both locations are mixing. These plots show that players mixed up these two locations. Figure 6 shows the most frequently selected images of these two locations, the left side is Zhengyangmen and the right side is Deshengmen Archery Tower, and you can see their similarity.

This similarity is difficult to detect by computers since the details of these images are not common. The similarity originates from higher level human perception, not only actual appearance of the images but also estimated shapes of landmarks based on the images and vague memory of the location. Our game logs also enable us to detect such higher-level similarity among landmarks, because they are actually based on human perception.

We conducted co-location pattern mining [15] on the players' guessed points at locations to find relationships between the locations. The algorithm requires setting two parameters. We set *distance threshold*, which decides whether two guessed points are neighbors or not, as 1km and *participation index threshold*, which decides at least how frequently the guessed points have to co-occur, as 0.7. It means that at least 70% of guessed points of two locations have to be observed within *distance threshold* to be a co-location pattern. By the nature of co-location pattern mining algorithm, it finds locations actually being close with each other as co-located since players' guessed points are very close. Therefore, we filtered out locations of which their actual distance is less than double of the *distance threshold*.

**Table 5. Co-location patterns based on players' guessed points**

| Co-location pattern | Participation index |
|---|---|
| {Imperial Palace, Zhengyangmen} | 0.783784 |
| {Beijing University of Aeronautics and Astronautics, Tsinghua University} | 0.755102 |
| {Deshengmen Archery Tower, Tiananmen} | 0.740741 |
| {China Millennium Monument, Temple of Heaven} | 0.722222 |
| {China Millennium Monument, West Beijing Railway Station} | 0.717949 |
| {Deshengmen Archery Tower, Imperial Palace} | 0.703704 |
| {Deshengmen Archery Tower, Zhengyangmen} | 0.703704 |

Table 5 lists the results; pairs of locations which were detected as co-location patterns, and their *participation indexes* [15], which represent the probability of these pairs can be observed in the data. A high *participation index* value indicates that the guessed points on the locations likely occur together, that is, the locations can be regarded as similar. As Table 5 shows, the Imperial Palace, Tiananmen, Zhengyangmen and Deshengmen Archery Tower are detected as co-located. This is because the appearances of these landmarks look similar, as a huge gate with traditional shape of Chinese buildings, and thus, players confounded them. Tsinghua University and Beijing University of Aeronautics and Astronautics are also detected as co-located. The landmark on both locations is a university and it is agreeable that players mixed them up. However, the China Millennium Monument, Temple of Heaven and West Beijing Railway Station do not have visual similarity. This is because the average discrepancy distances of the Temple of Heaven and West Beijing Railway Station are comparatively large, and thus, the players' guesses on these locations tended to co-occur with the guesses on other locations. We can eliminate them by adjusting the *distance threshold.*

Though our game is currently limited to inside a city, if we enlarge it, in other words, allow players to select locations outside the city, we can detect similar locations around the world. This knowledge gives us a new understanding of the world cities and cultures. More practically, it can help travel recommendation services by pushing locations similar to persons' preferences.

## 4.2 Image Relevance

We start by examining the overall relevance of images belonging to a location. We calculated entropy of players' image selection. We defined entropy as equation (2), where probability *p(i)* stands for the probability of selecting image *i* to play among all images belonging to *location*.

$$H(p) = -\sum_{i \in location} P(i) \cdot log_2 P(i) \qquad (2)$$

Entropy represents information outcome associated with the image selection, that is, the larger entropy means most images of a location were selected with similar probability, while the smaller entropy means some particular images were frequently selected than others.

The locations with a large value of entropy were the National Stadium (the value of entropy is 4.156), the National Aquatics Center (4.131), and the Summer Palace (3.907). The overall image relevance of these locations is high, that is, the image datasets contain many relevant images. The locations with a low value of entropy are North Sea Park (2.470) and Shanghai Center (2.477). For these locations, we have to think that their image datasets contain many irrelevant images. Based on entropy values, we can compare how many relevant images locations have. It is natural that the larger the number of images taken at a location, the larger the number of relevant images grows. Therefore, entropy is one index of the location's popularity.

To assess geographical relevance, we have three kinds of data extracted from our game logs; number of times of image selection, average number of clicks and average discrepancy distance of an image. Here, the average discrepancy distance is calculated separately for each image, as it is the sum of the discrepancy distance divided by the number of times of image selection.

We calculated the Pearson's product-moment coefficient between the number of times of image selection and average discrepancy distance per image, and found that they do not correlate with each other (coefficient of correlation is -0.0787). This is because the average discrepancy distance is affected by other external elements, such as players' knowledge and popularity of locations. Regarding the average number of clicks on the image, the smaller it is, the higher the relevance becomes, since players can infer the

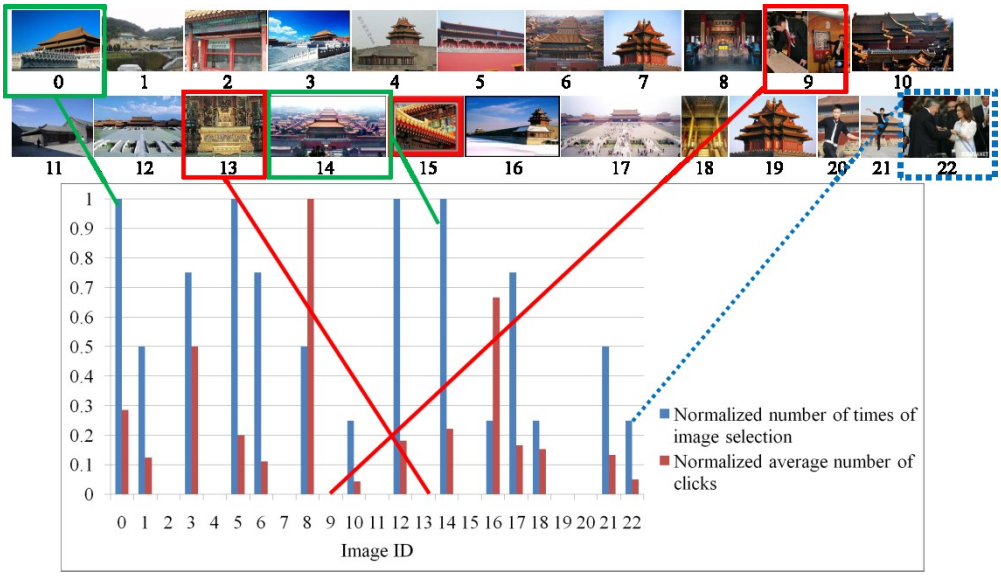


**Figure 7. Normalized number of times of image selection and normalized average number of clicks on Imperial Palace**

location by only viewing small portion of the image. Therefore, we can use the combination of number of times of image selection and clicks to decide the geographical relevance of images.

Here we show an example. Figure 7 presents the images on Imperial Palace used for the game, and a graph showing normalized number of times of image selection and normalized average number of clicks, of which we describe the definitions in Section 5.1.1. Both values' ranges are from 0 to 1, and a higher value is better. The images framed with green rectangles (images 0 and 14) can be regarded as highly relevant since they were selected to play frequently. Images framed with red rectangles (image 9 and 13) were not selected at all. As you can see, these images are irrelevant to Imperial Palace. Moreover, the normalized number of times of image selection on the image framed with a dotted blue rectangle (image 22) can be regarded as noise, since its normalized average number of clicks is quite small, that is, the Captain opened large part of the image since the Sailor cannot infer the location. Actually image 22 does not contain a figure of Imperial Palace at all.

In Section 5.1, we propose methods to decide the geographical relevance based on our game logs and conduct an experiment to verify their performance.

## 4.3 Image Region Relevance

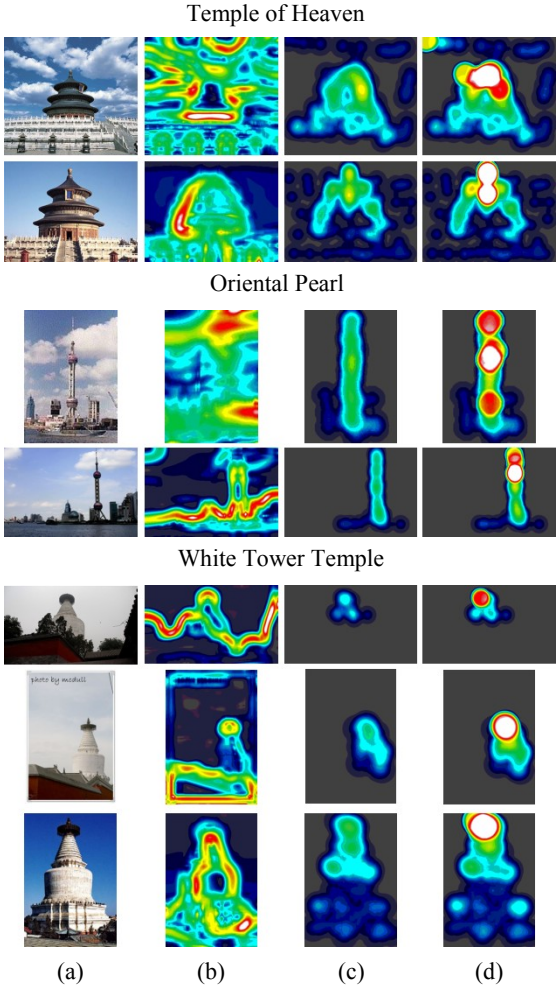


**Figure 8. Heat map examples; (a) original image (b) saliency map generated by local contrast analysis [21] (c) neutral heat map (d) weighted heat map**

In the game, the Captain opens image regions to the Sailor so that the Sailor can infer the location as quickly as possible to get a high score. Therefore, the opened image regions are the portion of high geographical relevance. In addition, because of the time constrain of the game, players usually open the most important regions first. Therefore, the initial opened regions can have more importance than regions that were opened later.

Figure 8 shows heat maps of players' opened image regions on the Temple of Heaven, Oriental Pearl and White tower temple, in which column (a) is original images, column (b) is conventional saliency maps generated by local contrast analysis [21], column (c) is *neutral heat maps*, and column (d) is *weighted heat maps*. Here, we define the *neutral heat map* as it considers all opened regions have the same weight and the *weighted heat map* as it weighs initial opened regions by each player. Specifically, we applied a power-law function on relation between players' order of clicks and its geographical relevance. The weight of the *i*th click is determined based on equation (3), where we set $a$ and ε as constants to adjust to range of heat, and set $k = 2$.

$$weight_i = \frac{a}{i^k} + \varepsilon \quad (3)$$

Based on a neutral heat map, we can detect characteristic objects of a location, that is, geographically relevant objects from the image. Furthermore, a weighted heat map reveals the most salient portion of the objects. In Figure 8, the top roof and the plate of the Temple of Heaven, the skewered red balls of the Oriental Pearl, and the top hat of the White Tower Temple are apparently *hot* portions, as people can recognize the whole object when they see that portion. These salient portions are useful for a number of applications, such as image location recognition and content matching as we discuss in Section 5.2. It is difficult to automatically detect the salient portions since they are independent from image features.

As Figure 8 (b) shows, conventional saliency maps cannot accurately detect the geographically relevant objects. They can detect the objects when the original images contain only the objects and the backgrounds are clear, such as second row of Temple of Heaven and Oriental Pearl and the third row of White Tower Temple. However, when the backgrounds are complicated, conventional saliency maps detect other unimportant objects as salient, such as clouds in Temple of Heaven and Oriental Pearl examples and trees and houses' roofs in White Tower Temple example. These saliency maps are generated based on low-level features of images, and thus, cannot indicate a relevant object. Furthermore, detecting objects' salient portions is almost impossible, since they are not distinguishable from low-level features and it is only humans that know the answer.

Additionally, our neutral and weighted heat maps accurately catch the geographically relevant objects and its salient portions, even though the sizes of the objects, the camera angles and backgrounds are different, since they are based on human perception.

## 4.4 Questionnaire Survey

We conducted a questionnaire survey to examine the quality of our game, regarding the game log and entertainment perspective. We sent the following questions to the top 55 active players by e-mail, and received answers from 48 of them.

1. What were the criteria to choose an image to play?
2. What were the criteria to select image regions to show the other player?

3. How did you feel using a map for the game compared to directly inputting the name of locations?
4. Do you think we can improve the game? Let us know your suggestions.
5. Do you want to play our game again?

Regarding the first two questions, all 48 respondents answered that they chose most relevant images for playing locations and opened their characteristic objects. They said that these were the most reasonable way to let their partner to recognize the locations and earn a high score. These results prove that we can basically rely on players' choices, that is, players' image selections directly reflect the geographical relevance of the images and their opened regions can be regarded as relevant objects on locations. However, two respondents commented that if the location was quite famous, they randomly chose images because the location was easy enough to recognize. Additionally, another three respondents also commented that they sometimes opened empty image regions to confuse their partners for fun. We can eliminate these noises easily by aggregating other large numbers of reliable game logs.

The third question is to ask the availability of a map for a game, which is one of our game's characteristic features. All respondents preferred to use a map, since clicking the surface of a map is more intuitive and easy than inputting text. Additionally, three respondents said that finding a location from a map was more challenging but it roused them to play the game.

From the fourth question, we received two major suggestions. First, 19 respondents suggested to add more cities to play. They wanted to play the game on other cities worldwide. Similarly, 15 respondents insisted that the number of images on each location should be increased because they wanted to see brand-new images anytime they selected the locations they had played before.

Another interesting direction is to enable players to add locations and images. As discussed in Section 2, there are some commercial game services that use geography. Considering the players on these games are actively creating their own locations, we can also expect that our players are willing to create locations and upload images. It will increase the scalability of our game data and we can collect images' geographical relevance not only for famous locations but also local ones.

The second suggestion is to dynamically display the other pairs' score, who are playing the game at the same time. In the game, the Captain and Sailor have to cooperate with each other to get a high score. They insisted that cooperation itself was enjoyable and it would be better to cooperate with the partner to win other pairs. We agree with this idea since competition is also effective to encourage players. We will add this function to the game in future.

To the fifth question, 31 respondents answered positive, while 13 answered negative and remaining four were neutral. 10 of the positive respondents commented that they could learn the famous locations of the city through the game. Although it is an unexpected effect of our game, we think that it will leverage players to play the game more. On the other hand, 15 out of the negative and neutral respondents insisted that they played the game many times and have already got familiar with locations and their images. However, they remarked that if the cities and images were increased, they would play the game again since it is fun. Given this remark, we can expect that our game, with enough amounts of locations and images, can attract players when it will be released as a public game service.

Since the ESP game's target is general Web images, its interaction cannot help being simple, while our game's target is specific. Therefore, we can design a more elaborate game. As the results of the questionnaire shows, map based interaction is intuitive to play, and players can show their rich knowledge of their geographical sense to other players. Furthermore, players can also learn additional knowledge on locations. These features are effective to attract people to play the game over a long period of time.

# 5. APPLICATIONS

## 5.1 Re-ranking Image Search Results

Image search engines use image captions and surrounding texts etc. to rank images, and thus, it is difficult to achieve reasonable ranking conforming to human perception. In this section, we propose methods to assess image geographical relevance and examine how they can improve current image search engines.

### *5.1.1 Methods Based on Single Feature*

As discussed in Section 4.2, we can derive the number of times of image selection, average number of clicks, and average discrepancy distance from our game logs, which we can expect to be effective to assess geographical relevance of images. We normalize them based on the following definition, and assign as geographical relevance to rank images. The range of all methods is from 0 to 1, and we hypothesize the larger the value is, the higher an image's geographical relevance is.

- Normalized number of times of image selection (*Freq*)

$$Freq = \frac{\text{Num. of times of image sel.}}{\text{Max. num. of times of image sel. on the location}} \quad (4)$$

- Normalized average number of clicks (*Click*)

$$Click = \frac{\text{Min. average num. of click on the location}}{\text{Average num. of click on an image}} \quad (5)$$

- Normalized average discrepancy distance (*Dist*)

$$Dist = \frac{\text{Min. average discrepancy distance on the location}}{\text{Average discrepancy distance of an image}} \quad (6)$$

To examine effectiveness of these methods, we conducted an experiment. We selected 15 frequently played locations and manually assigned ground-truth geographical relevance to images belonging to each location based on the following criteria.

- 4 (relevant): The main objects of an image are the landmark at a location and it is highly visible.
- 3: The main objects of an image are the landmark, but the image is blurred or the camera angle is not good.
- 2: The landmark is in an image, but not the main objects or there is high level of occlusion or distortion.
- 1: The image seems somehow related to the landmark, but the landmark does not appear on an image.
- 0 (irrelevant): The image is not related to the landmark.

We rank images on each location according to their geographical relevance determined by our methods. After that, we calculate the normalized discounted cumulative gain (NDCG) of each method using ground-truth relevance based on equation (7), where *n* is the number of images, $rel_i$ is the relevance value of the *i*th image, and *f(n)* is a normalization factor.

$$NDCG = \frac{1}{f(n)}\left(rel_1 + \sum_{i=2}^{n}\frac{rel_i}{log_2 i}\right) \quad (7)$$

To compare with a current image search engine, we conducted image search using Live Image Search [20] with the same query when we collected our images. We assign ground-truth geographical relevance to the images obtained by Live Image Search, and compare NDCG with our methods.

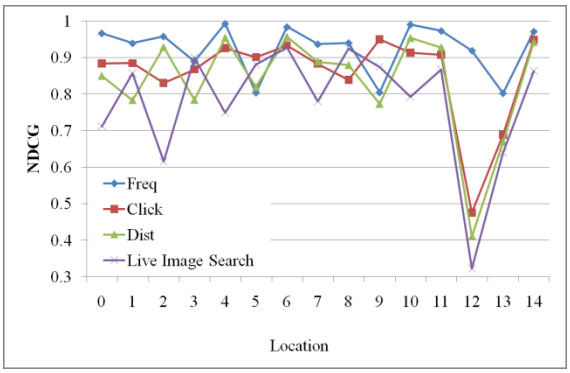


**Figure 9. NDCG of our methods and Live Image Search**

Figure 9 shows the graph of each method's NDCG. It is apparent that all of our methods outperform Live Image Search. Average NDCG is 0.9244, 0.8547, 0.8344 and 0.7799 for *Freq*, *Click*, *Dist*, and Live Image Search, respectively. This result shows that *Freq* achieves 0.1445, *Click* 0.0748, and *Dist* 0.0545 higher NDCG than Live Image Search on average. Among these methods, *Freq* especially performs the best. This is because players' interaction affecting the method is simpler, which is selecting a relevant image from candidate images. In other words, *Freq* is less dependent on players' knowledge of locations, e.g., players can select a relevant image even if they do not know its location. However, for *Click*, the interaction is selecting image region until the Sailor can make a decision of its location. That means if the Sailor is not familiar with the location, the Captain has to open many regions. Similarly, for *Dist*, players' knowledge and popularity of locations strongly affect its result.

However, *Freq* sometimes shows drops, such as locations 3, 5 and 9. This is because players tend to prefer images with a landmark's name, which is not always geographically relevant. Additionally, as we discussed in Section 4.2, players' image selection are sometimes noisy. *Click* achieves more stable performance since it is based on two players' consensus, though its NDCG is lower than *Freq*. Therefore, we can expect the combination of *Freq* and *Click* will improve the performance of *Freq* only. Among these three methods, *Dist* looks unstable. This is because *Dist* is directly affected by players' knowledge of locations and cannot be always proportional to geographical relevance.

Here, on location 12, the NDCG of *Click*, *Dist* and Live Image Search sharply drop. The location is Yuyuantan Park which is famous for facing a lake and cherry blossoms but lacks remarkable buildings and towers. Therefore, the data of number of clicks and the discrepancy distance became noisier than other locations. However, the NDCG of *Freq* is still high on this location, since to select a relevant image viewing whole figure is a more simple interaction than the other two. For Live Image Search, we can find many images taken *inside* the park because such images' captions and surrounding text contain the park's name, but there are few images focusing the park itself.

### *5.1.2 Methods Based on Combination Features*

In the former section, we confirmed that *Freq* drastically improves the image search results. However, it sometimes performs in an unstable manner, and thus, we improve it by combining it with *Click*. Since *Freq* basically outperforms *Click*, we have to lay stress on the *Freq* value to keep up overall performance. Therefore, we use the *Click* value to decide the rank for images with similar *Freq* values. We set a threshold to determine similar *Freq* values. If multiple images have similar *Freq* values, an image with a higher *Click* value is ranked higher. In the following, we call this method *Combination*.

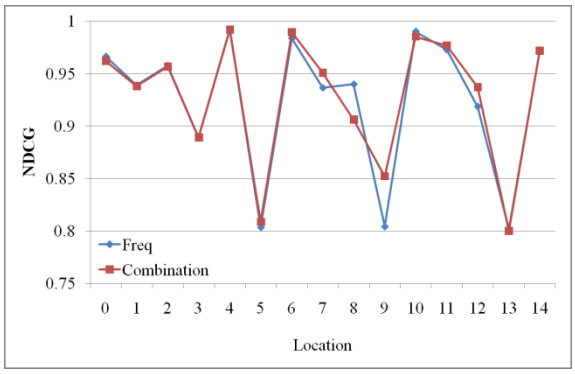


**Figure 10. NDCG of Freq and Combination**

Figure 10 shows the graph of the NDCG of *Freq* and *Combination*. The average NDCG is 0.9244 for *Freq* and 0.9277 for *Combination*. Although *Freq* achieves quite a high NDCG, *Combination* can achieve an even higher NDCG. Additionally, since our aim on *Combination* is to make the performance to be stable, we pay attention to minimum and median values of NDCG. Minimum and median values of NDCG are 0.8018 and 0.9400 for *Freq*, while 0.7997 and 0.9508 for *Combination*. These result shows that *Combination* can stabilize *Freq* with only a tiny loss.

The amount of our game logs is not large and our players behaved comparatively politely since the study was conducted internally in MSRA. Therefore, the percentage of noisy logs should be quite small compared to when releasing the game as a commercial game service. Therefore, the improvement by *Combination* is not so large on our logs, but we think that it will perform powerfully when releasing the game to public.

## 5.2 Image Location Recognition and More Applications

As we described in Section 2, image location recognition has been studied for decades in the computer vision field. They use various low-level features extracted from images to match a query image with one of the training dataset images and infer the location. Most works set aside object recognition techniques though it can improve the accuracy and efficiency, since automatic object recognition is still under research and manually labeling images' objects is too costly and lacks scalability. Our neutral and weighted heat maps, derivable from the game logs, can also help image location recognition. A neutral heat map can tell geographically relevant image regions, and a weighted heat map can tell their most salient portions, which should be considered as important. The *cool* regions on these heat maps can be ignored. Therefore, incorporating our heat maps can boost the location recognition accuracy and reduce computational cost. In our approach, people only enjoy the game and do not have to perform extra tasks for object recognition, and thus, we can expect to collect large amounts of game logs and generate heat maps when the game will be released as a public game service.

In addition, our neutral and weighted heat maps can be useful to various applications. They can help image content matching techniques, since they can tell the main object regions of an image.

It is also useful for image compression by firmly compressing unimportant regions and keeping the quality for important regions. Moreover, they can achieve smart image thumbnail generation. Since current image thumbnails simply minimize images, their visibility is low. Instead of minimizing the whole image, by cropping the important regions and minimizing it, thumbnails can show representative objects with better resolution. It is especially effective when presenting images on small screens such as a cellular phone's display, where techniques to effectively present large information on limited space are required.

# 6. CONCLUSION

Current image search engines rank images based on their <alt> attributes, captions, etc. instead of analyzing the image content. Due to lack of information of the image itself, it is difficult to achieve reasonable ranking conforming to a human point of view. However, there is still a big gap between low-level features used for image analysis and the semantic meaning of images. In this paper, we focused on image geographical relevance, since geographical context is important for image search, considering many popular Web services started to use the geographical information of images for their services, such as Flickr and Google Earth. We followed the concept of human computation and designed a game to collect information to decide image geographical relevance. We thoroughly analyzed the game logs obtained by a three week study and proposed methods to assign geographical relevance. As a result of the experiments, we confirmed that our methods can improve a commercial image search engine regarding NDCG by about 15 points on average. Furthermore, we showed that we can derive geographically relevant objects and their salient portions in images based on our game logs, which is useful for various applications such as image location recognition and thumbnail generation.

We plan to extend our game to allow players to compete with each other, since competition is a killer factor when attracting players. Although we currently use only positive feedback from players, such as image selection and clicks on images, negative feedback, such as voting for non-representative images, might also be useful to assess image geographical relevance. Moreover, we plan to add more cities and images to the game and release it to broader audiences. We will carry out experiments to study how the game logs can help more applications.

# 7. REFERENCES


[1] von Ahn, L. Games with a purpose. IEEE Computer, Vol. 39, Issue 6, pp. 92-94. 2006.

[2] von Ahn, L., Liu, R. and Blum, M. Peekaboom: A game for locating objects in images, In Proc. of CHI'06, pp. 55-64.

[3] von Ahn, L., Kedia, M. and Blum, M. Verbosity: a game for collecting common-sense facts, In Proc. of CHI'06, pp. 75-78.

[4] von Ahn, L. and Dabbish, L. Labeling images with a computer game. In Proc. of CHI'04, pp.319-326.

[5] Brockett, C. and Dolan, C. B. Echo Chamber: A game for eliciting a colloquial paraphrase corpus, In Proc. of KCVC'05.

[6] Carneiro, G. and Vasconcelos, N. A database centric view of semantic image annotation and retrieval, In Proc. of SIGIR'05, pp. 559 – 566.

[7] Dugan, C., Muller, M., Millen, D. R., Geyer, W., Brownholtz, B. and Moore, M. The dogear game: a social bookmark recommender system, In Proc. of Intl. ACM conf. on Supporting group work (2007), pp. 387-390.

[8] Epshtein, B., Ofek, E., Wexier, Y. and Zhang, P. Hierarchical Photo Organization using Geo-Relevance, In Proc. of ACM GIS'07.

[9] Flickr < http://www.flickr.com/ >

[10] Geocaching <http://www.geocaching.com/>

[11] Google Earth < http://earth.google.com/ >

[12] GPS Mission <http://gpsmission.com/ >

[13] Hays, J. and Efros, A. A. IM2GPS: Estimating geographic information from a single image. In Proc. of CVPR'08, pp. 1-8.

[14] He, X., Cai, D., Wen, J.-R., Ma. W.-Y. and Zhang, H.-J. ImageSeer: Clustering and searching WWW images using link and page layout analysis. Microsoft Technical Report, MSR-TR-2004-38, 2004.

[15] Huang, Y., Shekhar, S. and Xiong, H. Discovering colocation patterns from spatial data sets: a general approach. IEEE TKDE, Vol. 16, No. 12, pp. 1472-1485, 2004.

[16] Jing, Y. and Baluja, S. PageRank for product image search. In Proc. of WWW'08, pp. 307-315.

[17] Ke, Y., Tang, X. and Jing, F. The design of high-level features for photo quality assessment. In Proc. of CVPR'06, pp. 419-426.

[18] Lavrenko, V., Manmatha, R. and Jeon, J. A model for learning the semantics of pictures, NIPS'03.

[19] Lempel, R. and Soffer, A. PicASHOW: Pictorial authority search by hyperlinks on the Web. In Proc. of WWW'01, pp. 438-448.

[20] Live Image Search <http://www.live.com/?scope=images>

[21] Ma, Y.-F. and Zhang, H.-J. Contrast-based image attention analysis by using fuzzy growing. In Proc. of MM'03, pp. 374-381.

[22] Shamma, D. A. and Pardo, B. Karaoke callout: Using social and collaborative cell phone networking for new entertainment modalities and data collection, In Proc. of ACM workshop on Audio and music computing multimedia (2006), pp. 133-136.

[23] Voutsakis, E., Petrakis, E. G.M. and Milios, E. Weighted link analysis for logo and trademark image retrieval on the web. In Proc. of WI'05, pp. 581-585.

[24] Wang, L., Wang, C., Xie, X., Forman, J., Lu, Y., Ma, W.-Y. and Li, Y. Detecting Dominant Locations from Search Queries, SIGIR'05, pp. 424 -431.

[25] Wang, X.-J., Zhang, L., Jing, F. and Ma, W.-Y. AnnoSearch: Image auto-annotation by search. In Proc. of CVPR'06, pp. 1483-1490.

[26] Zhang, W. and Kosecka, J. Image based localization in urban environments. In Proc. of 3DPVT'06, pp. 33-40.