Detecting Cultural Differences using Consumer-Generated Geotagged Photos

Keiji Yanai Keita Yaegashi
The University of Electro-Communications
Chofu, Tokyo 182-8585 JAPAN
{yanai,yaegas-k}@mm.cs.uec.ac.jp

Bingyu Qiu
Beijing University of Posts and Technology
Beijing, 100876, China
bigyuqiu@bupt.cn

ABSTRACT
We propose a novel method to detect cultural differences over the world automatically by using a large amount of geotagged images on the photo sharing Web sites such as Flickr. We employ the state-of-the-art object recognition technique developed in the research community of computer vision to mine representative photos of the given concept for representative local regions from a large-scale unorganized collection of consumer-generated geotagged photos. The results help us understand how objects, scenes or events corresponding to the same given concept are visually different depending on local regions over the world.

Keywords
geotag, object recognition, representative image, Flickr

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Miscellaneous

1. INTRODUCTION
Recently, consumer-generated media (CGM) on the Web has become very popular. Especially, photo sharing sites such as Flickr and Picassa are representative CGM sites, which store a huge number of consumer-generated photos people uploaded, and make them accessible via the Web for everyone. Photo sharing sites collect not only photos but also metadata on uploaded photos. As metadata users add to photos, textual information such as keywords and comments is common. Recently, in addition to texts, some users attach “geo-tags” to their uploaded photos. Note that a “geo-tag” means metadata which represents a location where the corresponding photo was taken, which is expressed by a set of a latitude and a longitude.

An accurate geotag can be obtained with a GPS device or a location-aware camera-phone. However, since it forces us to use relatively special devices, GPS-based geotags have not been common so far. Instead, map-based geotags have become common, after Flickr, which is the largest photo sharing site in the world, launched an online geotagging interface in 2006. Then, Flickr also became the largest “geotagged” photo database in the world. According to [6], there are currently over 40,000,000 public geotagged photos on Flickr, and 100,000 geotagged photos have been added every month. These geotagged photos would be valuable not only for browsing and finding individual concepts, but also for helping us understand how local specific objects or scenes over the world are different.

Our objective is thus to facilitate a system which can automatically select relevant and representative photographs for the general object or scene concepts in the worldwide dimensions. In particular, we consider geotagged photos on Flickr, identify the representative image groups, and generate an aggregate representation based on locations that allows navigation, exploration and understanding of the differences of general concepts depending on local locations in the world visually. We employ the state-of-the-art object recognition technique developed in the research community of computer vision to mine representative photos of the given concept for each region from a large-scale unorganized collection of consumer-generated geotagged photos. The results help us understand how objects, scenes or events corresponding to the same given concept are visually different depending on local regions over the world.

2. APPROACH
Our approach for selecting the representative images for representative local regions from geotagged images consists of three main stages (Figure 1): (1) removing irrelevant images to the given concept, (2) estimating representative geographic regions, and (3) selecting representative images for each region.

First, we apply clustering techniques to partition the image set into similar groups, based on bag-of-visual-words feature vectors [1]. By evaluating the intra-cluster densities as well as the cluster member numbers, we discard most of the irrelevant images and obtain a reduced set of images which are visually similar each other. This stage could be regarded as the “Filtering Stage”.

Then, we geographically cluster the reduced set of images and select large geographic clusters as representative regions. Here we use the k-means clustering algorithm based on the geographic latitude and longitude of photos to obtain representative regions in the world for the given concept.

Finally, for each representative region, we perform the Probabilistic Latent Semantic Analysis (PLSA) [4] to identify the distinct “topics”, do additional clustering on the entire topic vectors, and select the “significant” cluster as the representative results for this geographic region. In addition, with the help of map service,
2.1 Filtering Irrelevant Images

2.1.1 Image Representation

We adopt bag-of-visual-words model from [1] as the image representation. This model was first proposed for the text document analysis and recently applied in visual object recognition which has been found to be extremely powerful in tasks of representing the image features. The construction of bag-of-visual-words feature vectors for images involves the following steps: (1) a set of points of interest are automatically detected in the image and local descriptors are computed over each point; (2) all the descriptors are quantized to form visual words; (3) for each image, we count the occurrences of each visual word to form a histogram of visual words which can be regarded as a bag-of-visual-words feature vector.

In our experiment, we first apply grid-based policy to detect the points of interest, and then compute the local descriptors by the Scale Invariant Feature Transform (SIFT) descriptor [7]. The SIFT descriptors are computed at 8 orientation directions over a $4 \times 4$ parts of spatial location, forming 128-dimensional vectors. Then we apply the k-means clustering algorithm over all extracted descriptors and computer the means to form visual words. Here we tried $k = 500$ and form a vocabulary of size 500. Finally, for each image, we assign all SIFT vectors to the nearest visual word and convert these vectors into one k-bin histogram which represents the bag-of-visual-words feature vector.

2.1.2 Visual Clustering

After building bag-of-visual-words representation for all raw images, we perform clustering using k-means algorithm over the bag-of-visual-words feature vectors to partition images into similar groups. In order to ensure a clear partition, we choose a high number of clusters $k$ ($\approx 200$ clusters for a dataset of about two thousand images). Since most irrelevant and visually unrelated photos tend to fall into the small clusters, we can discard such small clusters based on a minimum threshold (usually less than 10 cluster members in our experiment).

2.1.3 Selecting the Most Relevant Clusters

Since there may still exist some clusters with large noises (irrelevant images), in order to detect such irrelevant clusters and select the most relevant clusters, we employ the method of evaluating the intra-cluster similarity for the remaining clusters. The intra-cluster similarity is the average similarity between the images that belong to the cluster and the similarity between two images $P_i$ and $P_j$ can be calculated using the cosine metric between two image vectors $V_i$ and $V_j$:

$$sim(P_i, P_j) = \frac{V_i \cdot V_j}{\sqrt{||V_i|| \cdot ||V_j||}}$$

Then given a cluster of $n$ photos, $C = \{P_1, \ldots, P_n\}$, we can define the intra-cluster similarity as:

$$SIM(C) = \frac{\sum_{P_i, P_j \in C, i \neq j} sim(P_i, P_j)}{n(n-1)/2}$$

which denotes the average similarity between two photos within one cluster.

By computing the intra-cluster similarity value for each cluster and sorting all clusters in the descending order of the SIM values, we select several top ones as the most relevant clusters (We selected 40 clusters in our experiments).

2.2 Detecting Representative Regions

In this stage, given the remaining most relevant photos, we attempt to detect representative regions based on the photos’ geographic locations. For simplicity, we perform k-means clustering algorithm, based on the photos’ geographic latitude and longitude (with the help of geo-tags), using geographical distance as the distance metric. Then we select several largest geo-clusters to form the representative regions since they have more relevant photos and the number of photos taken in a region is an indication of the relative importance of that region for the particular concept. (In our experiment, for simplicity, we generally select about 4 or 5 representative regions for each concept.)

2.3 Generating Representative Photographs

At this point, we have obtained the most relevant or visually similar photos, and the corresponding representative regions. To generate a set of representative photos for these representative regions, we explore the Probabilistic Latent Semantic Analysis (PLSA) [4] model, which is recently applied to recognize object categories in an unsupervised manner.

As a generative model, PLSA was originally used to discover latent topics in the text documents represented by bag-of-words.
In a similar consideration, since images can be regarded as “documents” and represented by bag-of-visual-words, hence PLSA can be applied to images for discovering the object categories in each image. In terms of images, suppose we have a set of images \( D = \{ d_1, \ldots, d_n \} \), each containing the visual words from the visual vocabulary \( W = \{ w_1, \ldots, w_m \} \). By introducing a mediator known as latent topics \( Z = \{ z_1, \ldots, z_k \} \), we can build a joint probability model over images and visual words, defined as:

\[
P(w, d) = P(d) \sum_{z \in Z} P(w|z)P(z|d)
\]

where every image is modeled as a mixture of topics, \( P(z|d) \), and \( P(w|z) \) represents probability occurrence of visual words within a topic. We can learn the unobservable mixture parameters \( P(z|d) \) and topic distributions \( P(w|z) \) by the EM algorithm. Refer to [4] for a full explanation of the PLSA model.

As in our experiment, for each representative region, we apply the PLSA method to all the photos belonging to the region with a given number of topics, and get the probability distributions of all topics over each image, \( P(Z|d) \), which can be regarded as topic vectors to represent an image. In the experiment, the number of topics was set to 20. After that, we aggregate photos according to the distributions of mixture topics by doing an additional step of clustering the topic vectors, \( P(Z|d) \). In our experiments, we obtained the best results by applying k-means clustering with \( k = 5 \). Then the set of photos in the largest cluster are selected as representative photos of the given region, which is the final output of the proposed system.

### 3. EXPERIMENTAL RESULTS

To test and verify if our approach works in practice, we conducted preliminary experiments with photos collected directly from Flickr. In the experiments, we used seven “object” concepts and two “scene” concepts including “noodle”, “wedding cake”, “flower”, “castle”, “car”, “waterfall” and “beach”. For each concept, we collected about 2000 most relevant geotagged photos distributed evenly in the world wide areas. The precision of raw photos of these seven concepts is 43% on average, which is defined as \(# relevant\ images\)/(# all\ images\). After “Filtering Stage”, it was improved to 80%, which indicated that noise removal was effective.

We show the representative photos selected for several representative regions, while these regions were generated automatically based on geographic locations of the most relevant photos selected in the “Filtering Stage”. Figure 2 and Figure 3 show the results for the concept “noodle”, each of which presents the most representative photos generated for the approximate regions: Japan and Europe. Without doubt, these results can help us understand about the “noodle” in these local areas. For example, Figure 2 demonstrates many “ramen” photos in Japan and Figure 3 demonstrates “spaghetti” photos in the European area. In addition, South East Asia, Mideast US and Western US are obtained as other representative regions, representative photos of which also have characteristics such as “noodles” in the South East Asia area containing some Taiwanese style noodles and spicy Thai noodles.

Figure 4 and Figure 5 correspond to “wedding cake” in Europe and in Mid US, respectively. We can find many of the wedding cakes in Mid US are much taller than ones in Europe.

For the scene concept “waterfall”, we extracted the representative photos for four large regions: Asia, Europe, North America, and South America. Due to the space limitation, we show only two
4. RELATED WORK

Several recent researches have considered the problem of selecting representative or canonical photographs for online image collections. Jaffe et al. [5] selected a summary set of photos from a large collection of geo-tagged photographs based on only tags and geotags. By analyzing the correlations between tags and geotags, a map-based visualization “Tag Map” was developed to help indicate the most important regions and the concepts represented in those regions. Our work similarly identifies the most important regions and select representative photos for these regions. A key difference is that [5] used only tags and geotags, while in our work, we aim to select representative photographs for particular concepts by applying computer vision techniques. Simon et al. [11] have proposed a method to select canonical views for the landmarks by clustering images based on the visual similarity between two views. Like [11], Kennedy et al. [6] attempted to generate representative views for the world’s landmarks based on the clustering and on the generated link structure. Unlike the works [11] and [6], we choose the general concepts on objects or scenes as our target, but not the identical objects like landmarks which rely on 3D structure or viewpoint. Our goal is to select representative photos for geographic regions in the worldwide dimensions, and we concentrate on the general concrete concepts such as “noodle” and “wedding cake” as well.

Another similar work is from Raguram et al. [10]. They aim to select iconic images to summarize general visual categories, like “love”, “beauty”, “closeup” and “apple”. Since general visual or abstract concepts usually have many semantic “themes”, their canonical view selection is hence defined as select a small number of salient images for each semantic “theme”. Our goal is different as we aim to select representative photos for geographic regions in the worldwide dimensions, and we concentrate on the general concrete concepts such as “noodle” and “waterfall” as well [8]. On the other hand, [9] treated with generic concepts. However, we select canonical images on generic concepts regarding several regions in the worldwide dimension, while [9] treated with general concepts within only given regions. “IM2GPS” project [3] proposed a unique idea of estimating a place from just one non-geotagged image with 6 million geotagged images gathered from Flickr by comparing image similarity, the objective of which is different from ours.

Edwardes et al. examined which words are suitable for representing geographical concepts using online photo database [2]. Before that, subjectives were used in this kinds of research [12]. However, we do not limit the targets for only geographical concepts, but we can treat with generic concepts.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we present a novel topic which is for the purpose of generating representative photographs for typical regions in the world, and provide an approach to achieve it with the help of geo-tagged collections. The results help us understand how objects or scenes associated with the same concept are different depending on local regions in the world visually.

For future work, we plan to make extensive experiments for more concepts from a larger set of photos, and think out some other strategies in detecting more representative regions with a more precise scope. In addition, we will conduct some quantitative evaluations on the representativeness of the photos selected for the corresponding region and the differences of the tendency of representative photo sets among different local regions.

Acknowledgment

We appreciate helpful comments from the anonymous reviewers. This research was partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Exploratory Research, 19650027, 2008, and by Kayamori Foundations.

6. REFERENCES