A Visual Analysis of the Relationship between Word Concepts and Geographical Locations

Keiji Yanai Hidetoshi Kawakubo
The University of Electro-Communications
Chofu, Tokyo 182-8585 JAPAN
{yanai,kawaku-h}@mm.cs.uec.ac.jp

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

ABSTRACT

In this paper, we describe two methods to analyze the relationship between word concepts and geographical locations by using a large amount of geotagged images on the photo sharing Web sites such as Flickr.

Firstly, we propose using both image region entropy and geo-location entropy to analyze relations between location and visual features, and in the experiment we found that concepts with low image entropy tends to have high geo-location entropy and vice versa.

Secondly, we propose a novel method to select representative photographs for regions in the worldwide dimensions, which helps detect cultural differences over the world regarding word concepts with high geo-location entropy. In the proposed method, at first, we extracts the most relevant images by clustering and evaluation on the visual features. Then, based on geographic information of the images, representative regions are automatically detected. Finally, we select and generate a set of representative images for the representative regions by employing the Probabilistic Latent Semantic Analysis (PLSA) modelling. The results show the ability of our approach to mine regional representative photographs and cultural differences over the world.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Miscellaneous

General Terms

Algorithm, Experimentation

Keywords

geotag, entropy, representative image, Flickr

1. INTRODUCTION

In these days, there are a great number of images on the Web. Especially, photo sharing sites such as Flickr and Picassa are gathering a large number of photo images. They store a huge number of consumer-generated photos people uploaded, and make them accessible via the Web for everyone. Photo sharing sites collect metadata in addition to photos. While keywords and comments are common as metadata, recently some users attach "geo-tags" to their uploaded photos. A "geo-tag" means metadata which represents a location where the corresponding photo was taken, which is usually 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 [11], 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 specific objects or scenes are distributed and different over the world.

Our objective is exploring the relationship between word concepts and geographical locations by using a large amount of geotagged images on the photo sharing Web sites such as Flickr.

Firstly, we propose using both image region entropy and geo-location entropy to analyze relations between location and visual features, and in the experiment we found that concepts with low image entropy tends to have high geo-location entropy and vice versa.

Secondly, we propose a novel method to select representative photographs for regions in the worldwide dimensions, which helps detect cultural differences over the world regarding word concepts with high geo-location entropy. In the proposed method, at first, we extracts the most relevant images by clustering and evaluation on the visual features. Then, based on geographic information of the images, representative regions are automatically detected. Finally, we select and generate a set of representative images for the representative regions by employing the Probabilistic Latent Semantic Analysis (PLSA) modelling. The results show the ability of our approach to mine regional representative photographs and cultural differences over the world.

2. RELATED WORK

Until several years ago, researches on geotagged images focused on only location-based photo browsing for a personal geotagged photo collection [21, 15], since it is almost impossible to obtain a large number of geotagged images. However, the situation has been changed after Flickr launched an online geotagging interface in 2006. At the present, Flickr has become the largest geotagged

CIVR '09, July 8-10, 2009 Santorini, Greece.

photo database in the world. Geotagging with GPS devices is too expensive to spread, but Flickr online geotagging system allows users to indicate the place where photos are taken by clicking the online map. In addition, the Flickr database is open to everyone via FlickrAPI which allows users' program to search the whole Flickr photo database for geotagged images.

Therefore, some works on geotagged image recognition with huge Flickr geotagged image database has been proposed recently. [4] and [3] proposed methods on event recognition of geotagged images by integrating visual features and geographical information. In general, a geotag represents a pair of values on latitude and longitude. It is a just 2-dimensional vector. To convert a 2-d vector into more rich representation, [14] and [22] converted geotags into visual information from the sky using aerial images, and [10] transformed geotags to words using reverse geo-coding technique. On the other hand, [26] used GPS trace data which is a series of geotags instead of using just a pinpoint geotag in order to classify images into several pre-defined events. [25] used time and seasons for geotagged image recognition in addition to visual information and geo-location data. While event or scene recognition on geotagged images is common, "IM2GPS" project [7] proposed a unique idea of estimating a place from just one non-geotagged image with 6 million geotagged images gathered from Flickr.

As extension of location-based photo browsing, several recent researches have considered the problem of selecting representative or canonical photographs for online image collections. Jaffe et al. [9] select a summary set of photos from a large collection of geotagged photographs based on only tags and geo-tags. By analyzing the correlations between tags and geo-tags, a map-based visualization "Tag Map" is 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 in [9], the concepts are learn which could be mostly affected by users' photographic behavior. While in our work, we aim to select representative photographs for particular concepts by applying computer vision techniques. Simon et al. [20] have proposed a method to select canonical views for the landmarks by clustering images based on the visual similarity between two views. Like [20], Kennedy et al. [11] attempt to generate representative views for the world's landmarks based on the clustering and on the generated link structure. Unlike the works [20] and [11], we choose the general category objects or scenes as our target, but not the identical objects like landmarks which rely on 3D structure or viewpoint. One of the other similar works is the work by Raguram et al. [19]. They aimed 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 from theirs in terms of selecting representative photos for each geographic region. [18] treated with generic concepts like our work. However, we select canonical images on generic concepts regarding several regions in the worldwide dimension, while [18] treated with general concepts within only given regions.

Regarding entropy analysis for relations between word concepts and locations, as long as we know, there exists no similar works so far. To analyze the distribution of visual features regarding the given concepts, Yanai et al.[24] proposed "region image entropy" to measure visualness of word concepts with images gathered from the Web. They claimed that word concepts with low entropy are more appropriate for automatic image annotation [2]. Koskela et al.[12] used entropy to analysis the large-scale multimedia ontology, LSCOM [16]. In this paper, we use the modified "region image entropy"[24] which is different from the original methods in image features and learning methods.

3. ENTROPY ANALYSIS

In this section, we propose a new method to analyze relations between location and concepts in terms of image features. We compute both image region entropy [23] and geo-location entropy for many concepts using geotagged images gathered from the Flickr.

3.1 Image Region Entropy

Originally, "Image Region Entropy" was proposed in [23], which is a measure of "visualness" of concepts, that is, what extent concepts have visual characteristics. In the original method to compute "image region entropy", they perform probabilistic region selection for regions that can be associated with concept "X" from images which are labeled as "X" or "non-X", and then they compute a measure of the entropy of the selected regions based on a Gaussian mixture model for regions. By introducing a probabilistic region selection method, they can separate foreground regions from background regions, and compute the entropy using only the foreground regions. Intuitively, if such an entropy is low, then images associated with the concept have typical appearances, and the image features of the concepts are relatively concentrated. Alternatively, if the entropy is larger, the image features of the concepts are distributed, and the concept has no typical images.

In this paper, we modify the original method by using the bag-of-feature representation (BoF) [5] and the probabilistic latent semantic analysis (PLSA) [8] instead of color and texture features and the Gaussian mixture model (GMM), since it is regarded that the BoF representation has more semantically discriminative power than other representations [5] and PLSA is more appropriate for the BoF vectors than GMM. In addition, while the original method employed the probabilistic generative methods to select foreground regions which are used for computation of the region entropy, we use mi-SVM [1], which is a discriminative method, to select positive regions by taking account of the multiple instance learning setting, since discriminative method is superior to generative method in general in case that much training data is available.

We used the following iterative procedure based on mi-SVM [1] to select foreground regions:

1. Prepare a positive image set gathered from Flickr and a random background image set, carry out region segmentation with JSEG [6], and construct the region-based BoF vector for each region.
2. Sample one third of positive images and negative background images. Train SVM with them.
3. Classify all the regions of positive images with the trained SVM.
4. Select one third of regions in the descending order of the output values of the SVM. The selected regions can be regarded as positive regions.
5. If the number of iteration is more than the pre-defined value r, finish the selection of positive images. In the experiment, we set r as 5.
6. Otherwise select one sixth of positive regions in the ascending order of the output values of the SVM as negative samples. Sample one sixth of negative background images, and add them to negative samples.
7. Train SVM, and jump back to (3).

In the next step, to estimate the entropy of the image features of selected regions with respect to a generic distribution of image futures. To represent a generic model, we use the probabilistic Latent Semantic Analysis (PLSA) [8] which is the probabilistic method to identify latent topics with the give number of topics. PLSA was originally proposed as a probabilistic model to extract latent topics from text documents represented by bag-of-words. In a similar consideration, since images can be regarded as "documents" and represented by bag-of-features, hence PLSA can be applied to images for discovering the object categories in each image.

We need to obtain generic base topics in advance by the PLSA for computing the entropy. To obtain the generic base, we used about ten thousand images randomly picked up from the images gathered from the Web.

The PLSA model is represented as the generative model of each word w in a document d:

$$P(w,d) = P(d)\sum_{z\in Z} P(w|z)P(z|d) \quad (1)$$

where $z \in Z = (z_1, .., z_k)$ is a latent topic variable, k is the number of topics, $d \in D = (d_1, ..., d_N)$ is an image region expressed by the bag-of-features vector, and $w \in W = (w_1, ..., w_M)$ is one element of the BoF vector, which corresponds to a "visual" word. The joint probability of the observed variables, w and d, is the marginalization over the k latent topics Z. The parameters are estimated by the EM algorithm. In the experiments, we set 300 to the number of base topics k. We carry out this estimation of P(w—z) in advance which is regarded as training process of the PLSA.

For each positive region i for the concept "X", we estimate $P(z|d_i^X)$ employing "fold-in heuristics" [8]. The entropy for the concept "X" $H_{img}(X)$ is given by

$$H_{img}(X) = -\sum_k P(z_k|X)\log_2 P(z_k|X) \quad (2)$$

$$, where \;\; P(z_k|X) = \frac{1}{|I_{selected}|}\sum_{i\in I_{selected}} P(z_k|d_i^X) \quad (3)$$

and $|I_{selected}|$ is the number of selected positive regions. We summarize the procedure described above in Figure 1.

3.1.1 *Image Representation*

In this subsection, we describe about the region-based bag-of-features (BoF) representation [5] we use as an image representation to estimate image region entropy.

The main idea of the bag-of-features representation [5] is representing images as collections of independent local patches, and vector-quantizing them as histogram vectors. Before constructing the bag-of-features vector, we apply region segmentation for all the images. To obtain the region-based BoF vector, we extract the BoF vector from each region. As a region segmentation method, we use JSEG [6] after adjusting the parameters so as to generate about eight regions per image on average.

The main steps of the method are as follows:

1. Carry out region segmentation with the JSEG algorithm.
2. Sample 3000 patches per image randomly in the same way as [17].
3. Generate feature vectors for the sampled patches by the SIFT descriptor [13].
4. Construct a codebook with k-means clustering over all the extracted feature vectors. A codebook is constructed for each concept independently. We set k as 300.

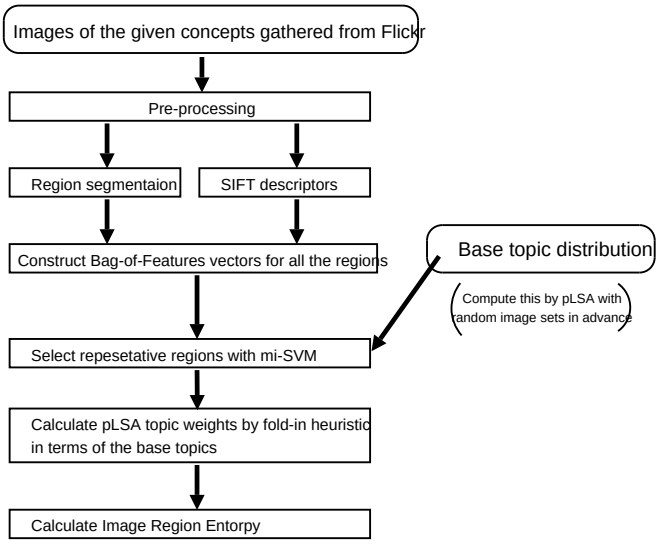

Figure 1: Flow diagram of the procedure to obtain the "image region entropy".

5. Assign all SIFT vectors to the nearest codeword of the codebook, and convert a set of SIFT vectors for each region into one k-bin histogram vector regarding assigned codewords. In addition, background images which are prepared as negative training samples in advance are also divided into regions and converted the sets of SIFT vectors extracted from regions into k-bin histograms based on the same codebook.

3.2 Geo-location Entropy

We can obtain location information of downloaded images which is represented by a set of an altitude and a longitude from Flickr with FlickrAPI. In this work, we calculate entropy regarding geo-location in addition to image region entropy. To estimate geo-location entropy, we build a histogram regarding location distribution on each concept by dividing latitude and longitude by every 10 degrees as shown in Figure 2.

Geo-location entropy $H_{geo}(X)$ is calculated by the following equation:

$$H_{geo}(X) = -\sum_i b_i \log_2 b_i \quad (4)$$

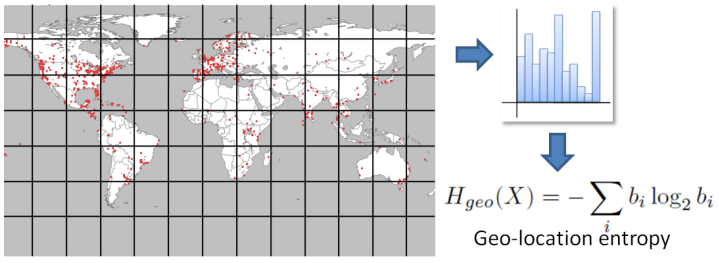

Figure 2: The overview of "Geo-location entropy".

3.3 Experiments

We made an experiment on image region entropy and geo-location entropy for 230 nouns shown in Table 1. As noun concepts, we selected abstract nouns, person names and location names as well as nouns related to objects and scenes in order to mix words which are likely to be related to location and words are not likely to be related to location. In the experiments, we gathered 500 images for each concept from Flickr using FlickrAPI.

Table 1: 230 noun concepts

africa	alexander	alligator	america	ant	asia	bach	banana
barbecue	battle	beach	beauty	beaver	bee	beer	beetle
board	boat	bob	book	box	bread	brother	bug
building	burger	bus	california	canada	candy	car	castle
cat	cedar	chair	chalk	cherry	chicken	china	circle
city	coffee	coke	color	computer	cookie	coral	crow
dandelion	daughter	desert	desk	dessert	deutschland	dish	doctor
dolphin	dragonfly	dream	eagle	edison	eel	egg	egypt
election	elevator	erica	europe	father	fern	field	fish
flea	flower	fly	fork	france	frog	fruit	game
gates	giraffe	goat	gorilla	grape	grass	grasshopper	gun
half	ham	hawk	height	hibiscus	hornet	house	ice cream
india	insect	italia	ivy	japan	jellyfish	jump	kangaroo
killer whale	lamp	lavender	lawn	leaf	lemon	level	library
light	lincoln	lion	lizard	locust	love	mangrove	manta
mantis	marriage	milk	mint	monkey	moon	mosquito	moss
moth	mother	mountain	mouse	mozart	museum	napoleon	new york
octopus	owl	oyster	palm	paris	park	parrot	party
pen	penguin	phone	pine	pizza	plant	pool	pope
potato	pride	rabbit	rice	rome	rose	salad	salmon
santa claus	school	sea	shakespeare	shark	ship	shrimp	sister
sky	snail	snake	son	sound	sport	square	starfish
steak	sun	sushi	sword	tea	teacher	temple	test
thomas	tiger	toad	tokyo	tool	town	tulip	tuna
turtle	usa	valley	village	whale	wine	worm	zoo
airplanes	backpack	bear	buddha	butterfly	cactus	cake	canoe
dice	dog	duck	eiffel tower	elephant	fireworks	goose	helicopter
horse	kayak	mars	mushroom	people	pyramid	rainbow	skyscraper
socks	spider	swan	tripod	watch	waterfall		

We show the top 20 and bottom 20 results in terms of region entropy and geo-location entropy in Table 2 and in Table 3, respectively.

Figure 3 shows the relations between image region entropy (x-axis) and geo-location entropy (y-axis) regarding 230 nouns. Table 4 represents the cross table between image region entropy and geo-location entropy regarding some concepts picked up from Figure 3. This table shows concepts which has relatively larger or smaller image region entropy and larger or small image geo-location entropy at the same time. Here, a concept with larger entropy means the concept is included in top 46 concepts in terms of entropy ranking, and a concept with smaller entropy means the concept is included in bottom 46 concepts in terms of entropy ranking.

3.4 Discussion

As a prominent tendency shown in Table 4, while geo-location entropy of location concepts such as "Deutschland", "Rome" and "Italia" and name concepts of historical persons such as "Mozart" and "Lincoln" were small, image region entropy of locations and person names were larger. "Sox" also belongs to this category, since the "Sox" image set gathered from Flickr includes many "Red Sox" photos which is a popular baseball team in US.

As shown in Figure 4, geo-location of location names and person names are strongly tied with the concepts themselves, while images related to them includes various appearance since they are relatively abstract concepts rather than physical concepts.

We found that for the concepts related to sky such as "sun" (Figure 5) and "rainbow" their image region entropy were smaller, while geo-location entropy were larger. This is because appearances related to such concepts tends to be very similar or almost the same everywhere over the world. So geo-location entropy became high, and image region entropy became low.

"Tulip" (Figure 6) was the only concepts which has low image entropy and low geo-location entropy. "Tulip" was mainly concentrated on the United States and Europe, especially , Holland, and most of the "tulip" photos included tulip flowers and tulip farms.

Both image and geo-location entropies on "Rice" (Figure 7) were large. Although "Rice" is a food concept which are common everywhere over the world, the way to cook is difference depending on the countries greatly. Moreover it is also the famous person's name at the same time. "Dolphin" also were distributed over the world, and its photos had diverse appearances as shown in Figure 8.

Table 2: Image region entropy $H_{img}(X)$ of top 20 and bottom 20 of 230 nouns

top 20		bottom 20	
concepts	$H(X)$	concepts	$H(X)$
sun	3.6497	horse	7.3057
rainbow	4.5538	pizza	7.3071
moon	4.6686	salad	7.3093
dragonfly	4.7550	africa	7.3101
sky	5.1049	japan	7.3387
mantis	5.1897	oyster	7.3435
egg	5.2288	flea	7.3590
airplanes	5.3851	tiger	7.3874
bee	5.4210	rice	7.3906
light	5.4524	rome	7.4013
fly	5.4916	usa	7.4020
coffee	5.6160	backpack	7.4086
bug	5.6407	italia	7.4111
mouse	5.6558	town	7.5177
butterfly	5.6785	santa-claus	7.5431
lemon	5.7096	house	7.5598
dream	5.7173	napoleon	7.5704
lamp	5.7347	school	7.6173
insect	5.7560	lincoln	7.7327
tulip	5.7700	mozart	7.8349

Table 3: Geo-location entropy $H_{geo}(X)$ of top 20 and bottom 20 of 230 nouns

top 20		bottom 20	
concepts	$H(X)$	concepts	$H(X)$
deutschland	0.2602	beetle	5.3225
rome	0.3843	grasshopper	5.3301
tokyo	0.6253	rice	5.3425
paris	0.6730	waterfall	5.3449
eiffel-tower	0.7461	monkey	5.3600
california	0.8776	boat	5.3755
new-york	1.0264	sun	5.4109
italia	1.3105	pool	5.4225
france	1.4833	banana	5.4674
egypt	1.8476	parrot	5.4767
japan	2.2973	sea	5.4936
mozart	2.6904	mother	5.5114
lincoln	2.7962	teacher	5.5417
europe	3.0379	lizard	5.5448
canada	3.2612	fruit	5.5779
castle	3.3948	hibiscus	5.5856
bach	3.4406	ant	5.6147
napoleon	3.4686	coral	5.6565
india	3.4874	fish	5.7831
shakespeare	3.5342	mosquito	5.9759

4. DISCOVERING CULTURAL DIFFERENCES

In this section, we propose a novel method to select representative photographs for regions in the worldwide dimensions, which helps detect cultural differences over the world regarding word con-

Figure 9: After collecting geotagged photo related to the given concept by the tag-based search, we remove noise images, cluster regions and select regional representative images.

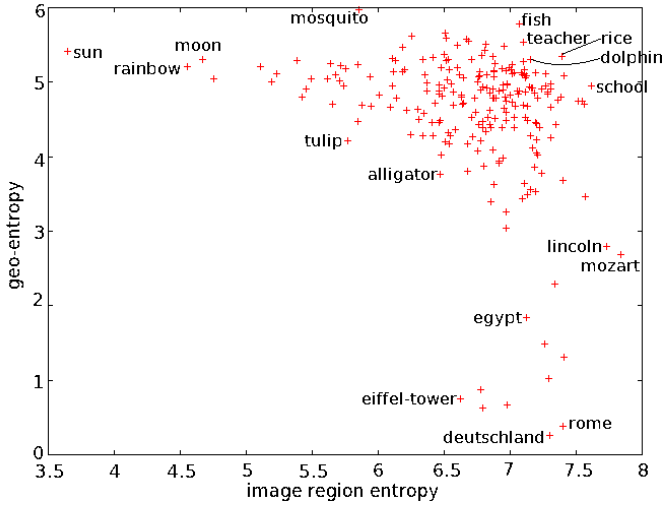

Figure 3: Relations of visual entropy and geo-entropy.

Table 4: Cross table between image region entropy and geo-location entropy

Geo-location entropy (GI)	Image region entropy (IE): Smaller IE	Image region entropy (IE): Larger IE
Smaller IG	tulip	france china new-york deutschland africa japan rome usa Italia napoleon lincoln morzart pope killer-whale chalk socks shakespeare thomas pride
Larger IG	sun rainbow sky moon airplane bug dream insect beach mosquito beetle beauty banana mangrove ant flower	dolphin rice

cepts with high geo-location entropy.

4.1 Approach

Our approach for selecting the representative images for representative local regions from geotagged images consists of three main stages as shown in Figure 9: (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-features (BoF) vectors [5]. 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". The method employed in this stage is based on the method proposed by Raguram et al. [19].

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) [8] 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, a UI is designed to support the browsing photos in context and understanding of the general object concepts.

4.2 Filtering Irrelevant Images

4.2.1 Visual Clustering

We adopt the bag-of-features representation [5] as the image representation the detail of which is described in the previous section. After building bag-of-features representation for all raw images, we perform clustering using k-means algorithm over the bag-of-features 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).

4.2.2 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:

$$sim(P_i, P_j) = \frac{V_i \cdot V_j}{\sqrt{\mid V_i \mid\mid V_j \mid}} \quad (5)$$

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

$$SIM(\mathbb{C}) = \frac{\sum_{P_i, P_j \in \mathbb{C}, i \neq j} sim(P_i, P_j)}{{}_nC_2} \quad (6)$$

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).

4.3 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.)

4.4 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 apply the Probabilistic Latent Semantic Analysis (PLSA) [8] model which we have already used in Section 2.

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 represent a joint probability model over images and visual words, defined as:

$$P(w, d) = P(d) \sum_{z \in Z} P(w|z) P(z|d) \quad (7)$$

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.

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.

Table 5: Evaluation results. This table describes the number of raw photos directly collected from Flickr (the numerical value in the () represents the precision), the number of photos selected by color-based method and our proposed method (two numerical values in the () represents the precision and the recall).

concepts	raw photos from Flickr	selection by color-based method	selection by proposed method
noodle	2080 (42)	769 (60, 54)	752 (90, 80)
flower	2225 (60)	703 (71, 37)	705 (85, 45)
castle	1848 (35)	780 (52, 61)	761 (70, 81)
car	1908 (43)	817 (56, 55)	800 (77, 75)
waterfall	1901 (39)	689 (63, 59)	672 (78, 70)
beach	1917 (38)	824 (51, 58)	813 (80, 90)
TOTAL/AVG.	11879 (43)	4582 (59, 54)	4503 (80, 74)

4.5 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.

4.5.1 Quantitative Evaluation

To evaluate our method for extracting the most relevant photos in the filtering stage, we use the precision, which is defined as $N_R/(N_R+N_{IR})$, and the recall, which is defined as $N_{R_{sel}}/N_{R_{col}}$, where N_R, N_{IR}, $N_{R_{sel}}$, and $N_{R_{col}}$ are the number of relevant photos, the number of irrelevant photos, the number of relevant photos in selected photos, and the number of relevant photos in raw collected photos, respectively.

For comparison, we still have applied simple color-historgam-based method to this selection task. First, we quantized the RGB color space into 64 (4 for each axis) bins, and made a color distribution histogram for each image. Based on the distance (histogram intersection) between images, we clustered all images into groups using k-means algorithm, and finally selected the largest clusters to form the most relevant image set. In addition, for more intuitively comparing, we kept the number of images selected by this method is almost equivalent to selecting by the proposed method.

In Table 5, we present the evaluation results of raw photos directly collected from Flickr, the precision and recall of photos selected by color-based method and our proposed method in terms of six keywords, "noodle", "flower", "castle", "car", "waterfall" and "beach". We obtain an average precision of 80% and an average recall of 74% by using our proposed method, which outperform the 43% precision of the rae images and the 59% precision and the 54% recall by the color-based method.

4.5.2 Examples of Regional Representative Photos

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 10 and Figure 11 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 10 demonstrates many "ramen" photos in Japan and Figure 11 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 12 and Figure 13 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 regions' results, Figure 14 for the region of Asia and Figure 15 for the region of South America. From the results, it is clear to find that waterfalls in South America seem to be more powerful, while waterfalls in the Asian area are somehow more beautiful. Such kinds of implications would be helpful in guiding travels around the world.

To see more results, please visit the following website: http://img.cs.uec.ac.jp/yanai/ASRP/ .

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed two methods to analyze relationship between word concepts and geographical locations by using a large amount of geotagged images on the photo sharing Web sites such as Flickr. Firstly, we proposed using both image region entropy and geo-location entropy to analyze relations between location and visual features, and in the experiment we found that concepts with low image entropy tends to have high geo-location entropy and vice versa. Secondly, we proposed a novel method to select representative photographs for regions in the worldwide dimensions, which helps detect cultural differences over the world regarding word concepts with high geo-location entropy.

Currently, the two works described in this paper are independent. For future work, we plan to integrate two works and 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 and specific scope. Finally, we will conduct methods to discriminate concepts which have larger cultural differences from concepts with low image entropy.

6. REFERENCES

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support Vector Machines for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems*, pages 577–584, 2003.

[2] K. Barnard, P. Duygulu, N. d. Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[3] L. Cao, J. Luo, H. Kautz, and T. Huang. Annotating collections of geotagged photos using hierarchical event and scene models. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2008.

[4] M. Cristani, A. Perina, U. Castellani, and V. Murino. Geo-located image analysis using latent representations. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2008.

[5] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Proc. of ECCV Workshop on Statistical Learning in Computer Vision*, pages 59–74, 2004.

[6] Y. Deng and B. S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):800–810, 2001.

[7] J. Hays and A. A. Efros. IM2GPS: Estimating geographic information from a single image. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2008.

[8] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 43:177–196, 2001.

[9] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 89–98, 2006.

[10] D. Joshi and J. Luo. Inferring generic activities and events from image content and bags of geo-tags. In *Proc. of ACM International Conference on Image and Video Retrieval*, 2008.

[11] L. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *Proc. of the International World Wide Web Conference*, pages 297–306, 2008.

[12] M. Koskela, A. F. Smeaton, and J. Laaksonen. Measuring concept similarities in multimedia ontologies: Analysis and evaluations. *IEEE Transaction on Multimedia*, 9(5):912–922, 2007.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[14] J. Luo, J. Yu, D. Joshi, and W. Hao. Event recognition: Viewing the world with a third eye. In *Proc. of ACM International Conference Multimedia*, 2008.

[15] M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *Proc. of ACM International Conference Multimedia*, pages 53–62, 2004.

[16] M. Naphade, J. R. Smith, J. Tesic, S. F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Transaction on Multimedia*, 13(3):86–91, 2006.

[17] E. Nowak, F. Jurie, W. Triggs, and M. Vision. Sampling strategies for bag-of-features image classification. In *Proc. of European Conference on Computer Vision*, pages IV:490–503, 2006.

[18] T. Quack, B. Leibe, and L. V. Gool. World-scale mining of objects and events from community photo collections. In *Proc. of ACM International Conference on Image and Video Retrieval*, pages 47–56, 2008.

[19] R. Raguram and S. Lazebnik. Computing iconic summaries of general visual concepts. In *Proc. of IEEE CVPR Workshop on Internet Vision*, 2008.

[20] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. In *Proc. of IEEE International Conference on Computer Vision*, 2007.

[21] K. Toyama, R. Logan, A. Roseway, and P. Anandan. Geographic location tags on digital images. In *Proc. of ACM International Conference Multimedia*, pages 156–166, 2003.

[22] K. Yaegashi and K. Yanai. Can geotags help image recognition ? In *Proc. of Pacific-Rim Symposium on Image and Video Technology*, 2009.

[23] K. Yanai. Image collector II : An over-one-thousand-image-gathering system. In *Proc. of the Twelfth International World Wide Web Conference*, 2003.

[24] K. Yanai and K. Barnard. Image region entropy: A measure of "visualness" of web images associated with one concept. In *Proc. of ACM International Conference Multimedia*, pages 420–423, 2005.

[25] J. Yu and J. Luo. Leveraging probabilistic season and location context models for scene understanding. In *Proc. of ACM International Conference on Image and Video Retrieval*, pages 169–178, 2008.

[26] J. Yuan, J. Luo, H. Kautz, and Y. Wu. Mining GPS traces and visual words for event classification. In *Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2008.

Figure 4: ”Mozart” images over the world.

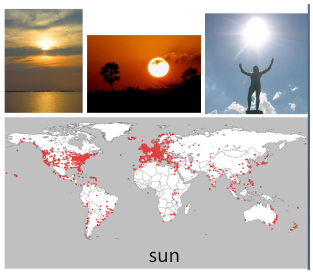

Figure 5: ”Sun” images over the world.

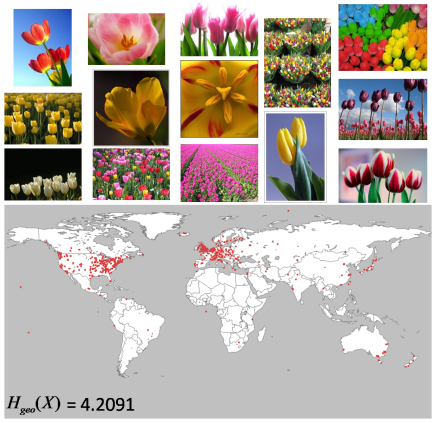

Figure 6: ”Tulip” images over the world.

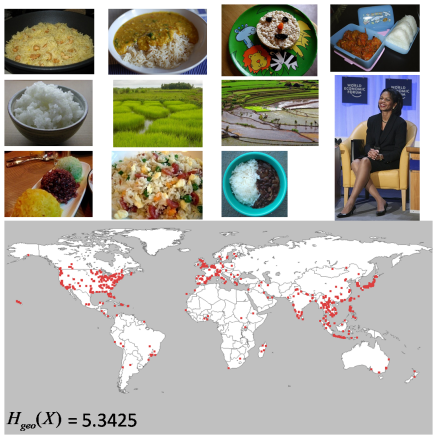

Figure 7: ”Rice” images over the world.

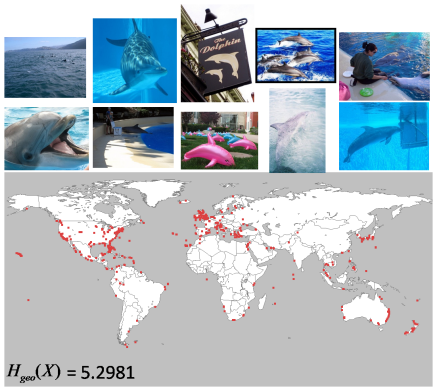

Figure 8: ”Dolphin” images over the world.

Figure 10: ”Noodle” in Japan. Chinese-style noodle “ramen” is popular.

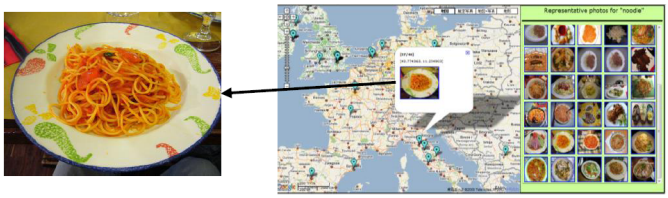

Figure 11: ”Noodle” in Europe. Most of photos are “Spaghetti”.

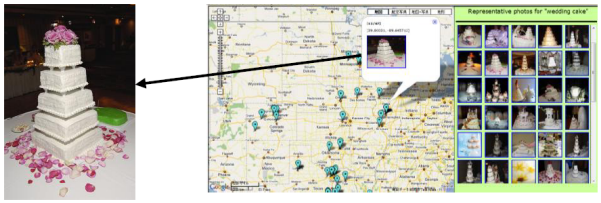

Figure 12: ”Wedding cake” in Mid US. Tall cakes are common. This is five-layered.

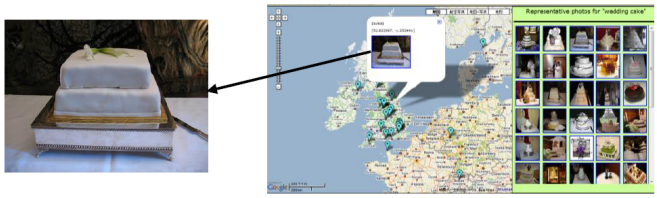

Figure 13: ”Wedding cake” in Europe. They are much shorter and simpler than US.

Figure 14: ”Waterfall” in Asia. It is somewhat beautiful.

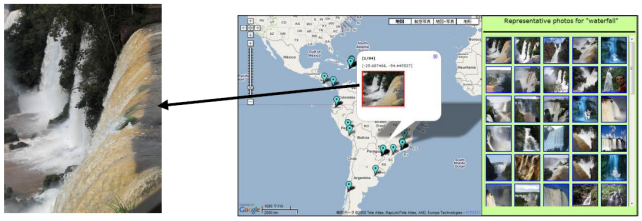

Figure 15: ”Waterfall” is South America. It is more powerful.