Can Geotags Help Image Recognition?

Keita Yaegashi and Keiji Yanai

Department of Computer Science,
The University of Electro-Communications
1–5–1 Chofugaoka, Chofu-shi, Tokyo, 182–8585 Japan
{yaegas-k,yanai}@mm.cs.uec.ac.jp

Abstract. In this paper, we propose to exploit geotags as additional information for visual recognition of consumer photos to improve its performance. Geotags, which represent places where the photos were taken, for photos can be obtained automatically by carrying a portable small GPS device with digital cameras. Geotags have potential to improve performance of visual image recognition, since recognition targets are unevenly distributed. For example, "beach" photos can be taken near the sea and "lion" photos can be taken only in a zoo except Africa.

To integrate geotag information into visual image recognition, we adopt two types of geographical information, raw values of latitude and longitude, and visual feature of aerial photos around the location the geotag represents. As classifiers, we use both a discriminative method and a generative method in the experiments.

The objective of this paper is to examine if geotags can help category-level image recognition. Note that we define an image recognition problem as deciding if an image is associated with a certain given concept such as "mountain" and "beach" in this paper. We propose a novel method to carry out geotagged image recognition in this paper. The experimental results demonstrate effectiveness of usage of geographical information for recognition of consumer photos.

1 Introduction

Due to the spread of consumer digital cameras and camera-equipped cell phones, we can easily take a large number of digital photos, while managing them is a troublesome job. To manage a large number of photos, word-tagging is one of popular methods, which enables us to search our personal photo storages with words. However, word-tagging by hand for a lots of photos is too boring and time-consuming task for many people. Therefore, automatic word-tagging is desirable.

In fact, in the research community of image recognition, visual recognition of generic consumer photos taken by people with usual digital cameras is one of hot topics. Recent progress on image representation [2,7], machine learning and computation power of computers have made visual recognition of consumer photos possible. Actually, 101 kinds of photo images can be classified automatically with the 87.8% classification rate by the state-of-the-art method [12]. However, since we have several thousands of kinds of targets to be recognized, visual image

T. Wada, F. Huang, and S. Lin (Eds.): PSIVT 2009, LNCS 5414, pp. 361–373, 2009.

recognition for consumer photos in which targets are not restricted is still far from practical use.

In this paper, we propose to exploit geotags as additional information for visual recognition of consumer photos to improve its performance. Geotags for photos can be obtained automatically by carrying a portable small GPS device with digital cameras. Geotags have potential to improve performance of visual image recognition, since recognition targets are unevenly distributed in the real world. For example, "beach" photos can be taken near the sea and "lion" photos can be taken only in a zoo except Africa. In this way, geotags can restrict concepts to be recognized for images, so that we expect geotags can help visual image recognition. In this paper, we examine if geotags can help visual recognition of consumer photos by experiments.

To utilize geotags in visual image recognition, we propose two methods: (1) combine values of latitude and longitude with visual feature extracted from a photo image. (2) combine visual feature extracted from aerial photo images with visual feature extracted from a photo image. The former method is relatively straightforward way, and it is expected to improve recognition performance for concepts associated with specific places such as "Disneyland" and "Mt. Fuji". On the other hand, in the latter method we utilize aerial photo images around the place where a photo was taken as information regarding that place. This will help more generic concepts such as "sea" and "mountain". Since "sea" and "mountain" are distributed all over the world, it is difficult to associate values of latitude and longitude with such generic concepts directly. Then, we regard aerial photo images around the place where the photo is taken as the information expressing the condition of the place, and utilize visual feature extracted from aerial images as yet another geographical information associated with geotags of photos. Especially, for geographical concepts such as "sea" and "mountain", using feature extracted from aerial photos is expected to be more effective than using values of latitude and longitude directly.

To collect geotagged images for experiments, we use Flickr. After Flickr launched an online geotagging interface in 2006, it became the largest geotagged photo database in the world. Flickr online geotagging system allows us to indicate the place where photos are taken by clicking the online map. In general, most of photos on the Web have no geospatial information, and photos in which GPS-based location information is embedded as the Exif data are very rare on the Web, People who like to add geotags their photos with GPS devices and upload them to the Web are very limited. Therefore, it was very difficult to collect large amount of geotagged images for research purpose so far. However, Flickr has changed this situation. They have a large number of images geotagged by Flickr's online geotagging system, and provide API to search Flickr photo databases for geotagged images. Everyone can access geotagged images on the Flickr very easily. From another point of view, in this paper, we propose to learn geotagged images from Flickr for visual recognition of consumer photos.

As related work related to geotagged photos, Kennedy et al. [5] proposed to select representative images by clustering based on visual feature regarding

a specific place. They used geottaged image collected from Flickr, and used geotags and word-tags to associate photos with a specific place. Snavely et al. [11] proposed to collect images associated with a specific place by sending the name of the place to Web image search engines and to estimate relative positions among the collected images by computer vision technique. They provided a new interface which enables us to see the given place from any direction of view.

Regarding recognition of aerial photo, it has been researched as "remote sensing" for more than thirty years [6]. To examine condition of the grounds effectively, aerial or satellite photos are analyzed with image recognition technique. Geographical features of the land such as the sea, rivers, mountains, city areas, islands and deserts in the photos are recognized. Therefore, in terms of recognition of aerial images, our work is related to remote sensing. The difference is that remote sensing aims at recognizing geographical features which appear in aerial photos directly, while the objective of our work is recognizing various kinds of concepts for consumer photos taken on the ground taking advantage of features which appear in aerial photos in addition to image features extracted from photos themselves. Since the concepts we intend to recognize are generic, that is, not restricted to geographical concepts such as rivers and roads, they do not always appear in aerial photos directly. For example, "flowers" do not appear in aerial photos directly in general. However, the places where "flower" photos are taken might have causal relationship to geographical features which appear directly in aerial photos. The places where "flower" photos are taken are unevenly distributed, and are usually not commercial areas or mountainous areas, but parks, farming areas or residential areas. We expect that this goes for many non-geographical concepts other than flowers. Then, in this paper, we take advantage of this indirect causal relation for geotagged image recognition.

The main objective of this paper is to examine if geotags can help image recognition by exploiting causal relation between aerial photos and concepts to be recognized. In this paper, we define an image recognition problem as judging if an image is associated with a certain given concept such as "mountain" and "beach". We propose a novel method to carry out geotagged image recognition in this paper, and we show the experimental results, which demonstrate effectiveness of usage of geographical information for recognition of consumer photos.

The rest of this paper is organized as follows: Section 2 describes basic idea of geotagged image recognition. Section 3 explains the procedure of geotagged image recognition for the experiments. Section 4 shows the experimental results and discusses them, and we conclude this paper in Section 5.

2 Geotagged Image Recognition

The objective of this paper is to examine if geotags can help image recognition. There are several types of image recognition. In this paper, we assume that image recognition means judging if an image is associated with a certain given concept such as "mountain" and "beach", which can be regarded as a photo detector for

a specific given concept. By combining many detectors, we can add many kinds of words as word-tags to images automatically.

As mentioned in the previous section, to integrate geotag information into visual image recognition, we adopt two types of geographical information, raw values of latitude and longitude, and visual feature of aerial photos around the geotagged location. To carry out experiments on the proposed geotagged image recognition, we need aerial photos corresponding to the geotags in addition to geotagged photos. We collect them from Flickr and an online aerial photo map site.

To perform geotagged image recognition, we need to extract feature vectors from images and geotags. As a representation of photo images, we adopt the bag-of-visual-words representation [2], which attracts much attention recently as a state-of-the-art method in the research community of image recognition. It has been proved that it has excellent ability to represent image concepts in the context of visual image recognition in spite of its simplicity. In the bag-of-visual-words method, an image is expressed by a high dimensional vector in the same way as a text document is expressed by a high dimensional bag-of-words vector. As a representation of geotags, we also adopt the bag-of-visual-words representation of aerial photos around the geotagged location in addition to raw values of latitude and longitude. After converting images and geotags into feature vectors, in this paper we adopt concatenation strategy, that is, combine them into one vector for each image.

After obtaining features vectors into which both visual and geographical information are mixed, we carry out two-class classification with two kinds of methods: a discriminative method and a generative method. As the discriminative method, we use Support Vector Machine (SVM), which is known as its excellent performance. As the generative methods, we use probabilistic latent topic mixture models [9]. In this paper, we use Probabilistic Latent Semantic Analysis (PLSA) [4] and Latent Dirichlet Allocation (LDA) [1] as latent topic models, while in [9] they used only PLSA.

3 Methods

In this section, we describe how to recognize images with visual features and geotags. First of all, we need to decide several concepts for the experiments. In this paper, we selected ten concepts for the experiments. Ideally, thousands kinds of concepts should be treated with as future work.

3.1 Data Collection

In this paper, we obtain geotagged images for the experiments from Flickr by searching for images which have Flickr tags corresponding to the given concept. Since the raw images fetched from Flickr include some noise images which are irrelevant to the given concepts, we select only relevant images by hand. In the experiments, relevant images are used as positive samples, while randomly-sampled images from all the geotagged images fetched from Flickr are used as

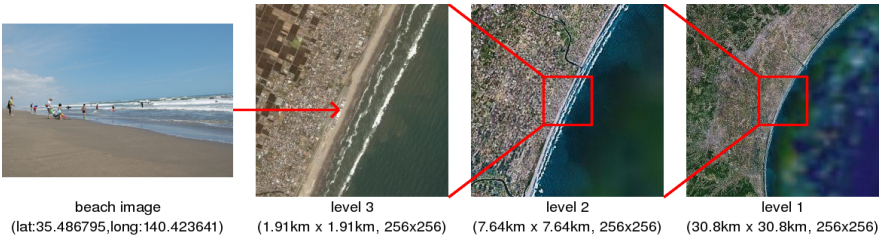

Fig. 1. Correspondence between a geotagged photo and aerial images

negative samples. We select 100 positive samples and 100 negative samples for each concept.

After obtaining geotagged images, we collect aerial photos around the points corresponding to the geotags of the collected geotagged image with several scales from an online aerial map site by screen-capturing so that the geotagged point is located at the center of an aerial photo. In the experiments, we collect 256×256 aerial photos in three different kinds of scales for one Flickr photo as shown in Figure 1. The larger-scale one (level 3) corresponds to an area of 1.91 kilometers square, the middle one (level 2) corresponds to a 7.64 kilometer-square area, and the smaller-scale one (level 1) corresponds to a 30.8 kilometer-square area. The level-1 and level-2 images are 16 times as large as the level-2 and level-3 images in terms of their size, respectively.

3.2 Extraction of Visual Features

To extract visual feature vectors from photos, we use the bag-of-visual-words method [2]. The main idea of the bag-of-visual-words is representing images as collections of independent local patches, and vector-quantizing them as histogram vectors. Note that the processing described below is carried out independently for each given concept.

The main steps to build a bag-of-visual-words vector are as follows:

1. Sample many patches from all the images. In the experiment, we sample patches on a regular grid with every 10 pixels.
2. Generate local feature vectors for the sampled patches by the SIFT descriptor [7] with four different scales.
3. Construct a codebook with k-means clustering over extracted feature vectors. A codebook is constructed for each concept independently. We set the size of the codebook k as 300 in the experiments.
4. Assign all feature vectors to the nearest codeword (visual word) of the codebook, and convert a set of feature vectors for each image into one k-bin histogram vector regarding assigned codewords.

SIFT Descriptors. Scale Invariant Feature Transform (SIFT) proposed by D. Lowe [7] provides a multi-scale representation of an image neighborhood. They

are Gaussian derivatives computed at 8 orientation planes over a 4×4 grid of spatial location, giving 128-dimension vector. The biggest advantage of SIFT descriptor is invariant to rotation. It has been shown that the SIFT descriptor is the best local patch descriptor for object recognition [8]. We compute SIFT vectors with the following four kinds of scales for regular grid points with every 10 pixels with the following four different scales: 4, 8, 12, and 16.

Generation of Codebook and Quantization. We obtain a collection of 128-dimension vectors for each image after the previous steps. Then, we apply vector quantization for them. Firstly, we compute a codebook by applying k-means clustering for all or randomly-sampled extracted SIFT vectors over both the positive training samples and negative training samples. In the experiment, we set the size of a codebook k as 300. Secondly, we assign all the SIFT vectors to the nearest codewords, which is sometimes called "visual words". This is the same as nearest neighbor search. Finally, we convert a set of the SIFT vectors for each image into one k-bin histogram of assigned codewords. Each histogram is represented by a k-dimension vector, so we have converted one image into one k-dimension feature vector based on the bag-of-visual-words representation.

3.3 Extraction of Geographical Features

As described before, we use visual features of aerial images around the point corresponding to the geotag, and raw values of latitude and longitude as geographical information.

Since a pair of latitude and longitude can be treated as a two-dimensional vector as it is, we need no conversion. On the other hand, since aerial photos are images, they should be converted into feature vectors. To do that, we adopt the bag-of-visual-words representation in the same way as extraction of visual features from photos. 256×256 aerial images the center of which correspond to the geotagged locations are converted into the bag-of-visual-words vectors. Note that the visual codebook for aerial images is constructed based of a set of SIFT vectors extracted from all the collected aerial images.

After converting both images and geotags into feature vectors, we combine them into one vector for each image by concatenating them.

3.4 Image Classification

After obtaining features vectors into which both visual and geographical information are mixed, we carry out two-class classification with two kinds of methods: a discriminative method and a generative method. As the discriminative method, we use Support Vector Machine (SVM). As the generative method, we use probabilistic latent topic mixture models [9].

Image Classification with SVM. As the first method, we use a Support Vector Machine (SVM) classifier with the RBF kernel. We train an SVM classifier with positive and negative training samples. Next, we classify test samples with the trained SVM one by one.

Image Classification with Latent Topic Mixture Models. As the generative method, we use probabilistic latent topic mixture models [9]. In this paper, we use Probabilistic Latent Semantic Analysis (PLSA) [4] and Latent Dirichlet Allocation (LDA) [1] as latent topic models, while in [9] they used only PLSA.

Recently, PLSA and LDA were applied to object recognition task as probabilistic generative models [10,3,9]. Since latent topic models such as PLSA and LDA were originally proposed for analyzing documents represented by bag-of-words, the mixture models of topics obtained by PLSA or LDA is more appropriate for classifying images represented by bag-of-visual-words than the Gaussian mixture model (GMM) which was commonly used as a probabilistic generative model before the bag-of-visual-words methods was proposed.

The main idea is that we apply probabilistic latent models to all the training samples to get latent topics, and decide "positive topics" and "negative topics" using the positive and negative training images.

The main steps are as follows:

1. Apply the latent topic method such as PLSA or LDA with the given number of topics to the bag-of-visual-words vectors of all the positive and negative training images, and get $P(z|d)$ where $z \in Z = (z_1, .., z_k)$ is the latent topic variable, and $d \in D = (d_1, ..., d_N)$ is an image.
2. Calculate the probability of being positive or negative over each topic, $P(pos|z)$ and $P(neg|z)$ using the pseudo-training images which are automatically selected in the collection stage.
3. Calculate $P(pos|d) = \sum_{z \in Z} P(pos|z)P(z|d)$, and evaluate relevancy of each image to the given keywords.

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

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

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 expressed by bag-of-visual-words, and $w \in W = (w_1, ..., w_M)$ is 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. For full explanation of the PLSA model refer to [4].

LDA: Latent Dirichlet Allocation (LDA) by Blei et al. [1] is also a probabilistic model to detect latent topics from text documents represented by bag-of-words. It was proposed as a method to resolve a drawback of PLSA that the number of parameters in the models grows linearly with the size of the data which leads to serious overfitting. LDA models each image as a mixture over topic, where each vector of mixture proportions is assumed to have been drawn from a Dirichlet distribution. The parameters are estimated by the variational EM algorithm. We also obtain $P(z|d)$ by applying LDA. For the detail refer to [1].

Next we estimate "positive topics" and "negative topics". A "positive topic" means that the latent topic is associated with images relevant to the given concept, and "negative topic" means that the latent topic is associated with irrelevant images. The probability of being positive and negative over a topic is calculated as follows:

$$p_0 = \frac{1}{|D_{pos}|} \sum_{d \in D_{pos}} P(d|z) \tag{2}$$

$$p_1 = \frac{1}{|D_{neg}|} \sum_{d \in D_{neg}} P(d|z) \tag{3}$$

$$P(pos|z) = p_0/(p_0 + p_1) \tag{4}$$

$$P(neg|z) = p_1/(p_0 + p_1), \tag{5}$$

where

$$P(d|z) = \frac{P(z|d)P(d)}{\sum_{d \in D} P(z|d)P(d)} \tag{6}$$

and, D_{pos} and D_{neg} are positive and negative samples, respectively.

Finally, we can calculate the probability of being positive over each image $P(pos|d)$ by marginalization over topics:

$$P(pos|d) = \sum_{z \in Z} P(pos|z)P(z|d) \tag{7}$$

We can rank all the candidate images based on this probability, $P(pos|d)$, and obtain the final result.

4 Experimental Results

4.1 Settings of the Experiments

We prepared the ten concepts shown in Table 1. The first two concepts, "mountain" and "beach" in Table 1 are geographical concepts which can be recognized in aerial images directly. The third and forth concepts, "road" and "train", are concepts related to social infrastructure which also is likely to be recognized in aerial photos. The fifth, "landscape", is relatively an abstract concept, which might corresponds to a broad area. The sixth, "shrine", is a concept related to architectures or religious places. The seventh concept, "flower", is an object concept, which is difficult to be recognized in aerial photos but existence of causal relation to geographical features is expected. The next one, "Chinese noodle", is a food concept. We do not know causal relation between it and aerial images. The last two concepts, "Disneyland" and "Tokyo Tower", represents specific places. For them, raw values of latitude and longitude are expected to be effective as an additional feature for image recognition. Note that we restricted the area of geotags attached to Flickr photos within Japan in the experiments.

Table 1. Ten concepts for the experiments

	concept	definition is this pape
1	mountain	a mountain landscape photo including mountain peaks
2	beach	a beach photo
3	road	a photo including roads clearly
4	train	a photo containing train vehicles
5	landscape	a landscape photo with no obstacles
6	shrine	architectures related to shrines
7	flower	a close-up photo for flowers or a photo mostly occupied with flowers
8	Chinese noodle	Chinese noodle with ready-to-east condition
9	Disneyland	photos taken inside the Disneyland
10	Tokyo Tower	The Tokyo Tower (in downtown Tokyo)

We collected geotagged images corresponding to the ten concepts from Flickr, and select 100 positive samples by hand. Table 1 shows the standard to select positive sample images by hand. Basically, we selected obvious positive images so that everyone agrees that selected images belong to the given concept. In addition, we prepare 100 randomly-sampled images as negative samples. After that, we collect three-different-scale aerial images of the places associated to all the positive and negative images.

In the experiments, we tried nine different combinations of visual feature of photos (V), raw values of latitude and longitude (R), and visual feature of aerial photos in three different level (L1, L2, L3). V can be regarded as a baseline. All the results were ranked by the output value of SVM or P(pos—d) computed by the probabilistic methods, and were evaluated by the average precision (AP) based on the following formula:

$$AP = \frac{1}{N} \sum_{i=1}^{N} Prec(i), \tag{8}$$

where $Prec(i)$ is the precision rate of the top i images which is defined as (number of positive images within the top i images)$/i$ and N is the number of test images for each fold.

We evaluate experimental results with five-fold cross validation, which means that all the data regarding one given concept are divided into five groups, four of them are used as training samples and the rest of them are used as to-be-recognized test samples. We perform classification and evaluate results repeatedly five times by exchanging test samples with the average precision. Finally we average the average precisions for five folds, and obtain the average precision for the given concept.

Table 2. Experimental results by SVM for nine combinations of visual feature of photos (V), raw values of latitude and longitude (R), and visual feature of aerial photos in three different level (L1, L2, L3). The red-colored bold value in each row represents the best result for each concept.

concept	V	V+L1	V+L2	V+L3	V+R	L1	L2	L3	R	**diff**
mountain	87.25	91.24	90.37	89.81	**91.84**	87.21	78.86	80.53	86.54	**+4.59**
beach	90.02	91.37	91.93	**93.32**	83.68	79.16	76.63	85.14	82.08	**+3.30**
road	71.27	72.11	73.08	**75.63**	69.28	62.71	65.85	59.09	69.62	**+4.36**
train	72.83	76.31	**77.38**	77.02	71.05	64.54	65.97	62.52	69.26	**+4.55**
landscape	77.16	79.16	**80.98**	**80.98**	77.75	64.52	65.30	67.35	66.04	**+3.82**
shrine	67.88	72.28	69.80	72.20	**72.89**	70.12	61.85	62.44	71.64	**+5.01**
flower	79.38	85.43	85.00	**86.63**	68.95	78.19	77.62	78.64	64.13	**+7.25**
Chinese noodle	86.49	87.31	**89.67**	87.71	86.65	68.01	73.13	68.29	82.28	**+3.18**
Disneyland	67.70	95.83	89.90	92.67	86.37	**98.56**	94.43	93.65	86.38	**+30.86**
Tokyo Tower	85.80	90.73	91.06	88.94	85.16	91.21	72.42	**91.70**	66.93	**+5.90**
AVG.	78.58	84.18	83.92	84.49	79.36	76.42	73.21	74.93	74.49	**+7.28**

4.2 Results

Table 2 shows the average precisions of the experimental results of visual image classification employing SVM on the given ten concepts regarding the following nine different combinations of features: V, V+L1, V+L2, V+L3, V+R, L1, L2, L3, and R. V represents the baseline with only visual features of images, while V+L1, V+L2, and V+L3 represent the combination of visual features of images and visual features of aerial images. V+R means the combination of visual features of images and the raw values, L1, L2, L3 and R represents only geographical features without visual features of the images. "Diff" in the table represents the difference on AP between the baseline and the best result incorporated with geospatial information.

Similarly, table 3 and Table 4 show the results in case of using the PLSA-based latent topic mixture and the LDA-based latent topic mixture, respectively. We set the number of topics as 20, which is selected from 10, 20 and 30 based on the preliminary experiments. Note that raw value of latitude and longitude cannot be incorporated with feature vectors in case of using probabilistic methods with PLSA or LDA, since LDA and PLSA assume that input vectors are represented by the bag-of-words representation. Therefore, results on V+R and R were omitted in Table 3 and Table 4.

4.3 Discussions

In case of PLSA, the average of APs over ten concepts are were degraded compared to the results by SVM and LDA. This is likely to come from the overfitting problem, which may also cause irregularly-biased results from concept to concept. On average, SVM outperformed PLSA and LDA for all kinds of the combinations of features except the baseline (V). Therefore, in this subsection, we discuss about the SVM results mainly.

Table 3. Experimental results by the PLSA mixture model

concept	V	V+L1	V+L2	V+L3	L1	L2	L3	**diff**
mountain	85.65	86.40	85.27	**87.50**	81.18	79.22	63.61	**+1.85**
beach	89.58	89.03	**90.03**	88.49	69.17	66.40	72.58	**+0.45**
road	62.22	**78.86**	67.56	63.13	61.30	61.91	48.84	**+16.64**
train	**71.07**	67.07	64.22	66.39	53.46	64.23	52.97	**+0.00**
landscape	**77.90**	72.43	73.76	76.82	48.57	60.85	59.60	**+0.00**
shrine	62.02	**77.13**	60.77	67.96	65.83	53.45	56.12	**+15.11**
flower	77.01	**86.79**	81.85	85.70	73.69	72.35	77.11	**+9.78**
Chinese noodle	75.76	74.02	73.60	**75.84**	50.70	55.01	62.23	**+0.08**
Disneyland	62.33	83.28	**90.81**	80.72	64.56	83.14	83.05	**+28.48**
Tokyo Tower	83.25	88.37	**91.63**	86.74	69.73	67.10	71.53	**+8.38**
AVG.	74.68	80.34	77.95	77.93	63.82	66.36	64.76	**+8.08**

Table 4. Experimental results by the LDA mixture model

concept	V	V+L1	V+L2	V+L3	L1	L2	L3	**diff**
mountain	86.64	**89.52**	88.24	88.72	84.60	79.98	83.10	**+2.88**
beach	89.93	90.48	91.13	**92.12**	79.01	76.53	76.87	**+2.19**
road	**71.70**	69.60	70.12	68.05	58.59	61.55	59.27	**+0.00**
train	74.87	**76.59**	74.34	74.30	66.69	64.64	58.50	**+1.72**
landscape	83.51	83.55	83.29	**86.13**	62.58	61.17	67.72	**+2.62**
shrine	66.13	**70.29**	68.93	68.76	68.76	62.81	56.45	**+4.16**
flower	80.08	**88.50**	85.69	87.60	76.87	76.24	78.79	**+8.42**
Chinese noodle	85.85	**89.25**	85.83	82.89	72.05	63.91	65.38	**+3.40**
Disneyland	64.80	86.11	92.02	92.60	**98.82**	94.96	97.05	**+34.02**
Tokyo Tower	85.28	91.04	91.43	87.59	96.00	70.91	**96.28**	**+11.01**
AVG.	78.88	83.49	83.10	82.88	76.40	71.27	73.94	**+7.04**

In case of SVM as a classifier, for all the ten concepts, the best results among eight combinations including geospatial features were superior to the baseline. Basically this is because the places where positive sample photos were taken are unevenly distributed, while the places where negative sample photos were taken are randomly distributed. Especially, all the SVM results by the combination of visual features of images and aerial photos (V+L1/L2/L3) outperformed the baseline results. This shows that incorporating visual features extracted from aerial photos with visual features extracted from images are effective and promising for image recognition.

The APs by SVM were improved by about 3% to 5% except for "Disneyland". For "Disneyland" which is a specific place name, geotags boosted the results greatly, and with only aerial photos and no visual information of the images the 98.56% average precision was obtained. From this result, to discriminate images associated with specific place names from randomly-sampled negative images, only geospatial information is enough. As we expected, for "Tokyo Tower", the

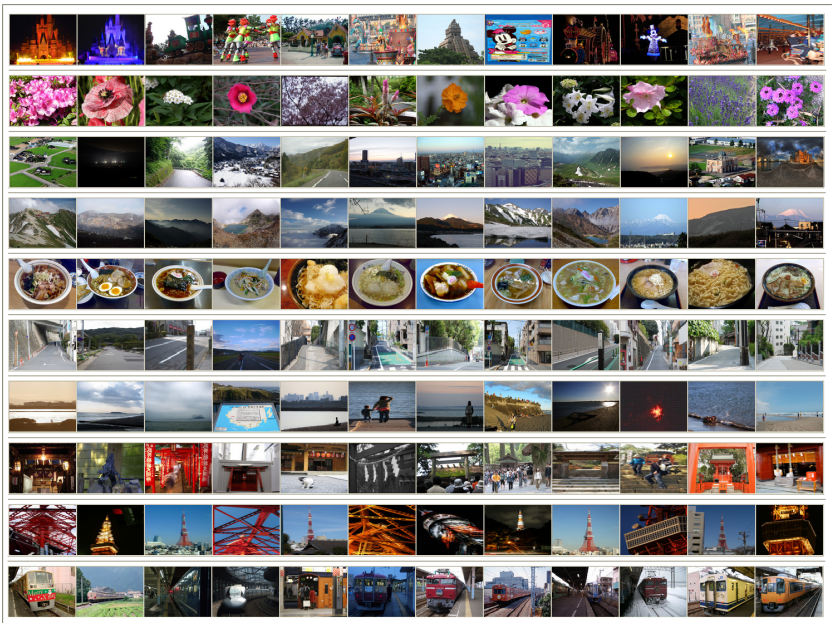

Fig. 2. Positive sample photos of ten categories: Disneyland, flower, landscape, mountain, Chinese noodle, road, beach, shrine, Tokyo Tower, and train

similar tendency was observed. However, the improvement is not as large as "Disneyland", since "Tokyo Tower" which is a 333 meter-high architecture can be seen from the relatively broad area of downtown Tokyo and the geotagged places are not as well-concentrated as "Disneyland".

Among concepts other than two specific location concepts, the result on "flower" were improved most. For "flower", while the result by visual features and raw coordinate values (V+R) was inferior to the baseline result (V), the results by the combinations of visual features and aerial photo features (V+L1/L2/L3) was much superior to the baseline (V). This shows "indirect causal relation" between "flower" concept and visual features extracted from aerial photos helped recognition of "flower" images.

From these results, we can conclude that geographical information has ability to help visual image recognition by using visual features of aerial images as additional features, although further experiments which should be more extensive are needed to examine effectiveness of this novel idea in detail.

5 Conclusions

In this paper, we proposed a novel method for "geotagged image recognition", which exploits aerial photos corresponding to the geotagged point as additional features for image classification. We made experiments so as to examine if geotags can help image recognition. The experimental results demonstrated effectiveness

of usage of geographical information for recognition of consumer photos. We believe this is the first attempt to utilize aerial photos where a photo was taken as additional features for image recognition.

In this paper, although we showed novel results that geotags helped performance of visual image recognition, the number of concepts examined in the experiments were limited. For future work, we plan to make more comprehensive experiments with several thousands of concepts and we also study more sophisticated method to integrate visual features of photos, visual features of aerial photos and raw values of latitude and longitude. In addition, it also should be investigated how to use aerial photos regarding levels and a range. Although we made the experiments on nine combinations of features in this paper, appropriate combinations for each concept should be selected automatically. Using several levels of aerial images at the same times will be possible. The final objective of this research project is to identify concepts for which geographical information helps image recognition effectively by examining several thousands of concepts.

References

1. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Proc. of ECCV Workshop on Statistical Learning in Computer Vision, pp. 59–74 (2004)
3. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proc. of IEEE Computer Vision and Pattern Recognition, pp. 524–531 (2005)
4. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 43, 177–196 (2001)
5. Kennedy, L., Naaman, M.: Generating diverse and representative image search results for landmarks. In: Proc. of the International World Wide Web Conference, pp. 297–306 (2008)
6. Lillesand, T.M., Kiefer, R.W., Chipman, J.W.: Remote sensing and image interpretation. John Wiley, Chichester (2004)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
8. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10), 1615–1630 (2005)
9. Monay, F., Gatica-Perez, D.: Modeling semantic aspects for cross-media image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(10), 1802–1817 (2007)
10. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their localization in images. In: Proc. of IEEE International Conference on Computer Vision, pp. 370–377 (2005)
11. Snavely, N., Seitz, S., Szeliski, R.: Photo tourism: exploring photo collections in 3d. ACM Transactions on Graphics (TOG) 25(3), 835–846 (2006)
12. Varma, M., Ray, D.: Learning the discriminative power-invariance trade-off. In: Proc. of IEEE International Conference on Computer Vision, pp. 1150–1157 (2007)