Geotagged Image Recognition by Combining Three Different Kinds of Geolocation Features

Keita Yaegashi and Keiji Yanai

Department of Computer Science,
The University of Electro-Communications
51–5–1 Chofugaoka, Chofu-shi, Tokyo, 182–8585 Japan
yaegas-k@mm.cs.uec.ac.jp, yanai@cs.uec.ac.jp

Abstract. Scenes and objects represented in photos have causal relationship to the places where they are taken. In this paper, we propose using geo-information such as aerial photos and location-related texts as features for geotagged image recognition and fusing them with Multiple Kernel Learning (MKL). By the experiments, we have verified the possibility for reflecting location contexts in image recognition by evaluating not only recognition rates, but feature fusion weights estimated by MKL. As a result, the mean average precision (MAP) for 28 categories increased up to 80.87% by the proposed method, compared with 77.71% by the baseline. Especially, for the categories related to location-dependent concepts, MAP was improved by 6.57 points.

1 Introduction

In these days, due to rapid spread of camera-equipped cellar phones and digital cameras with GPS, the number of geotagged photos is increasing explosively. Geotags enable people to see photos on maps and to get to know the places where the photos are taken. Especially, photo sharing Web sites such as Flickr and Picasa gather a large number of geotagged photos. Here, "geotag" means a two-dimensional vector consisting of values of latitude and longitude which represent where a photo is taken.

In this paper, we exploit geotags as additional information for visual recognition of consumer photos to improve its performance. Geotags have potential to improve performance of visual image recognition, since the distributions of some objects are not even but concentrating in some specific places. For example, "beach" photos can be taken only around borders between the sea and lands, and "lion" photos can be taken only in zoos except Africa. In this way, geotags can restrict concepts to be recognized for images, so that we expect geotags can help visual image recognition.

The most naive way to use geo-information for image recognition is adding 2-dim location vectors consisting of values of latitude and longitude to image feature vectors extracted from a given photo. However, this method is not effective except for some concepts associated with specific places "Disneyland" and "Tokyo tower" , because objects represented by general nouns such as "river" and "lion" can be seen at many places. To make location vectors to improve

image recognition performance, we have to prepare all the locations where they can be seen in the database in advance. This requirement is too unrealistic to handle various kinds of concepts to be recognized.

To make effective use of geo-information for image recognition, several methods to convert 2-dim location vectors into more effective features to be integrated with photo image features have been proposed so far [1–4]. Luo et al. [1] and Yaegas et al. [3, 4] proposed using aerial photos which are taken at the places corresponding to the geotags, and extracting visual features from them. On the other hand, Joshi et al. [2] proposed converting 2-dim geotag vectors into geo-information texts using reverse geo-coding Web services such as `geonemes.com`, and forming bag-of-words vectors. However, there exists no work to make use of both of aerial visual features and geo-information text features as additional features so far. Then, in this paper, we propose integrating both features extracted from aerial photos and geo-information texts with visual features extracted from a given image.

To fuse features of aerial photos and ones of geo-location texts with visual features of a given image, we use Multiple Kernel Learning (MKL). The MKL is also used to fuse aerial features and image features in [4]. Compared to [4], the main contributions of this paper are as follows: (1) We fuse various kinds of features derived from geotagged photos containing color visual features, geo-location texts, 2-dim geotag vectors and time-stamp features expressing the date and time when a photo is taken in addition to aerial features and image features extracted by grayscale-SIFT-based bag-of-features using MKL. (2) The experiments are more comprehensive regarding the number of concepts and combinations of features.

By the experiments, we have confirmed the effectiveness of using both aerial features and geo-text features for image recognition by evaluating not only recognition rates, but feature fusion weights estimated by MKL. As a result, the mean average precision (MAP) for 28 categories increased up to 80.87% by the proposed method, compared with 77.71% by the baseline. Especially, for the categories related to landmarks, MAP was improved by 6.57.

The rest of this paper is organized as follows: Section 2 describes about related work, and Section 3 and 4 explains the overview and the procedure of geotagged image recognition by fusing various kinds of features including ones extracted from photos, aerial photos and locate-related texts with Multiple Kernel Learning (MKL). Section 5 shows the experimental results and discusses them, and we conclude this paper in Section 6.

2 Related Work

To utilize geotags in visual image recognition, there are four papers proposed as mentioned in the previous section [1–4].

To utilize geotags in visual image recognition, three kinds of methods have been proposed so far: (1) combining values of latitude and longitude with visual features of a photo image [3], (2) combining visual feature extracted from aerial

photo images with visual feature extracted from a photo image [1, 3, 4], and (3) combining textual features extracted from geo-information texts obtained from reverse geo-coding services [2].

The method (1) is relatively straightforward way, and it improved recognition performance only for concepts associated with specific places such as "Disneyland" and "Tokyo tower" in the experiments of [3].

On the other hand, in the method (2), we utilize aerial photo images around the place where a photo was taken as additional information on the place. 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 geographic contextual information associated with geotags of photos. Especially, for geographical concepts such as "sea" and "mountain", using feature extracted from aerial photos was much more effective than using raw values of latitude and longitude directly [3].

Since in the method (2) of [3] they combined two kinds of feature vectors by simply concatenating them into one long vector, the extent of contributions of aerial photos for geotagged photo recognition was unclear. Since Luo et al. [1] focused 12 event concepts such as "baseball", "at beach" and "in park" which can be directly recognized from aerial photos, in their experiments the results by using only aerial photos were much better than the results by using only visual features of photos in addition to revealing that the results by both features combined by late SVM fusion [5] outperformed both of the results. However, this results cannot be generalized to more generic concepts such as "flower" and "cat", since most of generic concepts are unrecognizable directly from the sky.

In [4], they proposed introducing Multiple Kernel Learning (MKL) to evaluate contribution of both features for recognition by estimating the weights of image features of photos and aerial images. MKL is an extension of Support Vector Machine (SVM), and makes it possible to estimate optimal weights to integrate different features with the weighted sum of kernels. In the experiments, they evaluated the weights of both features using MKL for eighteen concepts. The contribution weights between aerial features and visual photo features in terms of image categorization are expected to vary depending on target concepts. For the concepts which can be directly recognized from aerial photos such as "beach" and "park", the contribution rates of aerial photos were more than 50%, while they are less than 30% for the unrecognizable concepts from aerial photos such as "noodle", "vending machine" and "cat".

As work to incorporate textual geo-information, Joshi et al. [2] proposed using textual features as additional features of image recognition. The textual features are obtained by reverse geo-coding which are provided by `geonames.com`. They showed the effectiveness of using textual geo-informations for object recognition.

However, the following questions are not explored: Which is more effective for geotagged image recognition, aerial images or geo-textual information? How about combining both of them with visual image features? Then, in this paper,

we propose combining various kinds of geo-information including 2-dim location vectors, aerial image features and geo-textual features by MKL, and reveal which features are effective for which concepts by analyzing the experimental results.

As another kind of image recognition with geotags, place recognition has been proposed so far [6, 7]. In these work, geotags are used in only the training step. In the recognition step, the system recognizes not categories of images but places where the photo are taken. That is, geotags themselves are targets to be recognized. IM2GPS [6] is a pioneer work of place recognition. Kalogerakis et al. improved the performance of place recognition by using statistics of travel trajectories [7].

3 Overview

The objective of this paper is to fuse various kinds of features derived from geotagged photos including visual photo features, aerial features, geo-text features, 2-dim location vectors and date/time feature vectors by using Multiple Kernel Learning (MKL) and to evaluate the recognition performance and the contribution weights of various features for image recognition regarding various kinds of concepts.

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 representation of photo images, we adopt the bag-of-features (BoF) representation [8] and HSV color histogram. BoF has been proved that it has excellent ability to represent image concepts in the context of visual image recognition in spite of its simplicity. Regarding HSV color histogram, the effectiveness as additional features of BoF are also shown in [1].

As representation of aerial photos, we also adopt the bag-of-features representation of aerial photos around the geotag location. To adapt various kinds of concepts, we prepare four levels of aerial photos in terms of scale as shown in Figure 1, and extract a BoF vector from each of them.

Regarding geo-information texts, we use Yahoo! Japan Local Search API to convert geotags into reverse geo-coded texts, and build 2000-dim Bag-of-Words (BoW) vectors by counting frequency of the top 2000 highly-frequent words.

After obtaining feature vectors, we carry out two-class classification by fusing feature vectors with MKL. In the training step of MKL, we obtain optimal weights to fuse both features.

4 Methods

In this section, we describe how to extract images with visual features and various kinds of features derived from geotags and how to use them for image recognition.

Fig. 1: Correspondences between a geotagged photo and aerial images.

4.1 Data Collection

In this paper, we obtain geotagged images for the experiments from Flickr by searching for images which have word tags corresponding to the given concept. Since sets of raw images fetched from Flickr always contain 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 negative samples. We select 200 positive samples and 200 negative samples for each concept. Note that in the experiments we collected only photos taken inside Japan due to availability of high-resolution aerial photos.

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. In the experiments, we collect 256×256 aerial photos in four different kinds of scales for each Flickr photo as shown in Figure 1. The largest-scale one (level 4) corresponds to an area of 497 meters square, the next 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 smallest-scale one (level 1) corresponds to a 30.8 kilometer-square area.

In addition, to obtain textual features, we gather geo-information texts using reverse geo-coding services via Yahoo! Japan Local Search API, which transform a 2-dim geo-location vector into landmark names around the place indicated by the location vector within 500 meters. For example, we can obtain names of elementary schools, hospitals, hotels, buildings, parks and temples.

4.2 BoF Features

We extract bag-of-features (BoF) [8] vectors from both photos and aerial images.

The main idea of the bag-of-features is representing images as collections of independent local patches, and vector-quantizing them as histogram vectors. The main steps to build a bag-of-features 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 [9] with four different scales: 4, 8, 12, and 16.
3. Construct a codebook with k-means clustering over extracted feature vectors. We construct a codebook for photo images for each given concept independently, while we construct a codebook for aerial images which is common

among all the aerial images for any concepts. We set the size of the codebook k as 1000 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.

4.3 Color Histogram

SIFT-based BoF does not include color information at all. Then, to use color information for recognition, we use a HSV color histogram in addition to BoF as visual features of photos. A color histogram is a very common image representation. We divide an image into 5×5 blocks, and extract a 64-bin HSV color histogram from each block with dividing the HSV color space into $4 \times 4 \times 4$ bins. Totally, we extract a 1600-dim color feature vector which is L1-normalized from each image.

4.4 Raw location vectors

As one of features to be integrated by MKL, we prepare raw location vectors consisting of the values of latitude and longitude. Since a geotag represents a pair of latitude and longitude, it can be treated as a location vector as it is without any conversion. As a coordinate system, we use the WGS84 system which is the most common.

4.5 Aerial photo features

As described before, we use

the bag-of-features (BoF) representation as features extracted from aerial photos.

The way to extract BoF is the same as visual features from photos. We convert each of four different scales of 256×256 aerial images (Figure 1) the center of which correspond to the geotagged location into a BoF vector. Note that the visual codebook for aerial images is constructed based of a set of SIFT vectors extracted from all the collected aerial images.

4.6 Geo-information textual features

To obtain textual features, we gather geo-information texts using reverse geocoding services via Yahoo! Japan Local Search API.

For all the gathered geotagged images, we get geo-location texts, and select the top 2000 highly-frequent words as the codewords for bag-of-words (BoW) representation. We count the frequency of each word regarding the selected 2000 words, and generate a BoW histogram for each image.

4.7 Date features

In addition to geo-location features, we prepare date/time features based on time stamp information embedded in image files. To represent date and time, we prepare two histograms on month and hour. For a month histogram, we divide 12 months into 12 bins, and for a hour histogram, we divide 24 hours into 24 bins. For soft weighting, we vote 0.5 on the corresponding bin and 0.25 on the neighboring bins. Finally we concatenate both a month histogram and a hour histogram into one 36-dim vector.

4.8 Multiple Kernel Learning

In this paper, we carry out two-class classification by fusing visual features of photo images and aerial images with Multiple Kernel Learning (MKL). MKL is an extension of a Support Vector Machine (SVM). MKL handles a combined kernel which is a weighted liner combination of several single kernels, while a standard SVM treats with only a single kernel. MKL can estimates optimal weights for a linear combination of kernels as well as SVM parameters simultaneously in the train step. The training method of a SVM employing MKL is sometimes called as MKL-SVM.

Recently, MKL-SVM is applied into image recognition to integrate different kinds of features such color, texture and BoF [10, 11]. However, the recent work employing MKL-SVM focuses on fusion of different kinds of features extracted from the same image. This is different from our work that MKL is used for integrating features extracted from the different sources, which are photos and aerial images.

With MKL, we can train a SVM with a adaptively-weighted combined kernel which fuses different kinds of image features. The combined kernel is as follows:

$$K_{comb}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{K} \beta_j K_j(\mathbf{x}, \mathbf{y}) \quad \text{with} \quad \sum_{j=1}^{K} \beta_j = 1,$$

where β_j is weights to combine sub-kernels $K_j(\mathbf{x}, \mathbf{y})$. As a kernel function, we used a χ^2 RBF kernel, which was commonly used in object recognition tasks, for all the histogram-based feature vectors except raw location vectors. It is represented by the following equation:

$$K\left(\boldsymbol{x},\ \boldsymbol{x}'\right) = \exp\left(-\gamma \sum_{i=1}^{D} |x_i - x_i'|^2 / |x_i + x_i'|\right) \tag{1}$$

where γ is a kernel parameter. Zhang et al. [12] reported that the best results were obtained in case that they set the reciprocal of the average of χ^2 distance between all the training data to the parameter γ of the χ^2 RBF kernel. We followed this method to set γ. For 2-dim location vectors, we use the following normal RBF kernel, because it is not a histogram-based feature:

$$K\left(\boldsymbol{x},\ \boldsymbol{x}'\right) = \exp\left(-\gamma \sum_{i=1}^{D} |x_i - x_i'|^2\right) \tag{2}$$

For this kernel, we also set the reciprocal of the average of L2 distance between all the training data to the parameter γ.

Sonnenburg et al. [13] proposed an efficient algorithm of MKL to estimate optimal weights and SVM parameters simultaneously by iterating training steps of a standard SVM. This implementation is available as the SHOGUN machine learning toolbox at the Web site of the first author of [13]. In the experiment, we use the MKL library included in the SHOGUN toolbox as an implementation of MKL.

5 Experiments

5.1 28-category dataset

For experiments, we prepared twenty-eight concepts, a part of which are shown in Figure 2. To select the twenty-eight concepts, we first define eight rough types of concepts and then select several concepts for each type as follows:

Location-specific concept (LS) [2 concepts]
Disneyland, Tokyo tower
The locations related to these concepts are specific to them. Since all the aerial photos related to these concepts are similar to each other, using aerial photos are expected to improve recognition performance much. Moreover, 2-dim raw location vectors are also expected to work well.

Landmark concept (LM) [5 concepts]
bridge, shrine, building, castle, railroad
Since there are possibility of recognizing these concepts on aerial photos directly, improvement of performance is expected.

Geographical concept (GE) [3 concepts]
lake, river, beach
These concepts are also expected to be recognizable from the sky directly.

Space concept (SP) [3 concepts]
park, garden, landscape
These concepts represents space which consists of various elements. It it difficult to expect if aerial features work or not.

Outdoor artifact concept (OA) [5 concepts]
statue, car, bicycle, graffiti, vending machine
These concepts are almost impossible to recognize on aerial images, since some are very small and the others are movable.

Time-dependent concept (TD) [5 concepts]
sunset, cherry blossom, red leaves, festival, costume-play festival
The first three concepts of five depend on time or seasons rather than locations, while "Festival" and "costume-play festival" depend on both time and locations.

Creature concept (CR) [3 concepts]
cat, bird, flower
These concepts are very difficult to recognize on aerial images.

Food concept (FD) [2 concepts]

sushi, ramen noodle

These concepts are impossible to recognize on aerial images. We cannot expect improvement by additional features derived from geotags.

Location-specific, landmark and geographical concepts are expected to have correlation to aerial images, while creature, food and time-dependent are to have no or less correlation to aerial images.

We gathered photo images associated to these concepts from Flickr, and selected 200 positive sample images for each concepts. As negative sample images, we selected 200 images randomly from 100,000 images gathered from Flickr.

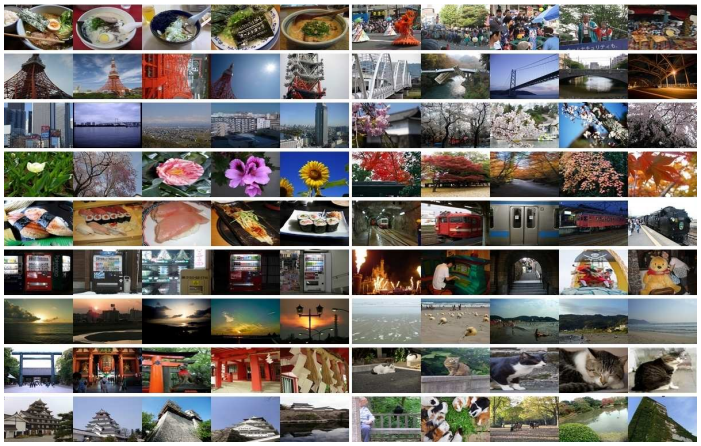

Fig. 2: Eighteen concepts of twenty-eight concepts for the experiments.

5.2 Combinations of features

Totally, in this paper, we use nine features containing BoF ans HSV color histogram of photos, BoF of aerial images in four scales, geo-text features, raw location vectors and time features. In the experiments, we prepare eleven kinds of their combinations shown in Table1, and provide each of them to Multiple Kernel Learning for training of it. Note that we regard four levels of aerial features as one feature for evaluation, although we provide four levels of aerial features to MKL independently.

5.3 Evaluation

In the experiments, we carried out two-class image classification and estimate weights to integrate BoF and color features of photos and various features derived from geotags using MKL for twenty-eight concepts.

Table 1: Eleven kinds of combinations of features

combinations	BoF	HSV	Aerial(A)	Geo-location(G)	Texts(T)	Date(D)
BoF (BoF)	O	-	-	-	-	-
HSV (HSV)	-	O	-	-	-	-
Vis (Vis)	O	O	-	-	-	-
Vis+G (VG)	O	O	-	O	-	-
Vis+T (VT)	O	O	-	-	O	-
BoF+A (BA)	O	-	O	-	-	-
Vis+A (VA)	O	O	O	-	-	-
Vis+A+T (VAT)	O	O	O	-	O	-
Vis+G+T+D (VGTD)	O	O	-	O	O	O
Vis+A+T+D (VATD)	O	O	O	-	O	O
All (All)	O	O	O	O	O	O

We evaluated experimental results with five-fold cross validation using the average precision (AP) which is computed by the following formula:

$$AP = \frac{1}{N} \sum_{i=1}^{N} Prec(i), \tag{3}$$

where $Prec(i)$ is the precision rate of the i positive images from the top, and N is the number of positive test images for each fold.

5.4 Experimental results

In this subsection, we describe the result of geotagged image recognition by MKL-SVM in terms of eleven kinds of the feature combinations as shown in Table1.

Table2 shows the average precision (AP) of each recognition result. Each column in this table corresponds to each abbreviation representing the combination of features shown in Table1.

The results by Vis(BoF+HSV) can be regarded as the baseline results, since they are obtained using only visual features extracted from photos. As a result, the mean average precision (MAP) for 28 categories increased up to 80.87% by the proposed method, compared with 77.71% by the baseline. Especially, for the categories related to location-dependent concepts, MAP was improved by 6.57 points.

To clarify the obtained gains by introducing geo-related features, we showed the differences of AP values between the baseline (Vis) results and the results obtained by combining photo features with some geo-related features in Table3.

In terms of the average of the gains, the gain in case of using raw location vectors was small value, 0.89, while we obtained the best gain, 3.15, in case of AT (aerial features and text features). This implies the effectiveness of transforming raw location vectors into aerial and textual features.

Next, we explain the obtained gain regarding each type of concepts. For "Location-dependent concepts", ATD (all the features except location features) achieved the largest value, 6.67, which are reasonable results. Regarding raw

Table 2: The average precision of geotagged image recognition for 28 categories with eleven kinds of feature combinations.

Type of concepts	concept	BoF	HSV	Vis	VG	VT	BA	VA	VAT	VGTD	VATD	All
Location-dep.	Disneyland	68.37	70.98	73.20	78.41	84.16	84.14	84.14	**84.17**	84.10	84.16	84.09
	Tokyo tower	79.34	80.59	81.54	82.28	**83.95**	83.67	83.73	83.75	83.78	83.91	83.78
	AVG	73.86	75.79	77.37	80.34	**84.05**	83.90	83.93	83.96	83.94	84.03	83.93
Landmark	bridge	69.78	63.87	69.78	71.95	74.76	74.32	75.43	**76.44**	74.65	74.83	75.12
	building	78.58	74.35	79.92	80.64	**81.77**	79.80	81.41	81.26	81.35	80.90	81.04
	castle	80.13	81.21	82.09	82.91	83.72	82.99	83.77	83.75	83.74	**83.79**	83.73
	shrine	70.32	70.60	71.34	73.11	**77.28**	75.01	76.72	77.15	76.08	77.05	76.20
	railroad	74.43	72.59	77.71	77.35	**79.25**	77.03	78.62	78.70	78.83	79.04	78.98
	AVG	74.65	72.52	76.17	77.19	79.36	77.83	79.19	**79.46**	78.93	79.12	79.02
Geographical	beach	81.02	80.06	81.70	81.99	**83.76**	83.50	83.54	83.61	83.72	83.62	83.55
	lake	78.27	76.25	79.41	80.53	**82.87**	81.42	82.42	82.67	82.67	82.78	82.39
	river	74.56	74.22	75.81	76.74	80.61	80.11	81.22	81.24	80.06	80.93	80.78
	AVG	77.95	76.84	78.97	79.75	82.41	81.68	82.39	**82.51**	82.15	82.44	82.24
Space	park	70.42	69.05	71.24	71.95	78.29	74.99	77.63	**78.76**	77.76	77.53	77.33
	garden	77.57	77.61	79.80	81.06	82.39	81.06	82.35	**82.72**	82.20	82.59	82.27
	landscape	74.23	74.28	75.75	76.73	78.27	75.16	78.12	**78.37**	78.00	78.20	78.06
	AVG	74.08	73.65	75.60	76.58	79.65	77.07	79.36	**79.95**	79.32	79.44	79.22
Outdoor artifact	bicycle	76.05	71.85	76.72	77.22	**79.76**	77.90	78.80	79.39	78.87	79.45	78.80
	car	70.80	70.74	75.72	76.18	**77.54**	74.84	76.54	76.72	77.01	77.05	76.70
	vending machine	79.88	82.81	83.26	83.11	83.22	81.59	83.08	83.03	**83.39**	83.10	83.28
	statue	66.45	65.94	67.78	68.43	72.81	70.78	72.94	**73.25**	72.34	72.27	72.49
	graffiti	72.59	75.60	76.68	78.27	81.27	79.53	81.27	81.53	81.81	**81.82**	81.42
	AVG	73.15	73.39	76.03	76.64	**78.92**	76.93	78.53	78.78	78.69	78.74	78.54
Time-dep.	costume-play	75.67	79.42	80.29	81.53	83.95	83.68	84.04	84.04	83.73	**84.09**	83.73
	red leaves	80.77	82.53	83.27	83.51	83.83	82.08	83.82	83.77	83.96	83.91	**83.97**
	festival	73.31	74.89	76.90	77.30	**80.19**	76.20	79.59	79.97	79.23	80.08	79.58
	cherry blossom	80.22	79.70	82.13	82.29	**82.91**	80.56	82.67	82.77	82.80	82.88	82.55
	sunset	82.90	83.41	83.68	83.71	83.83	82.83	83.83	83.78	83.86	**83.90**	83.84
	AVG	78.57	79.99	81.25	81.67	82.94	81.07	82.79	82.86	82.72	**82.97**	82.73
Creature	flower	77.98	77.70	80.71	81.02	81.53	78.30	81.31	81.24	82.23	**82.65**	82.22
	bird	69.75	71.32	73.00	74.02	**81.38**	78.89	80.89	81.21	80.36	81.20	80.26
	cat	67.96	67.24	71.72	73.43	**74.63**	69.62	74.60	**74.63**	74.23	73.93	74.14
	AVG	71.89	72.09	75.14	76.16	79.18	75.60	78.93	79.03	78.94	**79.26**	78.87
Food	ramen noodle	82.59	80.85	83.03	83.09	**83.28**	82.77	83.18	83.11	83.00	83.10	82.97
	sushi	79.85	79.51	81.76	82.09	83.18	82.09	83.15	83.07	83.20	**83.28**	83.19
	AVG	81.22	80.18	82.40	82.59	**83.23**	82.43	83.17	83.09	83.10	83.19	83.08
Total AVG		75.49	75.33	77.71	78.60	**80.87**	79.10	80.67	80.86	80.61	80.79	80.59

location vectors, they worked for only Location-dependent concepts. This also implied the requirement to convert them to aerial or geo-text features.

Next, we show the weight of each feature estimated by MKL in case of fusing all the features in Table5, which can be regarded as expressing relative discriminative power of each feature. Regarding the average weight over all the concepts, the weight for photo features were the largest, 0.4717, the second largest was the weight of geo-text features, 0.2915, and the third was the one of aerial features, 0.1411. This indicated that geo-text features contained more geo-context information which helps image recognition than aerial features in general. However, geo-text features has a problem that it depends heavily on the output of reverse geo-coding service. For the area where geo-texts are poorly available such as the sea or the undeveloped areas, it might be useless. On the other hand, aerial photos are available anywhere on the earth. In this sense, aerial features can be regarded as more solid and stable features than geo-text features. Actually, as show in Table3, the gain of aerial features is comparable to the gain of geo-

Table 3: The obtain gains calculated from the difference to the baseline(Vis) in terms of MAP.

combination	calculation	Overview
Geo	(Vis+G)−(Vis)	gain by raw location vectors
Text	(Vis+T)−(Vis)	gain by geo-text features
Air	(Vis+A)−(Vis)	gain by aerial features
AT	(Vis+A+T)−(Vis)	gain by aerial features and geo-text features
GTD	(Vis+G+T+D)−(Vis)	gain by all except aerial features
ATD	(Vis+A+T+D)−(Vis)	gain by all except raw location features
All	(All)−(Vis)	gain by all the features

genre	Geo	Text	Air	AT	GTD	ATD	All
Location-dep.	2.97	6.69	6.56	6.60	6.57	6.67	6.57
Artifact	0.61	2.89	2.49	2.75	2.65	2.71	2.51
Landmark	1.03	3.19	3.02	3.30	2.76	2.95	2.85
Space	0.98	4.05	3.77	4.35	3.72	3.84	3.62
Time-dep.	0.41	1.69	1.53	1.61	1.46	1.72	1.48
Food	0.19	0.84	0.77	0.69	0.71	0.80	0.68
Geographical	0.78	3.44	3.42	3.53	3.18	3.47	3.27
Creature	1.02	4.04	3.79	3.88	3.80	4.11	3.73
AVG	**0.89**	**3.16**	**2.96**	**3.15**	**2.89**	**3.08**	**2.88**

text features. The weights of location vectors were relatively small, because raw coordinate features are not effective to describe location context on the place. This shows that transformation of geotag vectors into aerial features or geo-text features are very effective methods to utilize geo-context information on the place. The weights of date features is also small but larger than raw geo-location features.

Regarding each concept, most of the concepts gathered the largest weights on photo features, while the eight concepts obtained the largest weights on geo-text features. For "beach", the weight of aerial features was largest, which is a reasonable result, because "beach" is easy to recognize in the aerial photos as the border between lands and the sea. For "statue", the weight of raw location vectors are biggest, since the locations of famous statues are limited and can be covered with the raw location vectors in the training data. For "cherry blossom", the weight of date features is the largest, since the season when cherry blossom blooms is only spring from late March to late April.

Table 4: Part of obtained average geo-text features from Yahoo! Japan Local API. We show the top ten words in the descending order of the value of the corresponding bins for the three concepts the geo-text kernel weights of which are relatively larger and smaller.

Concepts assigned with large weights					
costume-play		**Disneyland**		**castle**	
Ariake	0.0391	parking	0.117	Himeji	0.0324
building	0.0388	garage	0.1074	city	0.0252
Tokyo	0.025	Tokyo	0.0649	company	0.0201
parking	0.0223	Maihama	0.0424	building	0.0199
Harajuku	0.0222	resort	0.0416	school	0.0173
center	0.0181	park	0.0347	bridge	0.0161
park	0.0144	hotel	0.0321	post office	0.0158
apartment	0.013	Kasai	0.024	temple	0.0149
Makuhari	0.0128	bay	0.0197	park	0.0143
number	0.0124	entrance	0.0139	primay school	0.0133

Concepts assigned with small weights					
flower		**vending machine**		**ramen noodle**	
building	0.1089	building	0.0286	building	0.0522
embassy	0.0395	company	0.02	company	0.0173
hall	0.0255	post office	0.0187	post office	0.0169
Roppongi	0.0211	nursery	0.0168	city	0.0117
Kamiyacho	0.0199	Toyama	0.0154	center	0.0115
Nippon	0.0186	primary school	0.0141	hotel	0.0109
Azabu	0.0183	city	0.0132	primary school	0.0109
saint	0.0183	temple	0.0123	temple	0.0106
Toranomon	0.0181	center	0.0122	Tokyo	0.0102
Iikura	0.018	kindergarde	0.0122	nursery	0.0099

Table 5: Feature weights estimated by MKL in case of fusing all the features. The number written in the red ink is the largest weight among each concept.

category		photo(BoF+HSV)	aerial	location	texts	time
Location-dependent	Disneyland	0.0001(0.0001,0.0001)	0.3288(0.1460,0.1827,0.0001,0.0001)	0.0004	0.6706	0.0000
	Tokyo tower	0.1479(0.1474,0.0005)	0.4230(0.0000,0.0011,0.4217,0.0001)	0.0007	0.4280	0.0003
	AVG	**0.0740(0.0737,0.0003)**	**0.3759(0.0730,0.0919,0.2109,0.0001)**	**0.0006**	**0.5493**	**0.0002**
Landmark	bridge	0.3843(0.2690,0.1153)	0.3408(0.0529,0.0557,0.0295,0.2028)	0.0089	0.2289	0.0371
	shrine	0.4504(0.0999,0.3505)	0.0690(0.0012,0.0020,0.0001,0.0657)	0.0043	0.4307	0.0456
	building	0.6147(0.3897,0.2251)	0.1207(0.0500,0.0425,0.0004,0.0278)	0.0455	0.2123	0.0068
	castle	0.2441(0.0654,0.1787)	0.1094(0.0877,0.0041,0.0124,0.0053)	0.0521	0.5941	0.0002
	railroad	0.6918(0.3016,0.3903)	0.1004(0.0212,0.0151,0.0004,0.0637)	0.0001	0.1869	0.0208
	AVG	**0.4771(0.2251,0.2519)**	**0.1480(0.0426,0.0239,0.0085,0.0731)**	**0.0222**	**0.3306**	**0.0221**
Geographical	lake	0.3539(0.2184,0.1355)	0.2694(0.0749,0.0877,0.0150,0.0918)	0.0040	0.3435	0.0292
	river	0.3465(0.2492,0.0973)	0.3181(0.0063,0.0523,0.0614,0.1981)	0.0108	0.2447	0.0800
	beach	0.1888(0.1816,0.0073)	0.5003(0.0108,0.1632,0.0013,0.3250)	0.0015	0.2934	0.0159
	AVG	**0.2964(0.2164,0.0800)**	**0.3626(0.0307,0.1011,0.0259,0.2049)**	**0.0054**	**0.2938**	**0.0417**
Space	park	0.3971(0.1688,0.2283)	0.0881(0.0034,0.0101,0.0205,0.0541)	0.0169	0.4178	0.0800
	garden	0.4740(0.1338,0.3402)	0.1227(0.0080,0.0033,0.0170,0.0944)	0.0269	0.3512	0.0252
	landscape	0.5634(0.2895,0.2739)	0.0452(0.0000,0.0001,0.0078,0.0374)	0.0043	0.3587	0.0283
	AVG	**0.4782(0.1974,0.2808)**	**0.0853(0.0038,0.0045,0.0151,0.0620)**	**0.0161**	**0.3759**	**0.0445**
Outdoor artifact	statue	0.2223(0.2156,0.0067)	0.1381(0.0587,0.0041,0.0464,0.0288)	0.4492	0.0336	0.1568
	car	0.6772(0.3290,0.3481)	0.0285(0.0065,0.0000,0.0001,0.0219)	0.0146	0.2678	0.0119
	bicycle	0.5520(0.4410,0.1111)	0.0642(0.0076,0.0143,0.0003,0.0420)	0.0098	0.3391	0.0349
	graffiti	0.2304(0.0682,0.1621)	0.2846(0.0141,0.1064,0.1043,0.0598)	0.0493	0.3728	0.0630
	vending machine	0.8755(0.2851,0.5904)	0.0533(0.0064,0.0127,0.0094,0.0248)	0.0032	0.0085	0.0594
	AVG	**0.5115(0.2678,0.2437)**	**0.1137(0.0187,0.0275,0.0321,0.0355)**	**0.1052**	**0.2044**	**0.0652**
Time-dependent	red leaves	0.6158(0.2298,0.3861)	0.1490(0.0836,0.0297,0.0037,0.0318)	0.0066	0.0025	0.2261
	cherry blossom	0.4437(0.3291,0.1147)	0.0402(0.0002,0.0019,0.0381,0.0001)	0.0016	0.0093	0.5051
	sunset	0.8096(0.4022,0.4074)	0.0635(0.0001,0.0004,0.0000,0.0631)	0.0000	0.0002	0.1267
	costume-play	0.0485(0.0352,0.0133)	0.0036(0.0001,0.0001,0.0001,0.0033)	0.0126	0.9291	0.0062
	festival	0.5667(0.2347,0.3320)	0.1248(0.0012,0.0004,0.0005,0.1226)	0.0015	0.2154	0.0916
	AVG	**0.4969(0.2462,0.2507)**	**0.0762(0.0170,0.0065,0.0085,0.0442)**	**0.0045**	**0.2313**	**0.1911**
Creature	cat	0.6929(0.2690,0.4239)	0.0718(0.0002,0.0032,0.0007,0.0677)	0.0100	0.2217	0.0036
	bird	0.1869(0.0888,0.0980)	0.0296(0.0004,0.0039,0.0002,0.0252)	0.0000	0.7126	0.0708
	flower	0.8196(0.3356,0.4840)	0.0063(0.0001,0.0007,0.0015,0.0041)	0.0179	0.0269	0.1293
	AVG	**0.5665(0.2312,0.3353)**	**0.0359(0.0002,0.0026,0.0008,0.0323)**	**0.0093**	**0.3204**	**0.0679**
Food	ramen noodle	0.9317(0.4588,0.4729)	0.0563(0.0000,0.0000,0.0092,0.0470)	0.0018	0.0005	0.0097
	sushi	0.6789(0.2435,0.4354)	0.0021(0.0001,0.0005,0.0000,0.0016)	0.0043	0.2616	0.0532
	AVG	**0.8053(0.3512,0.4541)**	**0.0292(0.0001,0.0002,0.0046,0.0243)**	**0.0030**	**0.1310**	**0.0314**
Total AVG		**0.4717(0.2314,0.2403)**	**0.1411(0.0229,0.0285,0.0286,0.0611)**	**0.0271**	**0.2915**	**0.0685**

In addition, we show part of the geo-text features averaged over each concept in Table4. In this table, capitalized words represent place names. Note that this table includes many place names in Japan, because we limited geotagged images taken within Japan when fetching images via Flickr API. Geo-text features were effective for the concepts the representative location of which are fixed, while they are not effective for the concepts existing anywhere such as “flower” and “vending machine”.

From the above observations, we conclude that geo-text features are the most effective and aerial features are the second, while raw location features and date features are not so helpful.

6 Conclusion

In this paper, we proposed introducing Multiple Kernel Learning (MKL) into geotagged image recognition to estimate the contribution weights of visual fea-

tures of photo images and geo-information features. In the experiments, we made experiments with twenty-eight concepts selected from eight different types of concepts. The experimental results showed that using geo-textual features and aerial features can be regarded as very helpful for most of the concepts except "food" concepts. As a result, the mean average precision (MAP) for 28 categories increased up to 80.87% by the proposed method, compared with 77.71% by the baseline.

For future work, we plan to make more large-scale experiments with much more categories. We also plan to carry out multi-class classification experiments.

Acknowledgement. Keita Yaegashi is currently working for Rakuten Institute of Technology.

References

1. Luo, J., Yu, J., Joshi, D., Hao, W.: Event recognition: Viewing the world with a third eye. In: Proc. of ACM International Conference Multimedia. (2008)
2. Joshi, D., Luo, J.: 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)
3. Yaegashi, K., Yanai, K.: Can geotags help image recognition ? In: Proc. of Pacific-Rim Symposium on Image and Video Technology. (2009)
4. Yaegashi, K., Yanai, K.: Geotagged photo recognition using corresponding aerial photos with multiple kernel learning. In: Proc. of IAPR International Conference on Pattern Recognition. (2010)
5. Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Mei, T., Zhang, H.: Correlative multi-label video annotation. In: Proc. of ACM International Conference Multimedia. (2007) 17–26
6. Hays, J., Efros, A.A.: IM2GPS: Estimating geographic information from a single image. In: Proc. of IEEE Computer Vision and Pattern Recognition. (2008)
7. Kalogerakis, E., Vesselova, O., Hays, J., Efros, A., Hertzmann, A.: Image sequence geolocation with human travel priors. In: Proc. of IEEE International Conference on Computer Vision. (2010)
8. 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. (2004) 59–74
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
10. Varma, M., Ray, D.: Learning the discriminative power-invariance trade-off. In: Proc. of IEEE International Conference on Computer Vision. (2007) 1150–1157
11. Gehler, P., Nowozin, S.: On feature combination for multiclass object classification. In: Proc. of IEEE International Conference on Computer Vision. (2009)
12. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision **73** (2007) 213–238
13. Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: Large Scale Multiple Kernel Learning. The Journal of Machine Learning Research **7** (2006) 1531–1565