Geotagged Photo Recognition using Corresponding Aerial Photos with Multiple Kernel Learning

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

Abstract

In this paper, we treat with generic object recognition for geotagged images. As a recognition method for geotagged photos, we have already proposed exploiting aerial photos around geotag places as additional image features for visual recognition of geotagged photos. In the previous work, to fuse two kinds of features, we just concatenate them. Instead, in this paper, we introduce Multiple Kernel Learning (MKL) to integrate both features of photos and aerial images. MKL can estimate the contribution weights to integrate both kinds of features. In the experiments, we confirmed effectiveness of usage of aerial photos for recognition of geotagged photos, and we evaluated the weights of both features estimated by MKL for eighteen concepts.

1 Introduction

Recently, the number of geotagged photos on the Web is increasing rapidly, since people can see geotagged photos uploaded to some photo sharing Web sites on a map. Geotags for photos are easy to obtain with GPS-equipped cameras or carrying portable small GPS devices with digital cameras. 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 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.

To utilize geotags in visual image recognition, we have already proposed two methods in [7]: (1) combining values of latitude and longitude with visual features of a photo image, and (2) combining visual feature extracted from aerial photo images with visual feature extracted from a photo image.

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 [7].

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 [7].

Since in the method (2) of [7] we combine 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. They are expected to vary depending on target concepts. For the concepts which can be directly recognized from aerial photos such as "beach" and "mountain", the contribution of aerial photos is expected to be much, while it is expected to be less for the unrecognizable concepts from aerial photos such as "flower" and "cat". In this paper, we introduce 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, we evaluate the weights of both features using MKL for eighteen concepts.

As related work, J. Luo et al.[4] is the most similar to our work. They also proposed to exploit aerial photos corresponding to where a photo was taken for recognizing geotagged photos. However, they focused only event concepts such as "baseball", "at beach" and "in park", which can be directly recognized from aerial photos, and did not evaluate the contribution of aerial images for recognition with various kinds of concepts.

The rest of this paper is organized as follows:

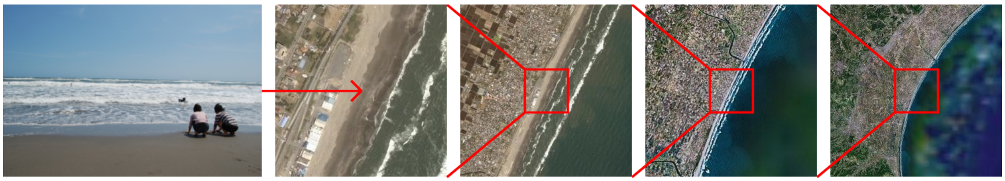

Figure 1. Correspondences between a geotagged photo and aerial images.

Section 2 and 3 explains the overview and the procedure of geotagged image recognition with MKL, respectively. Section 4 shows the experimental results and discusses them, and we conclude this paper in Section 5.

2 Overview

The objective of this paper is to evaluate the contribution weights of both visual features of photo images and aerial images for image recognition using Multiple Kernel Learning (MKL) 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 [1]. 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. As representation of geotags, we also adopt the bag-of-features representation of aerial photos around the geotag location.

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

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.

3.1 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 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 negative samples. We select 200 positive samples and 200 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. 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.

3.2 Visual Features

To extract visual feature vectors from photos and aerial images, we use the bag-of-features method [1]. 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 [3] 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 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.

3.3 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 treats with 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 [6, 2]. 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 chi-square RBF kernel.

Sonnenburg et al.[5] 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 [5]. In the experiment, we use the MKL library included in the SHOGUN toolbox as an implementation of MKL.

4 Experiments

In the experiments, we used eighteen concepts. To selected the eighteen concepts, we first define four rough types of concepts and then select several concepts for each type as follows:

Location-specific concept (l) [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.

Recognizable concept (r) [7 concepts]
beach, castle, bridge, railroad, park, shrine, landscape
Since there are possibility of recognizing these concepts on aerial photos directly, improvement of performance is expected.

Unrecognizable concept (u) [5 concepts]
flower, cat, sushi, ramen noodle, vendor machine
These concepts are very difficult to recognize on aerial images.

Time-dependent concept (t) [4 concepts]
sunset, cherry blossom, red leaves, festival
These concepts depend on time or seasons rather than locations.

Location-specific and recognizable concepts are expected to have correlation to aerial images, while unrecognizable 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.

In the experiments, we carried out two-class image classification and estimate weights to integrate features of photos and four kinds of aerial images in four different levels using MKL for the eighteen concepts, and as a baseline we also made experiments on image classification with only image features of photos without fusion using a standard SVM.

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

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

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.

Table 1. Average precision for eighteen concepts by only photos and by fusion of photos and aerial images.

(type) **concept**	photo	MKL	gain
(l) Disneyland	68.00	**84.06**	+16.06
(r) park	67.43	**76.04**	+8.61
(r) shrine	72.79	**78.53**	+5.74
(t) festival	72.32	**77.75**	+5.43
(r) bridge	69.51	**74.89**	+5.38
(r) landscape	73.71	**78.37**	+4.66
(r) beach	80.10	**83.85**	+3.75
(t) red leaves	79.18	**82.45**	+3.27
(l) Tokyo tower	80.84	**83.84**	+3.00
(r) castle	81.28	**83.53**	+2.23
(u) sushi	80.11	**81.93**	+1.82
(r) railroad	74.70	**76.20**	+1.50
(u) flower	77.00	**78.48**	+1.48
(t) cherry blossom	80.94	**81.61**	+0.67
(u) ramen noodle	82.34	**82.70**	+0.36
(u) cat	73.98	**74.26**	+0.28
(u) vendor machine	83.17	**83.43**	+0.26
(t) sunset	83.01	**83.11**	+0.10
AVERAGE	76.69	**80.28**	+3.59

4.1 Experimental Results

We show the results by MKL-based fusion and the baseline method in Table 1 and the estimated weights for features of photos and features of four kinds of aerial images in Table 2. In both the tables, the alphabets just before concept names represent the types of concepts. (l), (r), (u) and (t) means location-specific concepts, recognizable concepts,

Table 2. Weights of features of images and four kinds of aerial images. The results are sorted by the decending order of the weight of photos.

(type) **concept**	photo	level1	level2	level3	level4	(type) **concept**	photo	level1	level2	level3	level4
(u) ramen noodle	**0.873**	0.002	0.000	0.037	0.088	(r) bridge	**0.582**	0.077	0.044	0.070	0.226
(u) vendor machine	**0.794**	0.058	0.009	0.074	0.065	(t) red leaves	**0.523**	0.141	0.006	0.062	0.269
(t) cherry blossom	**0.774**	0.038	0.006	0.093	0.090	(r) castle	**0.523**	0.166	0.004	0.099	0.208
(u) cat	**0.743**	0.028	0.008	0.063	0.158	(t) festival	**0.518**	0.058	0.001	0.185	0.238
(t) sunset	**0.729**	0.055	0.058	0.016	0.142	(r) shrine	**0.507**	0.033	0.009	0.061	0.391
(u) flower	**0.658**	0.000	0.042	0.051	0.249	(r) park	**0.437**	0.073	0.012	0.045	0.433
(r) railroad	**0.604**	0.106	0.014	0.052	0.224	(r) beach	**0.392**	0.115	0.173	0.055	0.265
(r) landscape	**0.604**	0.078	0.024	0.093	0.199	(l) Disneyland	**0.384**	0.095	0.236	0.131	0.153
(u) sushi	**0.596**	0.062	0.015	0.062	0.266	(l) Tokyo tower	0.364	0.008	0.002	**0.396**	0.231

unrecognizable concepts, and time-dependent concepts, respectively.

In Table 1, all the AP values of all the concepts were improved by aerial image fusion. Recognizable concepts and location concepts tend to achieve higher gains, while unrecognizable concepts and time-dependent concepts tend to produce low gains.

Especially, the gain on "Disneyland" was by far the best among all the concepts, since "Disneyland" photos are alway taken inside Disneyland parks, and aerial photos corresponding to the geotag place are always the same. The gain on another location-specific concept "Tokyo tower" is not so much, since photos on "Tokyo tower" were not take inside it, but usually taken from many other places.

Regarding unrecognizable concepts such as "sushi" and "flower", small gains were obtained. This is because the places where "sushi" and "flower" photos are taken might have causal relationship to geographical features which appear directly in aerial photos. The places where "sushi" and "flower" photos are taken are unevenly distributed, and are usually commercial areas where there are many sushi restaurants for "sushi" or non-commercial areas and farming areas for "flower". We expect that this indirect causal relation goes for many unrecognizable concepts other than flowers and sushi.

Regarding the estimated weights shown in Table 2, most of the concepts have the highest weights on photo features and the second highest weights on level 4 aerial images which is the finest one among four levels of aerial images. For location-specific concepts such as Disneyland, the area covered by the finest level-4 image is too small to represent specific features to the Disneyland.

The weights of photos for unrecognizable concepts and time-dependent concepts tend to be larger, while the weights of photos for recognizable for concepts and location concepts tend to be smaller.

From these results shown in Table 1 and Table 2, using finer aerial photos for recognition is more effective for recognizable concepts, and using medium-level aerial photos rather than the finest ones are more helpful for location-specific concepts. Moreover, even for unrecognizable and time-dependent concepts, aerial photos is still a little beneficial due to indirect causal relation.

5 Conclusion

In this paper, we proposed introducing Multiple Kernel Learning (MKL) into geotagged image recognition to estimate the contribution weights of both visual features of photo images and aerial images. In the experiments, we made experiments with eighteen concepts selected from four different types of concepts. The experimental results showed that using aerial images can be regarded as very helpful for recognizable concepts such as "beach" and "park" and still a little beneficial even for unrecognizable concepts such as "cat" and "noodle" likely due to indirect causal relation.

For future work, we plan to make more extensive experiments with much more concepts and more precise aerial photos, and to use other kinds of data sources than aerial photos for geotags.

References

[1] 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.

[2] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Proc. of IEEE International Conference on Computer Vision*, 2009.

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

[4] 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.

[5] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large Scale Multiple Kernel Learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006.

[6] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Proc. of IEEE International Conference on Computer Vision*, pages 1150–1157, 2007.

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