Exploring Geotagged Images for Land-Use Classification

Daniel Leung
Electrical Engineering & Computer Science
University of California, Merced
Merced, CA 95343
cleung3@ucmerced.edu

Shawn Newsam
Electrical Engineering & Computer Science
University of California, Merced
Merced, CA 95343
snewsam@ucmerced.edu

ABSTRACT

On-line photo sharing websites such as Flickr not only allow users to share their precious memories with others, they also act as a repository of all kinds of information carried by their photos and tags. In this work, we investigate the problem of geographic discovery, particularly land-use classification, through crowdsourcing of geographic information from Flickr's geotagged photo collections. Our results show that the visual information contained in these photo collections enables us to classify three types of land-use classes on two university campuses. We also show that text entries accompanying these photos are informative for geographic discovery.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis; I.5.4 [**Pattern Recognition**]: Applications; H.2.8 [**Database Management**]: Database Applications—*spatial databases and GIS*

Keywords

Geotagged images, land-use classification

1. INTRODUCTION

On-line photo sharing websites such as Flickr [1] and Picasa [2] have become a popular means for people to share their precious memories. However, these datasets contain more than just memories; they potentially contain a wealth of information about the world. We usually think of the five W's and one H (Who, What, Where, When, Why, and How) as only applying to text documents but each of the photos in these collections along with its metadata can also provide us with these six types of information. As many researchers are discovering, innovative knowledge discovery is possible through analyzing these photo collections. With more than 180 million georeferenced photos available from Flickr, our goal in this paper is to map what-is-where on the surface of the Earth using the What and Where aspects of the information. In particular, we use georeferenced photos collected from two university campuses to perform land-use classification.

In traditional remote sensing, overhead imagery is used to distinguish different types of land-cover in a given region; however, it is more difficult to tell the type of land-use a certain land-cover class belongs to. For example it is easy to locate a region with large buildings and parking lots in the satellite view mode in Google Maps but the satellite view does not as easily indicate whether the region belongs to a shopping center or a warehouse. To find out the answer, one can switch to the street view mode and see the images of nearby objects and scenes taken from the ground level. In this work, we use the term "proximate sensing" to describe geographic discovery using ground level images of nearby objects and scenes.

The novel contribution of this work is to use proximate sensing to complement the shortcoming of remote sensing for land-use classification. We propose a novel framework using state-of-the-art techniques in multimedia content analysis, in particular automated image and text understanding, to perform geographic discovery in large collections of georeferenced photos.

2. RELATED WORK

We consider geographic discovery to be a process that derives knowledge about what-is-where on the surface of the earth in the broad sense of the term "what". Simply put, it can be used to generate maps not only of the physical aspects of our world, such as the terrain, but also of the cultural and behavioral aspects. While there has been relatively little work on using georeferenced images for geographic discovery, we feel it has significant potential for realizing the full worth of georeferenced photo collections as a repository for geographic imformation, particularly as an alternate to traditional means of geographic inquiry. Examples of work in this area include using large collections of georeferenced images to discover spatially varying (visual) cultural differences among concepts such as "wedding cake" [12]; to discover interesting properties about popular cities and landmarks such as the most photographed locations [4]; to estimate weather satellite images using widely distributed Web cameras [8]; and to create a map-like partitioning of a country-sized region into geographically coherent subregions [5]. Our work on proximate sensing as applied to georeferenced photo collections, however, represents a more comprehensive framework for geographic discovery particularly

GeoMM'12, October 29, 2012, Nara, Japan.

(a) Sample Academic images

(b) Sample Sports images

(c) Sample Residential images

Figure 1: Sample images used to perform land-use classification.

of phenomena often not observable through other means. We previously investigated proximate sensing for land-cover classification [10, 11]. We here investigate the more difficult problem of land-use classification.

The salient aspects of the work in this paper includes:

- To our knowledge, our work is the first to consider georeferenced images for land-use classification.
- Our approach does not require individual images to be manually labelled for training. Instead, labels are propagated from existing land-use maps to the images which greatly reduces the effort required to train the classifiers.

3. DATASET

Two university campuses (University of California, Berkeley and Stanford University) are selected as our study areas to learn the land-use classification models. We use the Flickr application programming interface (API) to download Flickr images located within the campus regions. For each campus, a land-use map is derived manually according to its campus map. Three land-use classes are considered: Academic, Residential, and Sports. Each downloaded image is then assigned a ground truth land-use class label according to its geographic location on the map. Figure 1 shows sample images from each class.

Besides training classifiers at the image level, we also train classifiers at the group-of-images level. Since the content of user contributed photos as well as the distribution of user contributions are very diverse, having many photos contributed by the same user may bias the training data. As a result, we partition the dataset into groups based on users (owners of photos), geographic locations, and the time when the photos are taken. For each campus, we first partition all images into 20 sub-regions based on their geographic locations using k-means clustering. These sub-regions are independent from the land-use classes. Finally, within each sub-region we group all the images taken by the same user on the same day. Our grouping methodology is based on the assumption that same user takes photos of similar scenes in a nearby location within a short period of time. Tables 1-3 provides the details of the three campus datasets, including the sizes of the training and test sets.

Table 1: Dataset used in visual image level classification experiments. Counts are the number of images.

	Academic	Sports	Residential	Total
Berkeley Training	5029	2153	463	7645
Berkeley Test	2000	1500	50	3550
Stanford Training	1524	2772	747	5043
Stanford Test	200	200	100	500

Table 2: Dataset used in visual group level classification experiments. Counts are the number of groups (of images).

	Academic	Sports	Residential	Total
Berkeley Training	1517	365	122	2004
Berkeley Test	200	50	30	280
Stanford Training	504	204	186	894
Stanford Test	50	30	30	110

3.1 Features

We use Jiang's implementation [9] of bag of visual words (BoW) with a soft-weighing scheme to extract a BoW feature from each image. Instead of assigning a visual word nearest to a keypoint detected, the soft-weighing scheme as-

Table 3: Dataset used in textual group level classification experiments. Counts are the number of groups (of images).

	Academic	Sports	Residential	Total
Berkeley Training	1425	348	123	1896
Berkeley Test	150	30	20	200
Stanford Training	421	193	141	755
Stanford Test	50	20	20	90

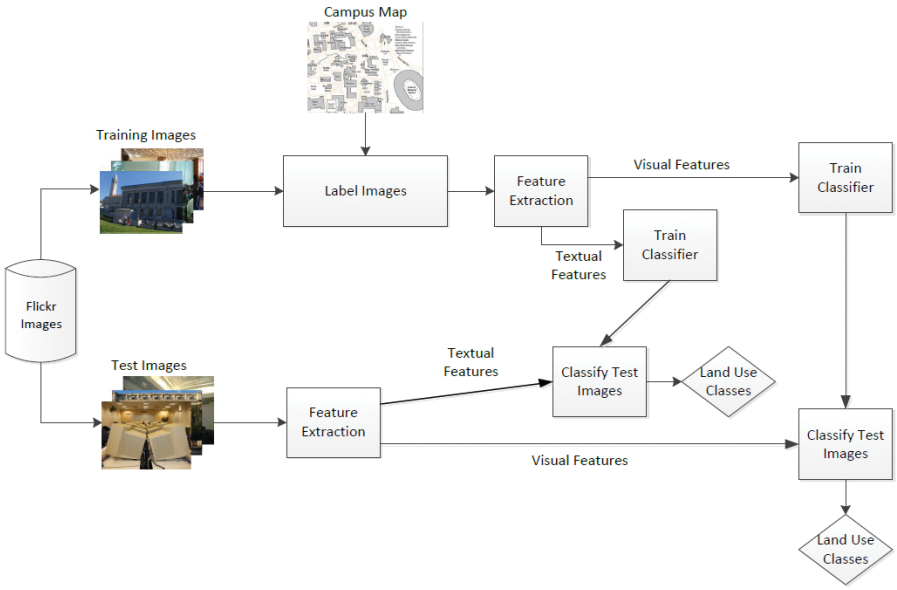


Figure 2: Framework of the proposed approach.

signs the four nearest visual words to a detected keypoint. A dictionary of 500 visual words is used in our implementation.

Since Flickr images commonly have user-supplied text associated with them, we also study the effectiveness of this text for land-use classification. To obtain the text features, we create a dictionary of terms based on the words extracted from the image title, descriptions, and tags associated with each image. After applying stopping and stemming, a total of 2457 unique terms are recorded, and out of these the 1949 most frequent terms are selected as the dictionary.

The text analysis is performed at the group level since there is typically not enough text associated with the individual images for effective classification. Each of the text components associated with an image is parsed into a set of terms (words) which are then pooled at the group level. At the moment, all terms are given equal weight although different weightings based on the relative importance of the components would be an interesting extension. As a result, each group of images is represented by a histogram of terms among the dictionary.

It is unlikely that classification at the term level would be effective due to the sparse occurrence of terms, so we apply a latent semantic approach from text document analysis in which a hidden topic $z \in Z = \{z_1, ..., z_K\}$ is associated with the observed occurrence of a word $w \in W = \{w_1, ..., w_M\}$ in a document (image group) $d \in D = \{d_1, ..., d_N\}$. This latent layer also helps overcome the problems of synonymy and polysemy.

We use a generative probabilistic technique termed probabilistic latent semantic analysis (pLSA) [6, 7] to learn the hidden topics. A pLSA model can be expressed as

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

where $P(w|d)$ is the observed word distributions over documents.

To learn the distribution of words over hidden topics, we use the Expectation Maximization (EM) algorithm. In the E-step, the posterior probabilities for the hidden topics are evaluated:

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')},$$

while in the M-step the parameters of the E-step are estimated based on the result of the E-steps:

$$P(w|z) = \frac{\sum_d n(d,w)P(z|d,w)}{\sum_{d,w'} n(d,w')P(z|d,w')},$$

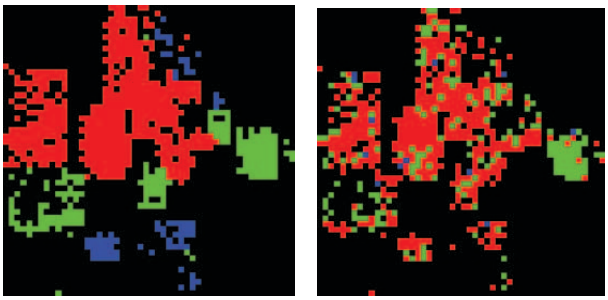
$$P(d|z) = \frac{\sum_w n(d,w)P(z|d,w)}{\sum_{d',w} n(d',w)P(z|d',w)},$$

$$P(z) = \frac{1}{R}\sum_{d,w} n(d,w)P(z|d,w), R \equiv \sum_{d,w} n(d,w).$$

Instead of defining the number of hidden topics as the number of ground truth classes as most pLSA approaches do, we use pLSA as a tool to reduce the dimensionality of the term histogram of each image group by computing the distributions over hidden topics for the image groups, $P(z|d)$. In our past experiments, we noticed that distributions of hidden topics provide an explicit representation of the image groups that is more robust than distributions over terms. To evaluate the hidden topic distribution of a novel image group, the EM algorithm is applied with a fixed $P(w|z)$ learned from the training step.

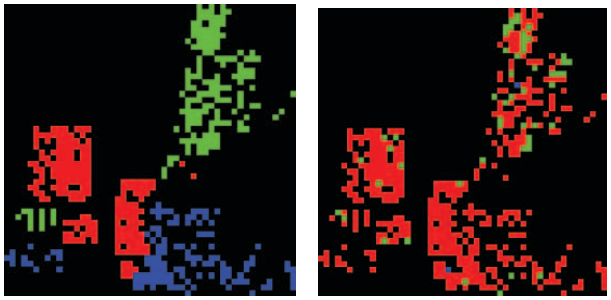
4. EXPERIMENTS

The goal of the experiments is to use the georeferenced photos as represented by their visual or text features to perform land-use classification. We formulate this as a supervised classification problem in which support vector machines (SVMs) are trained on a labeled subset of the data and then used to assign labels to a disjoint held-out set. We compare applying the SVMs 1) at the image level and 2) applying them at the group level. These two modes are



(a) (b)

Figure 3: Land-use classification of the Berkeley campus. (a) Ground truth map. (b) Predicted map using classifiers trained on the Stanford dataset. Academic, Sports, and Residential are denoted by red, green, and blue.



(a) (b)

Figure 4: Land-use classification of the Stanford campus. (a) Ground truth map. (b) Predicted map using classifiers trained on the Berkeley dataset. Academic, Sports, and Residential are denoted by red, green, and blue.

described in the following subsections. Performance is evaluated at two levels. First at the image or group level, and the second by comparing the predicted land-use maps to the ground truth maps derived manually from the campus maps. The workflow of the experiments is illustrated in figure 2.

The SVMs are implemented using the LIBSVM package [3]. We use radial basis function (RBF) kernels and determine optimal values for the two parameters, the penalty term C and the kernel width, through grid-search on a random partitioning of the training set.

4.1 Image Level Classification

In this set of experiments, the SVMs are used to classify individual images as being Academic, Sports, or Residential. We use a one-versus-all approach to train the SVMs for each campus and class. To generate a predicted land-use map, we first divide each campus into a map of 50x50 regions (tiles). The trained SVMs are then used to label each of the test images within each tile. As a result, each tile is represented by three ratio of images being classified into the three respective classes–e.g., Academic versus other. We use the label of the highest ratio to assign a land-use label to each tile for comparison with the ground truth maps.

4.2 Group Level Classification

In this set of experiments, the SVMs are used to label the groups directly. A single visual feature is computed for each group by averaging the features from all the images within the group. The text features are already computed at the group level so no aggregation is needed. Due to the fact that not all images have accompanying text, the size of the training sets for the text features is slightly reduced from that of the image features.

5. RESULTS

The different approaches are evaluated based on their accuracy, precision, and recall. Accuracy is the number of correctly predicted labels (both positive and negative) for a particular classifier normalized by the size of the particular test set. It is reported as a percentage. For example, if the binary classifier trained on the Berkeley Academic dataset classifies 320 of the 500 images in the Stanford test set correctly then the accuracy is 60%. An "accuracy" value can also be computed when a classifier is used to detect a class other than the one it is trained for. For example, the binary classifier trained on the Berkeley Academic dataset can be applied to the Stanford test with the Sports images as the positive labels. In this case, a low accuracy value would be a good result. Precision is the fraction of images that are assigned a particular class that actually have that class and recall is the fraction of the images with a particular class that are assigned that class.

Table 4 summarizes the visual image level classification accuracy of each classifier trained using one class and applied to detect another class for both the intra- and inter-campus cases. These results clearly demonstrate that the classifiers are learning discriminating visual features for the three different land-use classes. Classifiers trained on a particular class are always more likely to detect that class than another. As might be expected, the intra-campus results are better than the inter-campus ones. However, the approach is seen to generalize quite well from one campus to another. The high values for the Residential class can be explained by the relatively few images in this test set especially for the Berkeley campus.

Table 5 summarizes the visual group level accuracy. The classification performance at the visual group level is comparable to that at the visual image level. This is significant since grouping the images results in smaller number of training samples, greatly reducing the computational cost of the SVM learning.

Table 6 summarizes the textual group level classification accuracy. We can see that despite the diversity of text accompanying the images, pLSA is able to extract sufficient discriminating semantic information to distinguish the classes.

Table 7 summarizes the precision and recall rates for each approach. These results corroborate those of the accuracy results: the classifiers are able to distinguish between different classes; the intra-campus results are better than the inter-campus but the generalization is still good; and that the visual group and textual group results are comparable to that of the visual image. The poor precision and recall values for the Berkeley Residential class are again a result of there being too few images in this dataset.

Finally, we produce land-use maps using the visual image

Table 4: Visual image level classification accuracy

	Berkeley Test Sets			Stanford Test Sets		
Training Sets	**Academic**	**Sports**	**Residential**	**Academic**	**Sports**	**Residential**
Berkeley Academic	**82**	17	36	**62**	27	39
Berkeley Sports	18	**84**	68	42	**72**	65
Berkeley Residential	44	57	**97**	59	59	**80**
Stanford Academic	**64**	36	69	**75**	31	58
Stanford Sports	28	**73**	54	28	**85**	44
Stanford Residential	44	57	**96**	55	54	**84**

Table 5: Visual group level classification accuracy

	Berkeley Test Sets			Stanford Test Sets		
Training Sets	**Academic**	**Sports**	**Residential**	**Academic**	**Sports**	**Residential**
Berkeley Academic	**74**	19	18	**58**	24	33
Berkeley Sports	25	**86**	84	46	**78**	67
Berkeley Residential	28	82	**90**	55	73	**73**
Stanford Academic	**60**	36	39	**68**	30	45
Stanford Sports	29	**81**	76	40	**84**	60
Stanford Residential	29	82	**89**	55	73	**73**

Table 6: Textual group level classification accuracy

	Berkeley Test Sets			Stanford Test Sets		
Training Sets	**Academic**	**Sports**	**Residential**	**Academic**	**Sports**	**Residential**
Berkeley Academic	**80**	12	16	**66**	21	28
Berkeley Sports	21	**88**	85	39	**81**	70
Berkeley Residential	25	85	**90**	44	78	**78**
Stanford Academic	**66**	33	37	**72**	23	46
Stanford Sports	22	**87**	82	37	**83**	70
Stanford Residential	25	85	**90**	44	78	**78**

Table 7: Precision and recall rates

		Visual Image		Visual Group		Textual Group	
Training Sets	**Test Sets**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Berkeley Academic	Berkeley Academic	0.80	0.92	0.76	0.94	0.80	0.98
Berkeley Sports	Berkeley Sports	0.92	0.67	0.81	0.26	0.75	0.30
Berkeley Residential	Berkeley Residential	0.13	0.14	1.0	0.03	0	0
Berkeley Academic	Stanford Academic	0.52	0.93	0.52	0.98	0.62	0.96
Berkeley Sports	Stanford Sports	0.79	0.41	0.80	0.27	0.67	0.30
Berkeley Residential	Stanford Residential	0.50	0.05	0	0	0	0
Stanford Academic	Berkeley Academic	0.83	0.46	0.74	0.68	0.82	0.70
Stanford Sports	Berkeley Sports	0.67	0.71	0.48	0.38	0.61	0.37
Stanford Residential	Berkeley Residential	0.03	0.06	0	0	0	0
Stanford Academic	Stanford Academic	0.72	0.63	0.62	0.78	0.74	0.78
Stanford Sports	Stanford Sports	0.79	0.85	0.83	0.50	0.78	0.38
Stanford Residential	Stanford Residential	0.84	0.25	0	0	0	0

classifications. We first divide the test images into a map of 50x50 regions (tiles) according to their geographic locations. The trained SVMs are then used to label each of the test images. As a result, each tile is represented by three ratios of images being classified as the three respective classes. We use the label of the highest ratio to assign a land-use label to each tile label. Figures 3 and 4 show the classification maps compared to the ground truth maps of each campus. Since there are not enough test images to generate a map for each campus, we use the cross-campus classifiers to classify the entire image set of each campus. We note that most of the Academic class regions are correctly identified due to the strong performance of this classifier. The Stanford Sports classifier is also able to locate a significant amount of the Sports class regions correctly on the Berkeley campus. On the other hand, we can see that many labels of the Residential class are missing due to the failure of this classifier.

6. DISCUSSION

The work in this paper represents a proof-of-concept of land-use classification using geotagged images. Of course, land-use maps at least in the form of campus maps already exist for university campuses. We expect our approach to generalize to other areas for which land-use maps are not available. However, the ground truth for these regions will be more difficult to derive which is part of the reason we focused on university campuses here.

One issue with geotagged images is the accuracy of the location information. It is likely that some of the images will be incorrectly located due to placement error by the user, environmental constraints on the onboard GPS sensors of (phone) cameras, etc. Nevertheless, our results here demonstrate the feasibility of our approach even with potential location errors. The size of the dataset mitigates the noisy locations. Presumably, though, more accurate location information would improve the results. Improving the accuracy of the dataset as whole, though, is an interesting problem in itself.

7. FUTURE WORK

This paper represents our initial work on this interesting but challenging problem. There are many interesting directions in which it can be extended. Using high-level object and concept detectors instead of the low-level features used here will likely improve the classification performance. The problem of object and concept detection for geographic inference provides an interesting context in which to apply and evaluate existing techniques as well as develop novel ones.

It is likely that the visual and textual features are complementary. Therefore, it would be worthwhile to investigate combining these two modalities either at the feature or post-classification stage.

Finally, it would interesting to see if the approach can be used to classify subclasses of the three classes studied here. Subclasses of the Academic class such as classrooms, libraries, and laboratories, or subclasses of the Sports class such as stadiums, gymnasiums, and swimming pools, likely have their own distinctive visual and textual signatures in the images and thus could be distinguished given enough data.

8. CONCLUSION

In this paper, we proposed a novel framework to perform land-use classification using georeferenced images obtained from Flickr. We considered classification using visual and textual features, and at both the individual and group image level. The results from applying the approach to two university campuses represent promising first steps on this interesting but challenging problem.

9. ACKNOWLEDGEMENTS

This work was funded in part by an NSF CAREER grant (IIS-1150115) and a Department of Energy Early Career Scientist and Engineer/PECASE award.

10. REFERENCES

[1] Flickr photo sharing. http://www.flickr.com.

[2] Picasa web albums. http://picasa.google.com/.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[4] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the International World Wide Web Conference*, pages 761–770, 2009.

[5] M. Cristani, A. Perina, U. Castellani, and V. Murino. Geo-located image analysis using latent representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[6] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, 1999.

[7] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001.

[8] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. Geolocating static cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–6, 2007.

[9] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of ACM International Conference on Image and Video Retrieval*, 2007.

[10] D. Leung and S. Newsam. Proximate sensing using georeferenced community contributed photo collections. In *Procedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems: Workshop on Location Based Social Networks*, pages 57–64, 2009.

[11] D. Leung and S. Newsam. Proximate sensing: Inferring what-is-where from georeferenced photo collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2955–2962, 2010.

[12] K. Yanai, K. Yaegashi, and B. Qiu. Detecting cultural differences using consumer-generated geotagged photos. In *Proceedings of the International Workshop on Location and the Web*, 2009.