Proximate Sensing: Inferring What-Is-Where From Georeferenced Photo Collections

Daniel Leung and Shawn Newsam
Electrical Engineering & Computer Science
University of California at Merced
cleung3,snewsam@ucmerced.edu

Abstract

The primary and novel contribution of this work is the conjecture that large collections of georeferenced photo collections can be used to derive maps of what-is-where on the surface of the earth. We investigate the application of what we term "proximate sensing" to the problem of land cover classification for a large geographic region. We show that our approach is able to achieve almost 75% classification accuracy in a binary land cover labelling problem using images from a photo sharing site in a completely automated fashion. We also investigate 1) how existing geographic knowledge can be used to provide labelled training data in a weakly-supervised manner; 2) the effect of the photographer's intent when he or she captures the photograph; and 3) a method for filtering out non-informative images.

Figure 1. We conjecture that the visual content of georeferenced images can be used to derive maps of what-is-where on the surface of the earth.

1. Introduction

The growing availability of volunteered geographic information (VGI) is having a profound effect on geographical information systems and science. A range of new applications are being enabled by the georeferenced information contained in "repositories" such as blogs, wikis, social networking portals such as Facebook or MySpace, and, relevant to this work, community contributed photo collections such as Flickr [1]. The advantages of VGI are its temporal coverage which is often better both in terms of frequency and latency than traditional sources, and its size. The disadvantages are of course the quality of the data since it is usually made available without review and without accompanying provenance information.

Georeferenced photo collections in particular are enabling a new form of observational inquiry which we term "proximate sensing." If the traditional field of remote sensing is considered as using overhead images of distant scenes to derive geographic information then proximate sensing is using ground-level images of close-by objects and scenes. *The primary and novel contribution of this work is the conjecture that large collections of georeferenced photo collections can be used to derive maps of what-is-where on the surface of the earth.* Take, for example, the four photographs referenced on the map in figure 1. The content of these images provides rich geographic information about the locations at which they were taken.

This work investigates the application of proximate sensing to the problem of land cover classification. Rather than use overhead images to determine the distribution of land cover classes for a region, we investigate whether ground-level images can be used. We use this framework to investigate several interesting sub-questions such as 1) can existing geographic knowledge be used to provide labelled training data in a weakly-supervised manner? 2) what is the effect of the photographer's intent when he or she captures the photograph? and 3) does it help to filter out non-informative images? We quantitatively evaluate our results by comparing them to ground truth data from a land cover

978-1-4244-6985-7/10/$26.00 ©2010 IEEE

dataset. We show that our approach is able to achieve almost 75% classification accuracy in a binary land cover labelling problem using images from a photo sharing site in a completely automated fashion.

2. Related Work

While large collections of georeferenced photo collections have only recently become available through the emergence of photo sharing websites such as Flickr and Panoramio [3] and low-cost GPS technology, researchers have already investigated how these collections can help a number of image annotation and management tasks. A fundamental difference, however, between this previous work and ours is that the objective of this previous work has been to use location to infer something about one or more images, while the objective of our work is to use the images to infer something about a location.

In [10], Quack et al. describe a method for annotating photos whose location is only approximately known. A reference collection of georeferenced images is constructed by clustering a large number of Flickr images of a city scale region using both image features and textual annotation. Clusters are labelled as depicting events or objects (landmarks) using decision tree classifiers. Key phrases are determined for object clusters using frequent itemset mining. The phrases then provide links to Wikipedia articles. Novel images are annotated by identifying the visually similar clusters in the reference dataset and assigning the locations and Wikipedia links of the depicted objects.

In [8], Moxley et al. propose a system termed SpiritTagger which provides textual annotation for images whose geolocation is known. A novel image is annotated by transferring the tags of images from a large collection of georeferenced images (from Flickr) that are within a certain radius of the novel image's location. Tags from the top N matches based on visual similarity are assigned based on the ratio of their local to global (worldwide) frequency of occurrence– i.e. tags that are "unique" to the location should be weighed more heavily.

In [5], Hays and Efros tackle the challenging problem of estimating the unconstrained location of an image based solely on its visual characteristics. A reference dataset is constructed from over six million georeferenced Flickr images with geographic keywords. Novel images are geolocated by nearest neighbor indexing into the reference dataset using visual features. Once an image has been geolocated, a range of secondary annotation tasks are enabled by propagating mapped information, such as population density or elevation gradient, to the image. These labels can be used to organize photo collections.

In [4], Crandall et al. describe a method for organizing global-scale collections of georeferenced images. A large collection of Flickr images are spatially clustered at the landmark and metropolitan scales. The spatial clusters are used to 1) predict the location of a novel image based on its visual, textual, and temporal characteristics, and 2) provide interesting information on what people consider to be the most significant landmarks both in the world and within specific cities; which cities are most photographed; which cities have the highest and lowest proportions of attention-drawing landmarks; which views of these landmarks are the most characteristic; and how people move through cities and regions as they visit different locations within them.

While this last work by Crandall et al. does use a georeferenced image collection to derive information about landmarks and cities particularly with respect to how they are photographed, our work is fundamentally different from the above works in that we use the image collection to infer something about what-is-where on the surface of the earth. To the best of our knowledge, ours is the first work to do this.

3. The Problem

The primary and novel contribution of this work is the conjecture that large collections of georeferenced photo collections can be used to derive maps of what-is-where on the surface of the earth. We see this as an exciting new application of computer vision and image understanding that not only provides another context in which to investigate standard technical problems such as face recognition or indoor versus outdoor scene classification, but stands to motivate novel problems. Part of the excitement and allure of this problem is that it is not clear at the moment what kinds of things are geographically interesting in the image collections. There is significant opportunity for investigating whether it is informative and possible to infer things such as seasonal snowfall or flood severity from georeferenced (and temporally annotated) photo collections.

This paper presents our initial work which focuses on the task of using georeferenced image collections to perform land cover classification, a problem for which there is ground-truth for performing evaluation. Specifically, we investigate whether visual feature based classification of individual georeferenced images can be used to assign land cover labels to geographic regions to create a map.

This well-posed land cover classification problem also provides a framework in which to investigate a number of interesting broader questions such as 1) can existing geographic knowledge be used to provide labelled training data in a weakly-supervised manner? 2) what is the effect of the photographer's intent when he or she captures the photograph? and 3) does it help to filter out non-informative images? The specific approaches taken to answer these questions are described in the Experiments section below. First, we describe our dataset.

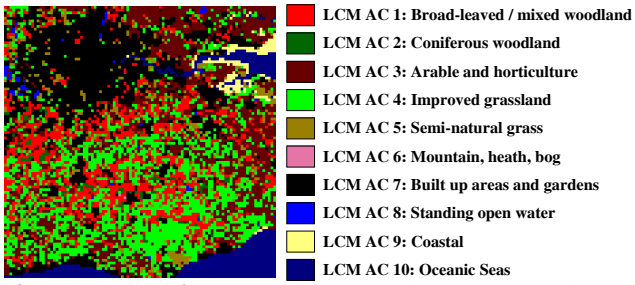


Figure 2. The dominant Land Cover Map 2000 Aggregate Classes (AC) for the TQ study area. This area measures 100x100 km and encompasses the London metropolitan area which appears towards the north-west.

4. Dataset

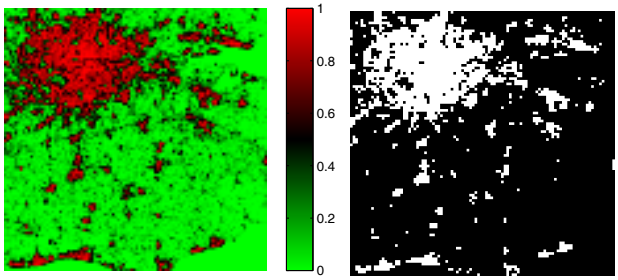
Our study area is the 100x100 km of Great Britain corresponding to the TQ square in the British national grid system. This region encompasses the London metropolitan area and thus includes a range of developed and undeveloped land cover classes.

We used the publicly accessible Countryside Information System (CIS) to download the Land Cover Map 2000 (LCM2000) of the United Kingdom's Centre for Ecology & Hydrology for the TQ study region. We focus on the LCM2000 Aggregate Class (AC) data which provides the percentage of ten land cover classes at the 1x1 km scale. Figure 2 shows the dominant classes for the TQ region.

We focus on binary classification into developed and undeveloped regions so we further aggregate the ten land cover classes into a developed superclass consisting of LCM AC:7 Built up areas and gardens, and an undeveloped superclass consisting of the remaining nine classes. We derive two ground truth datasets, one which indicates the percent developed for each 1x1 km tile in the TQ region and another which simply indicates a binary label for each tile by applying a 50% threshold to the percent developed. We refer to the first of these as the ground truth *fraction map* and the second as the ground truth *binary classification map*. Figure 3 shows the two ground truth maps.

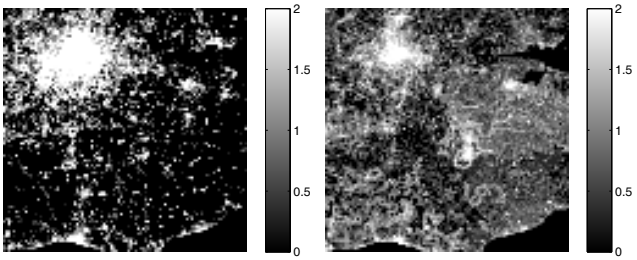
We compiled two georeferenced image collections for the TQ study area. First, we used the Flickr application programming interface (API) to download approximately 920,000 Flickr images located within the TQ region. Using the longitude and latitude information provided by the Flickr API, we were further able to assign each image to a 1x1 km tile. Figure 4 shows the distribution of the Flickr images. While Flickr contains a large collection of georeferenced images (over 90 million at time of writing), its spatial coverage is not uniform. For our study area, 5,420 of the 10,000 1x1 km tiles do not contain any Flickr images. The 4,580 tiles with images contain an average of 200.7, a median of 10, and a maximum of 53,840 images.

We contend that Flickr images represent "noisy" VGI



(a) (b)

Figure 3. Ground truth data derived from the LCM 2000 AC data. (a) Fraction map indicating the percent developed for each 1x1 km tile. (b) Binary classification map indicating the tiles labelled as developed (white) or undeveloped (black).



(a) (b)

Figure 4. The distribution of images for the TQ study region in the (a) Flickr and (b) Geograph datasets. On a base-10 logarithmic scale.

since the intentions of the photographers vary significantly and do not necessarily result in images that support geographic interpretation. Simply put, many Flickr images are not geographically informative. Our second dataset differs in this respect as it is derived from the Geograph British Isles (GBI) project [2] which aims to "collect geographically representative photographs and information for every square kilometre of Great Britain and Ireland." We consider this collection as a less noisy or purer example of VGI since the intent of the photographers who contribute to this collection is more likely to result in geographically informative images. We use the GBI API to download approximately 120,000 Geograph images for the TQ study area. While there are fewer Geograph images, they are more uniformly distributed than the Flickr images as shown in figure 4(b). Now, only 614 of the 10,000 1x1 km tiles do not contain any Geograph images and all but a few of these tiles correspond to ocean. The remaining 9,386 tiles contain an average of 12.9, a median of 5, and a maximum of 1,458 images.

Figure 7 shows sample images from the Flickr and Geograph datasets. Pairs of even/odd rows show Flickr/Geograph images for the same 1x1 km tiles. The top two pairs of rows are for tiles with a developed fraction of 1.0 while the bottom two pairs of rows are for tiles with a developed fraction of 0. These images indicate that

1) there is geographically relevant information contained in the two datasets thus supporting our conjecture that large collections of georeferenced photo collections can be used to derive maps of what-is-where on the surface of the earth; and 2) that the Geograph images are indeed generally more geographically representative.

5. Experiments

The goal in all experiments is to investigate how well the visual feature based classification of individual georeferenced images can be used to create developed/undeveloped land cover maps similar to the ground truth maps. We constrain the labels of the images to the same developed and undeveloped superclasses–that is each image is labelled as depicting a developed or undeveloped scene. The (developed) fraction assigned to a 1x1 km tile is then simply the ratio of the images with the label developed to the total number of images in the tile. Different approaches for labelling the images are compared based on how well the image generated fraction maps match the ground truth fraction map. Different quantitative measures of similarity are considered. We compute the correlation coefficient between the tile fraction values taken as observations of random variables. Specifically, if random variables X and Y represent the ground truth and image generated tile fraction values then the correlation coefficient for the image generated fraction map is computed as $\rho_{XY} = cov(X,Y)/\sigma_X\sigma_Y$ where $cov(X,Y)$ is the covariance of X and Y and σ_X (σ_Y) is the standard deviation of X (Y). ρ_{XY} ranges from -1 to 1 with a value of 0 indicating no correlation, and values of -1 and 1 indicating strong negative and positive correlation respectively. We also compute the mean absolute difference (MAD) and the root mean squared difference (RMSD) between the ground truth and image generated tile fraction values.

The binary label (developed or undeveloped) assigned to a 1x1 km tile is determined by applying a threshold to the image generated fraction for that tile. The similarity between the ground truth and an image generated binary classification map is measured in two ways. First, the overall classification rate is computed as the percentage of tiles with the same label. We also compute the average classification rate of the two classes (developed and undeveloped).

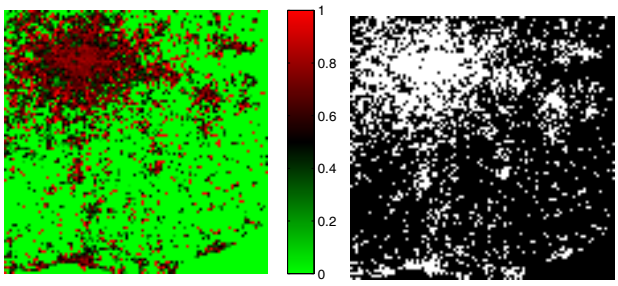
We deliberately choose simple features to characterize the visual content of the images. We annotate the georeferenced images using edge histogram descriptors which quantify the distribution of edges at different orientations. This is motivated by the observation that images of developed scenes typically have a higher proportion of horizontal and vertical edges than images of undeveloped scenes. This is evident in the sample images in figure 1. Following the method outlined in [7], we apply a set of five 2x2 linear filters to detect edges at roughly horizontal, vertical, 45° diagonal, 135° diagonal, and isotropic (non-orientation specific) directions. A threshold is applied to the outputs of these filters and the ratios of edges in various directions are summarized in a five bin L1 normalized histogram. In summary, each image is represented by a five dimensional edge histogram feature vector. Figure 1 indicates the edge histogram descriptors for the sample images.

We use a support vector machine (SVM) classifier to label individual images based on their edge histogram descriptors. Given a labelled training set, an SVM classifier with a Gaussian radial basis function kernel is trained using five fold cross validation and grid search for optimal parameter selection. Once trained, the classifier is used to label a set of target images which in all cases is disjoint from the training set. These labels are then used to generate fraction and binary classification maps which are compared with the ground truth maps. Even though the experiments below consider different training and target sets, the ground truth comparison is always based on the 4,553 tiles for which there are both Flickr and Geograph images. 38.9% of these tiles are developed in the ground truth so that the chance overall binary classification rate is 61.1% achievable by labelling all images and therefore all tiles an undeveloped.

Manually Labelled Training Set Here, the training set contains 2,740 Flickr images which have been manually labelled. A non-expert labelled an image as developed if it depicts a scene containing constructed materials such as used in houses, buildings, etc., and labelled it as undeveloped if it is of open areas and/or contains mostly trees and vegetation. These criteria will of course result in "incorrectly" labelled images such as indoor scenes always being labelled as "developed" even though they might have been taken inside isolated homes in rural regions. The SVM trained with the manually labelled training set is then used to classify a target set consisting of the remaining images from the 920K Flickr image set. The individual image labels are used to generate the fraction and binary classification maps shown in figure 5. Notice the similarity between these maps and the ground truth maps in figure 3.

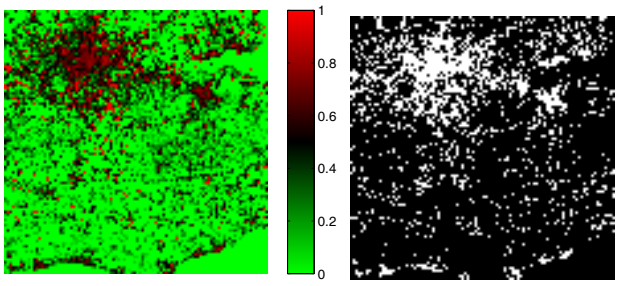
Line 1 of table 1 shows the quantitative similarity between the ground truth and image generated maps. The columns titled Fixed Threshold indicate the agreement between the binary classification maps when a fixed threshold of 0.5 is applied to the fraction values of the image generated fraction map to generate the binary classification map.

Prior Information The threshold used to derive the binary classification map can be adjusted so that the fraction of developed tiles matches that of the ground truth if such prior information is known. The columns titled Adaptive Threshold give the performance when the fraction of developed tiles in the image generated binary classification matches the 38.9% of the ground truth. For the manually labelled Flickr training set it results in decreased performance; subsequent experiments show that it can result in



(a) (b)

Figure 5. Land cover maps automatically generated using an SVM classifier trained with manually labelled Flickr images. The target set is also Flickr images. (a) Fraction map indicating the percent developed for each 1x1 km tile. (b) Binary classification map indicating the tiles labelled as developed (white) or undeveloped (black). Compare with the ground truth maps in figure 3



(a) (b)

Figure 6. Land cover maps automatically generated using an SVM classifier trained with a large set of Geograph images labelled in a weakly-supervised manner. The target set is also Geograph images. (a) Fraction map indicating the percent developed for each 1x1 km tile. (b) Binary classification map indicating the tiles labelled as developed (white) or undeveloped (black). Compare with the ground truth maps in figure 3

significant improvement.

Weakly-supervised Training This experiment investigates the performance of a classifier trained in a weakly-supervised manner. The training set is constructed without any manual labelling by selecting two images at random from each 1x1 km tile and labelling them with the majority class of the tile. Selection is limited to tiles with four or more images so that at least two images remain in the disjoint target set. For the Flickr dataset, this results in a training set termed "Flickr-small" containing 5,872 images. Line 2 of table 1 shows the results for the Flickr-small classifier when applied to the Flickr dataset. The performance is shown to be better than that of the classifier trained with the manually labelled dataset, an interesting and significant result indicating that training sets can be generated from regions for which maps exists and then used to train classifiers for mapping unmapped regions. That the results are better than the manual case suggests that the automatically generated training set more accurately characterizes the differences between images from developed and undeveloped regions than the intuition humans use when labelling images.

Photographer Intent This experiment investigates the effect photographer intent has on the image generated maps. The training and target sets are now selected from the Geograph dataset which contains images captured by photographers who intend their photographs to be geographically representative. We again train the classifier with a weakly-supervised dataset which now has 10,576 images as there are more tiles with four or more Geograph images than tiles with four or more Flickr images. Line 3 of table 1 shows the result of applying this Geograph-small classifier to a held out Geograph target set. The maps generated from the Geograph dataset are significantly better than those generated from the Flickr dataset indicating that photographer intent is a significant factor. We comment more on the implication of this result this in the Discussion section below.

Training Set Size This experiment investigates the effect of the size of the weakly-supervised training set. We construct a 11,465 Flickr-large training set by selecting five images from tiles that contain more than ten Flickr images. Line 4 of table 1 shows the results of using this training set. The performance is worse than that of the Flickr-small training set likely because of a bias present in the smaller number of tiles with more than ten images. These tiles are more likely to be of developed regions as confirmed by the higher ratio of images labelled developed in the training set (indicated in parenthesis in the column titled Training Set Size in table 1). A Geograph-large training set is also constructed and shown to perform better than the Geograph-small set (see line 5 of the table). Figure 6 shows the maps generated using a classifier trained with the Geograph-large training set. Tables 2 and 3 show the confusion matrices for the Geograph-large training set for the fixed and adaptive threshold cases respectively.

Training Set Quality This experiment investigates the effect of the quality of the weakly-supervised training set. We now select training images from tiles that have very high or very low developed fractions according to the ground truth map. The intuition here is that such tiles should result in more accurate training sets. Lines 6 and 7 of table 1 show that these Flickr-good and Geograph-good training sets do not result in improved performance. This finding indicates that it is not necessary to constrain the weakly-supervised training sets in this way.

Relative Importance of Training and Target Sets The results above clearly indicate that the Geograph dataset is more effective than the Flickr dataset. This experiment investigates whether this improvement is due to the training or target set. Lines 8 through 14 in table 1 list the results when the training and target sets are from different image collec-

tions. These results make it clear that photographer intent is more important for the target set than the training set. While this finding is somewhat unfortunate since the overall (worldwide) coverage of the Flickr dataset is broader than that of the Geograph dataset, it does identify some interesting research challenges which will be discussed in the Discussion section below.

Filtering Images With Faces This experiment investigates whether removing images with faces improves the results. The motivation here is that photographs of people are less likely to be geographically informative, especially close-in portraits. The fact that few of the Geograph images contain people empirically suggests this is true. We used a standard face detection algorithm [12, 6] to filter Flickr images containing one or more faces. We then repeated a set of experiments using this face-free target set. Unfortunately, as lines 15 through 17 in table 1 show, this did not provide any improvement over the target set with faces.

6. Discussion

The results from the experiments above substantiate our conjecture that large collections of georeferenced photo collections can be used to automatically derive maps of what-is-where on the surface of the earth. The image generated fraction and binary classification maps were shown to be similar to the ground truth LCM 2000 data. The results also provide initial answers to some of the broader questions that were posed.

We demonstrated that weakly-supervised labelled training data resulted in better performance than manually labelled training data. This has clear benefits in the amount of effort required to train the classifiers. An interesting open question remains as to how well a classifier trained with weakly-supervised data from one region generalizes to other regions. We plan to investigate this in future work.

Perhaps the most conclusive result from the experiments is the effect that photographer intent has on the performance. The maps generated using the Geograph target data were significantly more accurate than the maps generated using the Flickr target data. This confirmed our expectations. The fact that the images in the Geograph dataset are contributed by photographers whose intent is to geovisually annotate Great Britain and Ireland should result in images which are more effective at producing land cover maps. While this does restrict the types of target data which the proposed approach can be applied to, and thus restrict the geographic regions since there are fewer intent driven georeferenced photo collections, it does pose interesting research challenges. Specifically, is it possible to use a "clean" dataset such as the Geograph images to filter or process a "noisy" target dataset such as the Flickr images. We are beginning to investigate this problem.

Our analysis framework potentially allows us to investigate the subtler question of the effect of the charge given to photographers in visually documenting what-is-where. It is clear that asking photographers to capture geographically relevant photographs is effective for producing land cover maps but how about variants of this question. Is it better to ask them to photograph the most geographically relevant thing, for example.

Our investigation into whether removing images with faces from the target dataset improves performance was one attempt to improve the noisy Flickr dataset. We stipulated that images with faces were less likely to be geographically informative because the intent behind photographing people likely runs contrary with the intent to visually describe a landscape and because the people often simply obscure other parts of the scene. Unfortunately, filtering images with faces did not result in better performance. We plan to delve further into this null result investigating, for example, filtering images based on the relative size of the face–i.e., the larger the face, the more likely it is simply a portrait. We also plan other top-down filtering approaches such as using existing research on detecting whether sky, water or other geographic features are present in an image.

Clearly our edge histogram image features limit our performance. We will obviously explore other general image descriptors such as color and texture as well as more specialized descriptors such as Oliva and Torrabla's gist descriptor [9]. There is a large research corpus on visual image annotation which we can revisit in the context of this new problem.

Many of the Flickr images and all of the Geograph images have textual annotations. We plan to investigate how these can be integrated with the visual features to improve performance. These two datasets will also allow us to investigate the effect of the annotator's intent.

We plan to extend the number of superclasses we model beyond the current developed and undeveloped groupings.

Finally, we plan to investigate spatial models for improving our approach. Tobler's first law of geography states that all things are related, but nearby things are more related than distant things [11]. Clearly there are opportunities for modelling the spatial distribution of land cover classes with Markov random fields, for example, and using these models to improve the derived maps.

7. Acknowledgements

This work was funded in part by NSF grant IIS-0917069 and a Department of Energy Early Career Scientist and Engineer/PECASE award. The authors would like to thank Nathan Graves for implementing the edge histogram extraction.

Table 1. The experimental results. The number in parenthesis in the Training Set Size column indicates the fraction of images labelled as developed in the training set. Please see the text for other details.

				Binary Maps						
				Overall Class. Rate		Avg. Class. Rate		Fraction Maps		
	Training Set	Target Set	Training Set Size	Fixed Threshold %	Adaptive Threshold %	Fixed Threshold %	Adaptive Threshold %	ρ	MAD	RMSD
1	Manual (Flickr)	Flickr	2740 (0.51)	66.4	64.9	68.8	63.0	0.374	0.287	0.383
2	Flickr small	Flickr	5872 (0.52)	67.2	66.9	68.7	65.2	0.380	0.279	0.373
3	Geograph small	Geograph	10576 (0.26)	68.2	74.0	60.8	72.6	0.520	0.271	0.358
4	Flickr large	Flickr	11465 (0.56)	57.7	61.5	64.0	59.6	0.372	0.336	0.441
5	Geograph large	Geograph	13374 (0.36)	73.8	**74.7**	70.2	**73.2**	**0.552**	0.235	0.313
6	Flickr good	Flickr	5070 (0.49)	67.0	68.1	67.4	66.6	0.329	0.285	0.374
7	Geograph good	Geograph	5603 (0.47)	**74.2**	74.6	**71.5**	73.1	0.551	**0.231**	**0.308**
8	Geograph small	Flickr	10576 (0.26)	60.0	72.3	49.8	70.9	0.230	0.354	0.457
9	Geograph large	Flickr	13374 (0.36)	60.0	68.7	51.4	66.9	0.301	0.312	0.404
10	Geograph good	Flickr	5603 (0.47)	60.7	68.3	53.8	66.6	0.330	0.294	0.381
11	Manual (Flickr)	Geograph	2740 (0.51)	66.1	73.5	70.1	72.0	0.531	0.273	0.356
12	Flickr small	Geograph	5872 (0.52)	67.8	74.1	70.8	72.3	0.526	0.264	0.345
13	Flickr large	Geograph	11465 (0.56)	56.3	72.6	63.3	71.2	0.486	0.340	0.428
14	Flickr good	Geograph	5070 (0.49)	69.9	73.1	**71.5**	71.7	0.496	0.2545	0.3310
15	Flickr small	Flickr (no faces)	5872 (0.52)	66.8	66.7	66.8	64.2	0.367	0.301	0.414
16	Geograph small	Flickr (no faces)	10576 (0.26)	59.8	72.2	49.0	69.7	0.225	0.377	0.493
17	Geograph good	Flickr (no faces)	5603 (0.47)	59.9	68.0	52.0	65.2	0.312	0.321	0.428

Table 2. Confusion matrix for the Geograph-large training set for the fixed threshold case.

		Prediction	
		Dev.	Undev.
Ground Truth	Dev.	962 (54.3%)	810
	Undev.	384	2397 (86.2%)

Table 3. Confusion matrix for the Geograph-large training set for the adaptive threshold case.

		Prediction	
		Dev.	Undev.
Ground Truth	Dev.	1180 (66.6%)	592
	Undev.	560	2221 (79.9%)

References

[1] Flickr Photo Sharing. http://www.flickr.com.

[2] Geograph British Isles Project. http://www.geograph.org.uk.

[3] Panoramio Photo Sharing. http://www.panoramio.com.

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

[5] J. Hays and A. Efros. IM2GPS: estimating geographic information from a single image. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[6] R. Lienhart and J. Maydt. An extended set of Haar-like features for rapid object detection. In *Proceedings of the IEEE International Conference on Image Processing*, pages 900–903, 2002.

[7] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11:703–715, 1998.

[8] E. Moxley, J. Kleban, and B. S. Manjunath. SpiritTagger: a geo-aware tag suggestion tool mined from Flickr. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pages 24–30, 2008.

[9] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[10] T. Quack, B. Leibe, and L. Van Gool. World-scale mining of objects and events from community photo collections. In *Proceedings of the International Conference on Content-based Image and Video Retrieval*, pages 47–56, 2008.

[11] W. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2):234–240, 1970.

[12] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 511–518, 2001.

(a) Sample Flickr images from a 1x1 km tile with a developed fraction of 1.0.

(b) Sample Geograph images from the same tile as above.

(c) Sample Flickr images from a 1x1 km tile with a developed fraction of 1.0.

(d) Sample Geograph images from the same tile as above.

(e) Sample Flickr images from a 1x1 km tile with a developed fraction of 0.

(f) Sample Geograph images from the same tile as above.

(g) Sample Flickr images from a 1x1 km tile with a developed fraction of 0.

(h) Sample Geograph images from the same tile as above.

Figure 7. Sample images from the Flickr and Geograph datasets.