2014 22nd International Conference on Pattern Recognition

Semi-Supervised Learning of Geospatial Objects Through Multi-Modal Data Integration

Yi Yang and Shawn Newsam
Electrical Engineering and Computer Science
University of California, Merced, CA, 95343
Email: snewsam@ucmerced.edu

***Abstract*—We investigate how overhead imagery can be integrated with non-image geographic data to learn appearance models for geographic objects with minimal user supervision. While multi-modal data integration has been successfully applied in other domains, such as multimedia analysis, significant opportunity remains for similar treatment of geographic data due to location being a simple yet powerful key for associating varying data modalities, and the growing availability of data annotated with location information either explicitly or implicitly.**

We present a specific instantiation of the framework in which overhead imagery is combined with gazetteers to compensate for a recognized deficiency: most gazetteers are incomplete in that the same latitude/longitude point serves as the bounding coordinates of the spatial extent of the indexed objects. We use a hierarchical object appearance model to estimate the spatial extents of these known object instances. The estimated extents can then be used to revise the gazetteers.

A particularly novel contribution of our work is a *semi-supervised learning regime* which incorporates weakly labelled training data, in the form of incomplete gazetteer entries, to improve the learned models and thus the spatial extent estimation.

I. Introduction

The automated analysis of overhead imagery remains an open problem especially for complex geospatial objects. We here investigate ways in which multi-modal data integration can help with computer-based image understanding.

Multimodal data integration has been successfully applied to other information discovery problems particularly in multimedia analysis. Researchers have exploited connections between image and non-image data such as image annotations or video transcripts, to improve image understanding and other challenging tasks. We contend that there is significant opportunity for similar analysis of geographic data due to 1) location being simple yet powerful key for associating varying data modalities; and 2) the growing availability of data annotated with location information, either explicitly or implicitly.

We describe a general framework in which georeferenced overhead imagery is integrated with possibly inaccurate and/or incomplete geographic object instance data to 1) learn appearance models for the geographic objects with minimal user supervision; and, in turn, 2) use the learned appearance models to revise the instance data.

We focus on combining high-resolution aerial imagery with gazetteers. Gazetteers are geographic dictionaries of what-is-where on the surface of the Earth which specify, at a minimum, a type and a location for each record. Despite their extensive coverage, most, if not all gazetteers, are deficient in that *the spatial extents of the archived objects are limited to a single point*. As the development team of the University of California at Santa Barbara Alexandria Digital Library (ADL) gazetteer points out [1], "for a digital library application, the spatial extent of the feature, either approximately with a bounding box or more accurately with a polygonal representation, is better, but there are no large sets of gazetteer data with spatial extents." They go on to state that "establishing the standards that will enable the sharing of gazetteer data will help harvest data from many sources, but ultimately deriving spatial locations and extents from digital mapping products and other sources automatically will be needed."

Fig. 1. Gazetteers are deficient in that they specify the spatial extents of the indexed objects using a single point. We propose a semi-supervised learning framework that leverages these points, such as shown above for this golf course, to learn object appearance models. These models can, in turn, be used to estimate the true spatial extents of the objects and update the gazetteer.

The ADL entry for a golf course in Southern California is shown at the top of figure 1. Note the same latitude/longitude point is used for the two bounding coordinates. The bottom of the figure shows this point mapped on an aerial image.

A fundamental contribution of this paper is to do just as the ADL gazetteer development team proposes. We leverage readily available high-resolution overhead imagery to esti-

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.696

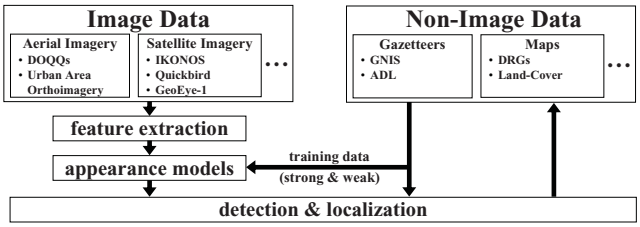


Fig. 2. The proposed framework in which georeferenced overhead imagery is integrated with possibly inaccurate and/or incomplete geospatial object instance data to 1) learn appearance models for the object with minimal user supervision; and, in turn, 2) use the learned appearance models to revise the non-image data to make it more accurate and/or complete.

mate the spatial extents of known object instances to revise the gazetteers as indicated by the upwards pointing arrow in figure 1. *The other fundamental contribution is that we treat the incomplete gazetteer entries as weakly-supervised training data for learning object appearance models in a semi-supervised manner.* This is shown by the downward pointing arrow, and represents a novel way in which to integrate these two geographic sources not proposed by the ADL gazetteer development team nor, to our knowledge, by other researchers.

II. Related Work

Researchers in multimedia analysis have combined image and non-image data to improve image understanding. Computer vision researchers have exploited various forms of metadata associated with image collections to learn visual object models. Berg et al. [2] data mine a large collection of captioned images of faces from online news sources to train a recognition system for commonly occurring people. Barnard et al. [3] develop an object recognizer using 10,000 images of works of art along with associated free text which varies greatly from physical description to interpretation and mood. And, Li et al. [4] turn the search paradigm around by using search results from the Google image search engine to learn visual models for a variety of object categories.

Researchers working on geographic information systems have likewise proposed a number of ways to leverage non-image data sources to improve overhead image understanding. Road extraction has been improved by using existing vectorized road networks as seeds [5], [6], [7] and by using digital surface models to account for gaps between road segments due to shadows [8]. Agouris et al. [9] propose a SpatioTemporal Gazetteer that incorporates aerial imagery as well as existing vector datasets of extracted outlines and thematic datasets (building blueprints, building usage records) to automatically detect changes to the spatial footprints of buildings using template matching.

Our work differs from previous work on integrating different modalities of geographic data in the following aspects:

- We model more complex object types than the previous approaches.
- We use a hierarchical object appearance model that has a latent land use/land class level.
- We incorporate weakly labelled training data in a semi-supervised learning framework. We show this semi-supervised learning framework improves upon a fully-supervised *particularly when very little labelled training data is available*.

This paper builds upon our earlier work on this problem. In [10], we showed that the distributions of quantized local features extracted from image regions centered on the gazetteer point locations were more similar for intra-class object instances than they were for inter-class instances. This provided initial indication that the gazetteer entries could be used as weakly labelled training data. That work, however, did not propose any appearance models, did not propose how those models would be learned in a semi-supervised manner, nor did it use ground truth spatial extents for evaluation. In [11], we developed the hierarchical model used in this work. The learning in that paper is completely supervised, however, using only manually labelled examples, and thus does not exploit gazetteers as a source of weakly labelled training data. The work presented in this paper extends that work to incorporate weakly labelled training data in a semi-supervised learning framework. This is a significant development which allows gazetteers and other non-image data sources to be integrated with image data to advance automated image understanding.

III. Framework Overview

Figure 2 presents an overview of the proposed framework for integrating different modalities of geographic data. The sources are separated into image and non-image data. Image sources include aerial imagery such as Digital Orthophoto Quarter Quads (DOQQs) and urban aerial orthoimagery which is freely available at the USGS National Map, and commercial satellite imagery from the IKONOS, Quickbird, GeoEye-1, and other space-borne sensors which is available for cost. The non-image data includes gazetteers, maps, such as digital raster graphics (DRGs), and other repositories which provide type-location tuples. The overarching goal is to use the imagery to update the non-image records. This includes detecting novel object instances as well as localizing known instances (estimating the spatial extent or correcting for general location errors). This is accomplished through feature extraction, object appearance modelling, and detection and localization modules. We describe below the features, object models, and localization procedure used to estimate the spatial extents of known gazetteer entries which is the focus of this paper. The non-image data provides both strongly and weakly labelled training data. The *strongly labelled* data is in the form of accurate and complete records, or, as in the case of this paper, manually labelled data. The *weakly labelled* data is the incomplete and possibly inaccurate records.

The rest of this section describes the data sources used to estimate the spatial extents of known gazetteer entries.

A. Data Sources - Gazetteers

We utilize two gazetteers in this work. First is GeoNames[1], an online world-wide gazetteer compiled from several dozen sources including other gazetteers. Queries to GeoNames return a single point as the spatial extent of an object.

[1] http://www.geonames.org

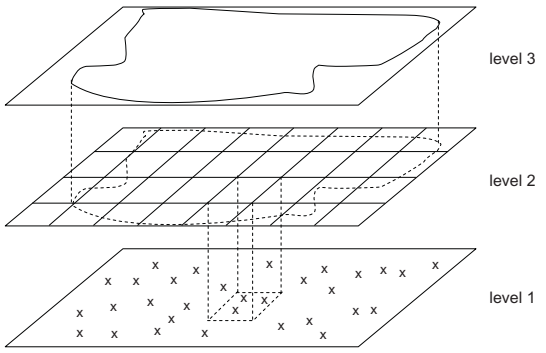


Fig. 3. The three levels of our hierarchical model. Level 1 represents the object using quantized SIFT features shown here as x's. BOVW histograms are computed for image tiles and SVM classifiers are used to assign LULC labels to the tiles in level 2. The distribution of the LULC classes in level 3 constitutes the final object model.

We also treat Google Maps as a gazetteer in that it allows us to perform location-based searches for geospatial objects such as Costco shopping centers. We further use the Google Maps Geocoding API[2] to translate the street addresses provided by Google Maps into latitude/longitude points.

B. Data Sources - Image Repositories

We use the USGS National Map Seamless Data Server[3] interface to automatically download high-resolution overhead imagery. Images are retrieved from the National Map using a simple rectangular query region specified by its bounding latitude and longitude values. In our case, the single latitude/longitude point from the gazetteer serves as the center of a region whose size is chosen to ensure that the retrieved image should contain the target object. This size is chosen empirically in the experiments below based on the observed sizes of sample objects. A single size is picked for each object type and then fixed for all the retrievals. Note that the gazetteer point does not always fall inside the object due to data collection, geo-registration, or other errors.

IV. Hierarchical Object Model

This section describes the three levels of the hierarchical appearance model we use to estimate the spatial extent of known gazetteer entries. See figure 3 for an illustration of the model.

A. Level 1 - Local Invariant Features

We use local invariant features to characterize the objects at the lowest level of the hierarchy. These features are designed to be robust to image variations caused by geometric image transformations, such as scaling and rotation, as well as to photometric distortions caused by variation in illumination, etc. They have proven to be effective for a range of computer vision applications over the last decade.

We choose David Lowe's Scale Invariant Feature Transform (SIFT) [12] as our local invariant feature detector and descriptor. The SIFT detector, like most local feature detectors, results in a large number of feature points. This density is important for robustness but presents a representation challenge particularly since the SIFT descriptors have 128 dimensions. We adopt a standard bag-of-visual-words (BOVW) [13] approach to summarize the descriptors by quantizing and aggregating the features without regard to their location. We first construct a visual dictionary by performing k-means clustering on a large number of SIFT features (from a dataset different from that used to train the object models). This dictionary is then used to quantize the individual SIFT points into "visual words" by simply assigning the label of the closest cluster centroid. We aggregate the quantized features at the image tile level using a BOVW histogram

$$BOVW = [t_1, t_2, \ldots, t_V] \ ,$$

where t_v is the number of occurrences of visual word v in a tile and V is the dictionary size. The BOVW histogram is normalized to have unit L1 norm to account for the difference in the number of interest points between tiles.

We use 256x256 pixel tiles in all the experiments below.

B. Level 2 - Latent LULC Classes

An intermediate, latent level bridges the gap between the low-level local invariant features and the high-level objects. Specifically, land use/land class (LULC) labels are assigned to image tiles using support vector machines (SVMs).

We leverage our recent work [14] on LULC classification. In that work, we used a large ground truth dataset to train SVM classifiers for a number of LULC classes. We use the probabilistic output option of the LIBSVM package [15] to compute, for each tile i in an image, the probability distribution over the M LULC classes as

$$P(tile_i) = [p_1, p_2, ..., p_M] \ ,$$

where p_m corresponds to the probability that tile i is assigned to the mth class by the SVM classifiers. The SVM classifiers take as input the BOVW histograms from level 1. We normalize $P(tile_i)$ so that $\sum^M p_m = 1$.

In order to reduce the effect of tile (mis)alignment, we perform the LULC labeling on tiles which overlap by 50 percent. Thus, each 128x128 pixel *block* appears in four 256x256 pixel tiles. We apply a smoothing mechanism to the LULC class distribution at the block level

$$P(block_j) = \frac{1}{4} \sum P(tile_i) \ , \qquad (1)$$

where the sum is taken over the four tiles in which block j appears.

To summarize, our final representation at level 2 in the hierarchy is a probability distribution $P(block_j)$ over M LULC classes for each 128x128 pixel block j.

C. Level 3 - Object Model

The top level of our representation also models the objects as probability distributions over LULC classes. For an object

[2] http://code.google.com/apis/maps/documentation/geocoding
[3] http://seamless.usgs.gov

region encompassing a set of $\mathbb{U}$ blocks labelled at level 2, we compute

$$P(object) = \frac{1}{|\mathbb{U}|} \sum_{block_j \in \mathbb{U}} P(block_j) \ , \tag{2}$$

where $P(block_j)$ is computed using equation 1 and $|\mathbb{U}|$ is the cardinality of $\mathbb{U}$.

V. Spatial Extent Estimation Via Relevancy

We use a relevance function to determine whether image tiles near a known object instance are actually part of the object or not. This relevance function is based on the top level of our hierarchical object model, specifically the LULC class probability distribution. We compare two techniques for learning this relevance function: fully-supervised, in which only strongly labelled examples are used, and semi-supervised, in which both strongly and weakly labelled examples are used.

We use a recent bipartite ranking function to incorporate unlabelled data in the learning phase [16]. The goal in bipartite ranking is to learn a scoring function H which assigns higher scores to relevant instances than to irrelevant ones. In our problem, we consider regions belonging to the object of interest as being relevant, and regions belonging to the background as being irrelevant. Our goal is to mark as relevant those regions within the true spatial extent.

In the fully-supervised case, the function is learned using a set of examples which are either labelled as relevant or irrelevant. In the semi-supervised case, the training set includes not only the labelled data but also unlabelled data. The unlabelled data is typically incorporated into the learning by assuming that unlabelled examples which are similar (in feature space) to labelled ones should have a similar (relevancy) label. We use the bipartite ranking function of Amini et al. [16] which provides robustness to error-prone propagation of labels to unlabelled training data by minimizing the ranking errors on the labelled and unlabelled training sets separately.

Once a relevance function has been learned, we estimate the spatial extent of a known object instance in a target image as follows. First, we extract and quantize SIFT features from the target image. We then compute the BOVW histograms for overlapping 256x256 pixel tiles and the multi-class SVM classifiers are used to compute the LULC class distributions for each tile. The LULC class distributions are then computed for each 128x128 pixel block using equation 1.

We slide a square window of size wxw blocks over the image in increments of one block. For each window location, we compute the probability distribution of the window over the LULC classes:

$$P(window) = \frac{1}{w^2} \sum_{block_j \in window} P(block_j) \ , \tag{3}$$

where $P(block_j)$ is computed using equation 1. We use the relevance function to rank all the windows in the target image in order of decreasing relevancy. The input to the relevance function for a window is the probability distribution over the LULC classes as computed in equation 3 and the output is a relevancy score. The estimated spatial extent is then the union of all blocks in all windows whose relevancy is above a relevancy threshold θ. We discuss the setting of θ below.

VI. Experimental Results

A. Dataset

The GeoNames gazetteer is used to identify 44 high schools, 27 golf courses, and 23 mobile home parks, and Google Maps is used to identify 18 Costco shopping centers. The National Map Seamless Data Server is then used to download 1-foot resolution orthoimagery using a large query region to ensure the images contain the target objects.

A ground truth dataset is created by manually delineating the target objects using a polygon representation. This labeling was done by undergraduates in our lab with no knowledge of the proposed approach.

We use the hierarchical object model to represent the ground truth objects as follows. SIFT features are extracted from each of the images and quantized using a visual dictionary consisting of 100 visual words. In previous work [17], we showed that a dictionary of this size represents a good balance between efficiency and accuracy. A BOVW histogram is computed for overlapping 256x256 pixel tiles.

Tile-level LULC distributions are computed using a set of SVMs corresponding to 18 LULC classes: agricultural, airplane, baseball diamond, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, runway, sparse residential, and tennis courts. Finally, block-level LULC distributions are computed using equation 1 and object-level distributions are computed using equation 2.

B. Evaluation

We use a retrieval/detection paradigm to evaluate how well a learned model estimates the spatial extent of a known object instance. That is, instead of selecting a particular value for the cutoff threshold θ, we vary this parameter and compute precision and recall values.

Given a target image with ground truth spatial extent L_{true}, and estimated spatial extent L_{est} corresponding to a specific setting of the relevancy threshold θ, we compute precision as the fraction of the estimated region that actually belongs to the true spatial extent:

$$precision = \frac{|L_{est} \bigcap L_{true}|}{|L_{est}|} \tag{4}$$

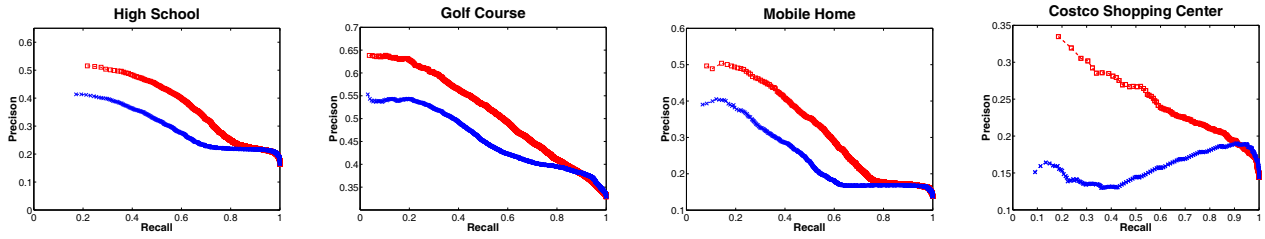
We compute recall as the fraction of the true spatial extent that appears in the estimated region:

$$recall = \frac{|L_{est} \bigcap L_{true}|}{|L_{true}|} \ , \tag{5}$$

In these equations, $|\cdot|$ indicates the area of a region in pixels and $\bigcap$ indicates set intersection. As usual, precision and recall range from 0 to 1.

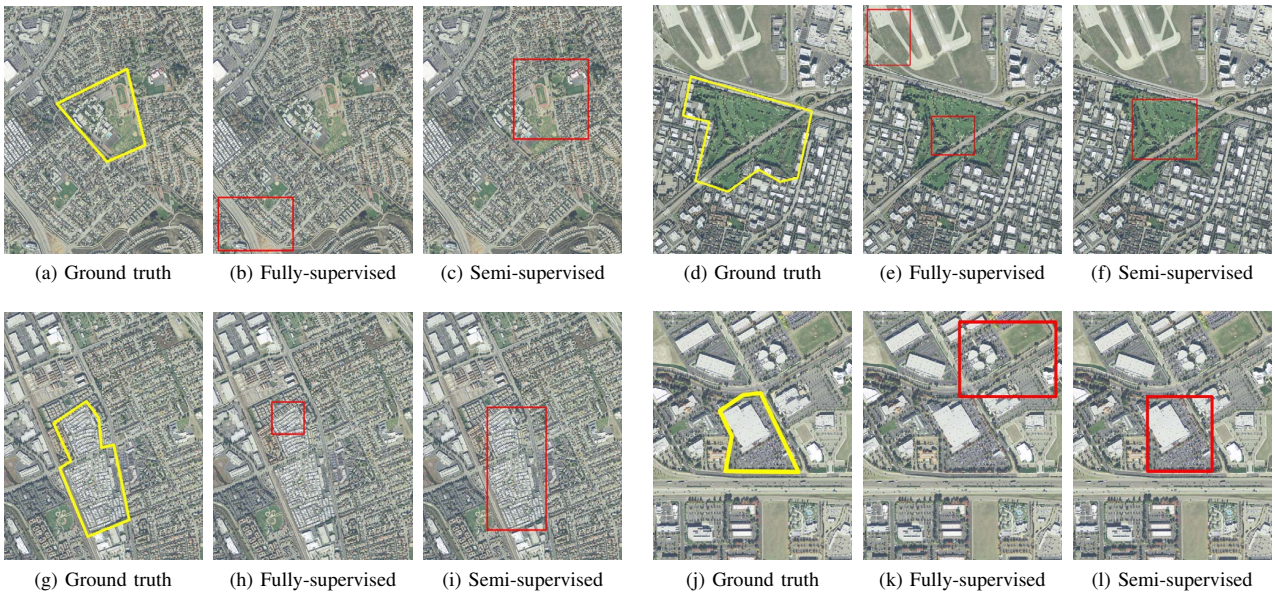
C. Experiments

We compare two different training regimes. First, the fully-supervised case where the ranking function is learned using strongly labelled training data; i.e., images in which the spatial extent of an object has been manually delineated.



(a) (b) (c) (d)

Fig. 4. Precision-recall curves for the four object types. The results for the fully-supervised learning regime are shown using blue x's. The results for the semi-supervised regime are shown using red squares.



(a) Ground truth (b) Fully-supervised (c) Semi-supervised (d) Ground truth (e) Fully-supervised (f) Semi-supervised

(g) Ground truth (h) Fully-supervised (i) Semi-supervised (j) Ground truth (k) Fully-supervised (l) Semi-supervised

Fig. 5. Results for instances of the high school (a-c), golf course (d-f), mobile home park (g-i), and Costco (j-l) classes. In each triple, the left image is the manually delineated ground truth, the middle image is the bounding box(es) as detected using the fully-supervised approach, and the right image is the bounding box as detected using the proposed semi-supervised approach.

The second case is the semi-supervised case where the ranking function is learned using a combination of strongly and weakly labelled training data. The weakly labelled data are images in which the object has not been delineated. The significance here is that such weakly labelled training data can be automatically generated using existing gazetteers which represent the spatial extent using only a single point. This point can be used to retrieve imagery from the National Map or other image repository which should contain the object roughly centered. The query region is chosen to be larger than the typical size of the particular object type.

To show the improvement provided by the weakly labelled data, the strongly labelled data consists of *only one manually labelled image* in both learning regimes. The LULC distribution of the ground truth region as computed using equation 2 is the single relevant example. The LULC distributions over windows outside the object region are the irrelevant examples.

The unlabelled examples in the semi-supervised learning regime are the LULC distributions over windows from a set of weakly labelled images. We equally weight the labelled and unlabelled data in the learning as described in [16].

We evaluate performance using cross-validation. Each image in the ground truth dataset is taken separately as the labelled image. The rest of the images are separated equally into unlabelled training data and test data. Both learning regimes see the single labelled image. The semi-supervised regime also sees the unlabelled training data. The ranking function is then applied to each of the test images separately. For each test image, a set of precision-recall values are computed as the relevancy threshold, θ, is varied. A set of precision-recall values is computed by averaging over all test images. The final set of precision-recall values is computed by averaging over all trials in the cross-validation (one for each image in the ground truth dataset). We also compute an average precision (AP) from this final set.

D. Results

Our results clearly show that incorporating the weakly labelled training samples provided by the gazetteers improves the object appearance models, and thus the spatial extent estimation. Figure 4 shows the precision-recall curves for the four object types. The results for the fully-supervised learning regime are shown using blue x's. The results for the semi-

TABLE I. AVERAGE PRECISION VALUES FOR THE TWO LEARNING REGIMES.

Learning	HS	GC	MHP	Costco
Fully-Supervised	0.316	0.460	0.260	0.190
Semi-Supervised	0.401	0.518	0.340	0.260

supervised regime are shown using red squares. The proposed semi-supervised regime results in higher precision at almost all values of recall, the exceptions being at very high recall values where there is not much difference.

The average precision for the four object types are listed in table I. These results again demonstrate that the semi-supervised regime improves over the fully-supervised one.

Figure 5 shows the results for instances of each of the classes. The left panels indicate the manually delineated ground truth, the middle panels show the results of the fully-supervised case, and the right panels show the results of the semi-supervised case. The rectangles in the middle and right panels are the bounding boxes of the windows detected for an empirically chosen relevancy threshold. (The same threshold value is used for the fully- and semi-supervised cases.) Clearly, the bounding boxes computed using the proposed semi-supervised approach more accurately depict the spatial extents of the objects than those computed using the fully-supervised approach.

E. Discussion

The results above are notable given that *only a single manually labelled training image* is used to learn the ranking functions. Even in the fully supervised case in which this single image is the only training data, we achieve a recall rate of 0.5 while the precision is still over 0.3 for three of the four object types. That is, we are able to estimate more than half the true spatial extent while keeping the estimated region reasonably sized. While not perfect, this level of accuracy for the spatial extent is a big improvement over the single latitude/longitude point currently present in gazetteers. And, these precision-recall values would be even higher if we used a bounding box as the ground truth (as opposed to the manually delineated polygon).

The main contribution of this paper, however, is the semi-supervised framework for incorporating weakly labelled training data. The results above clearly demonstrate this improves upon the standard fully supervised framework. This is significant since there are a growing number of sources from which to obtain such weakly labelled data. We use a standard gazetteer and Google Maps in this paper but one could easily imagine expanding this other sources such as digital raster graphics maps or even georeferenced social media.

VII. CONCLUSION

We described a general framework in which georeferenced overhead imagery is integrated with possibly inaccurate and/or incomplete non-image geographic data to learn appearance models for geographic objects with minimal user supervision.

We demonstrated a particular instantiation of this framework which integrates imagery with gazetteers to improve the spatial extents of the gazetteer records. We are motivated by the recognized deficiency of these records currently specifying the same latitude/longitude pair as the bounding coordinates.

A particularly novel aspect of our approach is that we leverage weakly supervised training data which can be automatically generated using the deficient gazetteer records. We show that a semi-supervised learning regime greatly improves upon a fully-supervised one. This is important because manually labelled training data is expensive to generate.

VIII. ACKNOWLEDGEMENTS

This work was funded in part by NSF grants 0917069 and 1150115, and a Department of Energy Early Career Scientist and Engineer/PECASE award.

REFERENCES

[1] L. L. Hill, J. Frew, and Q. Zheng, "Geographic names: The implementation of a gazetteer in a georeferenced digital library," *D-Lib*, vol. 5, no. 1, 1999.

[2] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. Forsyth, "Names and faces in the news," in *CVPR*, vol. 2, 2004, pp. 848–854.

[3] K. Barnard, P. Duygulu, and D. Forsyth, "Clustering art," in *CVPR*, vol. 2, 2001, pp. 434–441.

[4] L.-J. Li, G. Wang, and L. Fei-Fei, "Optimol: Automatic Online Picture collecTion via Incremental MOdel Learning," in *CVPR*, 2007, pp. 1–8.

[5] C. Zhang, "Towards an operational system for automated updating of road databases by integration of imagery and geodata," *P&RS*, vol. 58, no. 3-4, pp. 166–186, 2004.

[6] P. Agouris, S. Gyftakis, and A. Stefanidis, "Using a fuzzy supervisor for object extraction within an integrated geospatial environment," *International Archives of Photogrammetry and Remote Sensing*, vol. 32, no. III/1, pp. 191–195, 1998.

[7] P. Doucette, P. Agouris, M. Musavi, and A. Stefanidis, "Automated extraction of linear features from aerial imagery using Kohonen learning and GIS data," in *ISD '99: Selected Papers from the International Workshop on Integrated Spatial Databases, Digital Inages and GIS*, 1999, pp. 20–33.

[8] A. Baumgartner, W. Eckstein, H. Mayer, C. Heipke, and H. Ebner, "Context-supported road extraction," *Automatic Extraction of Man-Made Objects from Aerial and Space Images*, vol. II, pp. 299–308, 1997.

[9] P. Agouris, K. Beard, G. Mountrakis, and A. Stefanidis, "Capturing and modeling geographic object change: A spatiotemporal gazetteer framework," *PE&RS*, vol. 66, no. 10, pp. 1241–1250, 2000.

[10] S. Newsam and Y. Yang, "Integrating gazetteers and remote sensed imagery," in *SIGSPATIAL*, 2008, pp. 26:1–26:10.

[11] Y. Yang and S. Newsam, "Estimating the spatial extents of geospatial objects using hierarchical models," in *WACV*, 2012, pp. 305–312.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[13] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.

[14] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *SIGSPATIAL*, 2010, pp. 270–279.

[15] C.-C. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines*, 2001.

[16] M. R. Amini, T. V. Truong, and C. Goutte, "A boosting algorithm for learning bipartite ranking functions with partially labeled data," in *SIGIR*, 2008, pp. 99–106.

[17] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *TGARS*, vol. 51, no. 2, pp. 818–832, 2013.