Cross-view Convolutional Networks

Nathan Jacobs jacobs@cs.uky.edu
Scott Workman scott@cs.uky.edu
Menghua Zhai ted@cs.uky.edu
Computer Science, University of Kentucky

Abstract

Billions of geotagged ground-level images are available via social networks and Google Street View. Recent work in computer vision has explored how these images could serve as a resource for understanding our world. However, most ground-level images are captured in cities and around famous landmarks; there are still very large geographic regions with few images. This leads to artifacts when estimating geospatial distributions. We propose to leverage satellite imagery, which has dense spatial coverage and increasingly high temporal frequency, to address this problem. We introduce Cross-view ConvNets (CCNs), a novel approach for estimating geospatial distributions in which semantic labels of ground-level imagery are transferred to satellite imagery to enable more accurate predictions.

1. Introduction

The wide availability of geotagged ground-level imagery, from sources such as social media and Google Street View, has created an opportunity to better understand the world from a human perspective. A specific example is the work by Arietta et al. [2] in mapping the human perception of safety in urban areas. Such approaches combine techniques from computer vision and geospatial statistics to estimate geospatial distributions from ground-level imagery.

Estimating geospatial distributions from ground-level imagery alone presents several challenges. The foremost challenge is automatically extracting human perceptions from a single image. The approach to this problem is well established: collect sample imagery, have each image annotated by one or more people, and then train a model to replicate these predictions. Recent studies include estimating memorability [7], virality [5], and urban perception [6]. There are two key remaining challenges: noise and sparsity. The noise problem arises because of the diversity of input imagery available via social media, including heavily edited, artistic, and non-photographic imagery. Such imagery can result in inaccurate estimates of human perceptions. The sparsity problem arises because there are relatively few images uploaded to social media away from major urban areas and tourist attractions. Together, these problems lead to errors in the resulting geospatial distributions.

To address this problem, we propose an alternative approach for estimating geospatial distributions, which we call Cross-view ConvNets (CCNs). Our method does not require manually annotating the satellite imagery. During training time, a satellite imagery understanding network is trained to predict a target label extracted from a co-located ground-level image. At test time, inference is performed using only satellite imagery. The CCN approach can be used to estimate geospatial distributions for many different quantities of interest. Our approach is similar to cross-view approaches used for ground-level image localization [12, 13, 22, 23]. The key difference is that we focus on the mapping not localization.

There are two main motivations for using ground-level imagery as a source of annotations instead of manually annotating the satellite imagery. First, ground-level imagery does not typically require an expert annotator, therefore low-cost services, such as Amazon Mechanical Turk, can be used to obtain labels. The relatively low cost for obtaining annotations for ground-level imagery has lead to the creation of large datasets for many tasks. Second, the annotations can be finer grained and more human relevant because many quantities of human interest are more readily apparent from a ground-level view than they are from a nadir viewpoint.

1.1. Background

The explosive increase in the amount of imagery available via social media platforms has spurred the creation of automatic methods for parsing, understanding, and exploiting this data. Crandall et al. [4] explore methods for organizing and analyzing large photo collections, including techniques for automatically identifying places of interest. Lu et al. [14] propose an automatic trip planning framework. Li et al. [11] propose a method for camera pose estimation that uses structure-from-motion models of famous landmarks.

Many methods have been proposed for deriving geo-

graphic information from georeferenced photo collections. Leung and Newsam [10] address the task of land cover classification. Wang et al. [20] estimate time-varying properties of the natural world, such as snow cover. Lee et al. [9] estimate several geospatial attributes, including population, gross domestic product, and population density. Zhu and Newsam estimate land usage [25]. Unlike these works, we use satellite imagery to avoid the limitations in estimating geospatial distributions using noisy estimates from sparsely distributed ground-level images.

Despite the growing availability of geotagged ground-level images, earth observation from satellite and aerial imagery continues to be an active research area. Mnih and Hinton [18] learn to detect roads in high-resolution aerial imagery. Cohen et al. [3] describe learning-based algorithms for detecting objects in geospatial imagery. Marcos et al. [16] propose a cross-domain feature which is sensor-invariant and demonstrate its functionality for updating land cover maps and change detection. Our approach of using ground-level images as a supervisory signal eliminates the need for manual annotation of satellite imagery.

Some work has explored jointly reasoning over ground-level and satellite image pairs in order to improve the performance of traditional vision tasks: Luo et al. [15] address the task of event recognition; Sakurada et al. [19] estimate large-scale land surface conditions; Mattyus [17] propose to label roads and sidewalks; and Wegner et al. [21] construct a catalog of street trees. Unlike these approaches, our proposed approach only requires ground-level imagery at training time. This means that our approach works well in urban and rural areas.

1.2. Problem Statement

We address the problem of using imagery to automatically estimate a geospatial distribution, or, alternatively, a map. This distribution could represent many different quantities of interest, including land cover, land use, or population density. We assume that the quantity of interest is apparent in both a nadir view and a ground-level view of a particular geographic location. In the following section, we define our strategy for estimating such a distribution by combining ground-level and satellite imagery. In Section 3, we show how to use this strategy to estimate geospatial distributions for scenicness and ground-level scene category.

2. Approach

We propose cross-view convolutional networks (CCNs), a strategy for transferring semantic labels of ground-level imagery to satellite imagery. The goal of this strategy is to enable us to make higher-resolution maps than if we only used ground-level imagery. See Figure 1 for a visual overview of the approach. We assume we have a method for obtaining labels for ground-level imagery, and that the

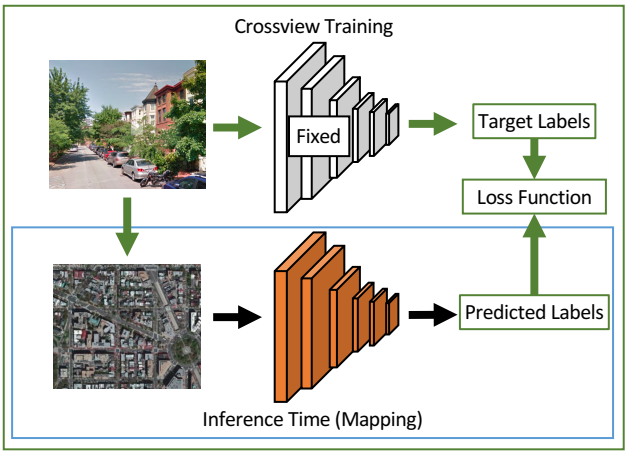

Figure 1. Approach Overview.

ground-level images are geotagged (their geographic location is known). There are many ways of ensuring the existence of ground-level labels: 1) manually annotating all the imagery, 2) using an existing method for automatically labeling ground-level imagery, and 3) labeling a subset of the imagery, learning a model to predict the label, and then automatically labeling the remaining ground-level imagery. These approaches enable us to take advantage of the ease of labeling ground-level imagery by non-expert users and the vast number of existing methods for ground-level image understanding, some of which achieve human-level performance.

A high-level overview of our training process is as follows. First, collect co-located pairs of satellite and geotagged ground-level images. Second, obtain labels for each of the ground-level images, using one the strategies described above. Finally, train a convolutional neural network, which takes only the satellite image as input, to predict the label of the ground-level image. The result of this process is a method for estimating a quantity of interest using only satellite imagery. To obtain a final dense map, we apply the network convolutionally to the satellite image of a region of interest. Specific details, such as network architecture, loss function, and dataset, will vary depending on the application.

3. Examples

We have proposed a high-level approach for estimating geospatial distributions using satellite imagery. In this section, we apply our approach to two real-world tasks. For each task, we define the method for obtaining ground-level image labels, the neural network architectures for the ground-level and satellite networks, the training procedure, and the evaluation dataset. While both of these results are purely qualitative, they highlight the potential of the CCN framework.

3.1. Mapping Scene Categories (San Francisco)

We consider the task of constructing a map of ground-level scene categories, which is closely related to the traditional remote-sensing task of land cover/land use classification. We use the *Places* CNN [24], which predicts the scene category of a given image, as our fixed ground-level image understanding method. The output of the *Places* CNN, which was trained on a large-dataset of ground-level images, is a categorical distribution over 205 scene classes.

We follow the approach outlined in Section 2 to learn a CCN which operates on satellite imagery and predicts the *Places fc8* features of the corresponding ground-level image. To train our model, we used imagery from the CVUSA dataset [23], which consists of approximately 1.5 million geotagged ground-level images sparsely sampled from across the United States. For each ground-level image, there is a satellite image centered around the ground-level image location. For this experiment, we use the *AlexNet* CNN [8] architecture, and minimize the L_2-norm. To optimize the parameters of our model we used stochastic gradient descent, with hyperparameters set to those used for training the *Places* CNN.

For visualization purposes, we use ground-level and aerial imagery captured around San Francisco [22]. We manually selected three categories: urban (the *parking-lot* class), rural (the *field/wild* class), and water (the *ocean* class). Figure 2 (left) shows a scatter plot of these categories for the ground-level images. To construct the final map of scene categories (Figure 2), we applied our learned CNN convolutionally to the dense grid of satellite images. We reduced the dimensionality of our output predictions from 205 dimensions to 3 by extracting the logits for each category and applying a softmax activation function per geographic location. The output of the softmax was used as the RGB pixel value for the corresponding map location. This process results in a map that captures high resolution local structure with minimal artifacts.

As a baseline for comparison, we interpolated the sparse set of samples estimated using only the ground-level images. To perform the interpolation, we use locally weighted averaging (LWA), with latitude/longitude as the input features and a Gaussian kernel. The values for each category were interpolated separately, resulting in a three-channel output which is visualized as an RGB image. The results show that no kernel bandwidth is free of noticeable artifacts (it is either too smooth or too noisy).

3.2. Mapping Scenicness (United Kingdom)

We demonstrate the CCN training approach for the task of creating a map of scenic places in the United Kingdom. Such a map could be used, for example, to plan a road trip or a hike. We use a large dataset of ground-level images, each of which was annotated by at least three individuals [1]. Each annotation is an integer value from one (not scenic) to 10 (highly scenic). For each image, we aggregate the scenicness votes by computing the average value. Figure 3 (left) shows a scatter plot of the average value of the scenicness score. The middle column shows the result of using locally weighted averaging to interpolate the sparse samples. As before, we used a Gaussian kernel but in this example we solved for an optimal kernel bandwidth across the full dataset.

For our CCN approach, we trained an *AlexNet* CNN to predict the ground-level scenicness using the corresponding satellite image. We used the same settings as in the previous example to train the model. We then applied this network convolutionally across the entire region of interest to obtain the maps shown in Figure 3. As in the previous example, the CCN approach has significantly higher spatial resolution and fewer artifacts than the baseline approach.

4. Conclusion

We proposed a novel approach for estimating geospatial distributions from satellite imagery that does not require manual labeling. Our approach is quite general; it can be applied to many quantities of interest. Cross-view ConvNets combine ground-level and satellite imagery at training time and uses only satellite imagery during inference. For future work, we are exploring ways of fusing information extracted from ground-level imagery and satellite imagery during inference. In addition, we plan to apply this strategy to a range of real-world application domains using both visible and non-visible light satellite imagery.

Acknowledgements

We gratefully acknowledge the support of NSF CAREER grant (IIS-1553116) which partially supported this work.

References

[1] http://scenicornot.datasciencelab.co.uk/. 3

[2] S. M. Arietta, A. A. Efros, R. Ramamoorthi, and M. Agrawala. City forensics: Using visual elements to predict non-visual city attributes. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2624–2633, 2014. 1

[3] J. P. Cohen, W. Ding, C. Kuhlman, A. Chen, and L. Di. Rapid building detection using machine learning. *Applied Intelligence*, 2016. 2

[4] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *International World Wide Web Conference*, 2009. 1

[5] A. Deza and D. Parikh. Understanding image virality. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1

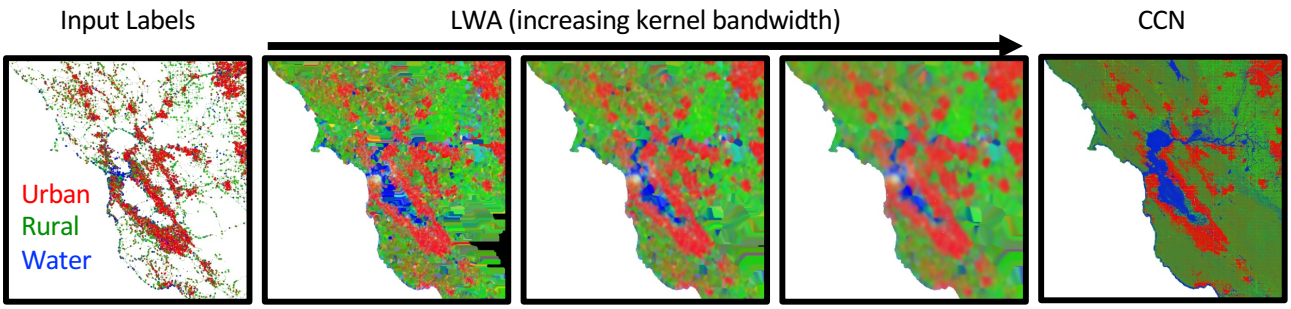

Figure 2. Comparing methods for mapping scene categories. (left) Each dot represents the location of a ground-level image. The colors correspond to the scene category of the image. (middle) Three locally weighted average (LWA) interpolations of the ground-level image labels. All kernel bandwidth settings have significant artifacts. Regions in the ocean are colored white, because satellite imagery over the ocean was not included in the source dataset [22]. (right) Our cross-view mapping approach results in higher quality, higher resolution maps than the LWA approach.

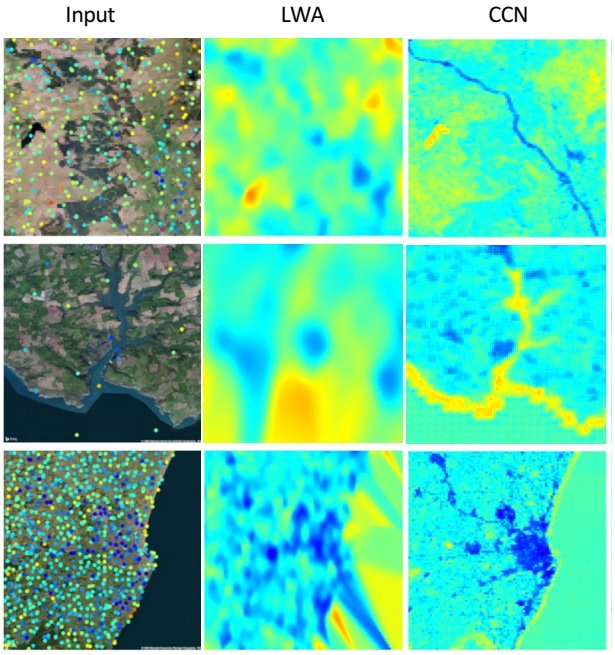

Figure 3. (left) Satellite images of a region of interest, each dot corresponds to the location of a ground-level image. The color of each dot reflects scenicness of the ground-level image (blue=not scenic, green=moderately scenic, red=highly scenic). (middle) A locally weighted average (LWA) of the scenicness values of the ground-level images. (right) The predictions from our cross-view convolutional network. Note the higher spatial resolution compared to the LWA method, especially when ground-level images are sparsely distributed.

[6] A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *European Conference on Computer Vision*, 2016. 1

[7] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva. Understanding and predicting image memorability at a large scale. In *IEEE International Conference on Computer Vision*, 2015. 1

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 3

[9] S. Lee, H. Zhang, and D. J. Crandall. Predicting geo-informative attributes in large-scale image collections using convolutional neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2014. 2

[10] D. Leung and S. Newsam. Proximate sensing: Inferring what-is-where from georeferenced photo collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 2

[11] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3d point clouds. In *European Conference on Computer Vision*, 2012. 1

[12] T.-Y. Lin, S. Belongie, and J. Hays. Cross-view image geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 1

[13] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays. Learning deep representations for ground-to-aerial geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1

[14] X. Lu, C. Wang, J.-M. Yang, Y. Pang, and L. Zhang. Photo2trip: generating travel routes from geo-tagged photos for trip planning. In *ACM International Conference on Multimedia*, 2010. 1

[15] J. Luo, J. Yu, D. Joshi, and W. Hao. Event recognition: viewing the world with a third eye. In *ACM International Conference on Multimedia*, 2008. 2

[16] D. Marcos, R. Hamid, and D. Tuia. Geospatial correspondences for multimodal registration. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[17] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun. Hd maps: Fine-grained road segmentation by parsing ground and aerial images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[18] V. Mnih and G. E. Hinton. Learning to detect roads in high-resolution aerial images. In *European Conference on Computer Vision*, 2010. 2

[19] K. Sakurada, T. Okatani, and K. M. Kitani. Massive city-scale surface condition analysis using ground and aerial imagery. In *Asian Conference on Computer Vision*, 2014. 2

[20] J. Wang, M. Korayem, and D. Crandall. Observing the natural world with flickr. In *ICCV Workshop on Computer Vision for Converging Perspectives*, 2013. 2

[21] J. D. Wegner, S. Branson, D. Hall, K. Schindler, and P. Perona. Cataloging public objects using aerial and street-level images - urban trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[22] S. Workman and N. Jacobs. On the location dependence of convolutional neural network features. In *IEEE/ISPRS Workshop: Looking From Above: When Earth Observation Meets Vision*, 2015. 1, 3, 4

[23] S. Workman, R. Souvenir, and N. Jacobs. Wide-Area Image Geolocalization with Aerial Reference Imagery. In *IEEE International Conference on Computer Vision*, 2015. 1, 3

[24] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, 2014. 3

[25] Y. Zhu and S. Newsam. Land use classification using convolutional neural networks applied to ground-level images. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2015. 2