Cross-View Image Geolocalization

Tsung-Yi Lin and Serge Belongie
University of California, San Diego
tsl008@ucsd.edu, sjb@cs.ucsd.edu

James Hays
Brown University
hays@cs.brown.edu

Abstract

The recent availability of large amounts of geotagged imagery has inspired a number of data driven solutions to the image geolocalization problem. Existing approaches predict the location of a query image by matching it to a database of georeferenced photographs. While there are many geotagged images available on photo sharing and street view sites, most are clustered around landmarks and urban areas. The vast majority of the Earth's land area has no ground level reference photos available, which limits the applicability of all existing image geolocalization methods. On the other hand, there is no shortage of visual and geographic data that densely covers the Earth – we examine overhead imagery and land cover survey data – but the relationship between this data and ground level query photographs is complex. In this paper, we introduce a cross-view feature translation approach to greatly extend the reach of image geolocalization methods. We can often localize a query even if it has no corresponding ground-level images in the database. A key idea is to learn the relationship between ground level appearance and overhead appearance and land cover attributes from sparsely available geotagged ground-level images. We perform experiments over a 1600 km^2 region containing a variety of scenes and land cover types. For each query, our algorithm produces a probability density over the region of interest.

1. Introduction

Consider the photos in Figure 1. How can we determine where they were taken? One might try to use a search-by-image service (e.g., Google Images) to retrieve visually similar images. This will only solve our problem if we can find an instance-level match with a known location. This approach will likely succeed for famous landmarks, but not for the unremarkable scenes in Figure 1. If instead of instance-level matching we match based on scene-level features, as in im2gps [9], we can sometimes get a coarse geolocation based on the distribution of similar scenes. Is this our best hope for geolocalizing photographs? Fortunately, no – there are numerous additional levels of image understanding that can help constrain the location of a photograph. For instance, in the popular "View from Your Window" contest[1], humans discover a litany of geo-informative visual evidence related to architecture, climate, road markings, style of dress, etc. However, recognizing these properties and then mapping them to geographic locations is at the limit of human ability and far beyond computational methods.

Figure 1: The above ground level images were captured in Charleston, SC within the region indicated in this satellite image. What can we determine about their geolocations? In this work, we tackle the case for which ground level training imagery at the corresponding locations is not available.

In this paper, we take a small step in this direction by exploiting two previously unused geographic data sets – overhead appearance and land cover survey data. For each of these data sets we must learn the relationship between ground level views and the data. These data sets have the benefits of being (1) densely available for nearly all of the Earth and (2) rich enough such that the mapping from ground level appearance is sometimes unambiguous. For instance, a human might be able to verify that a putative match is correct (although it is infeasible for a human to do exhaustive searches manually).

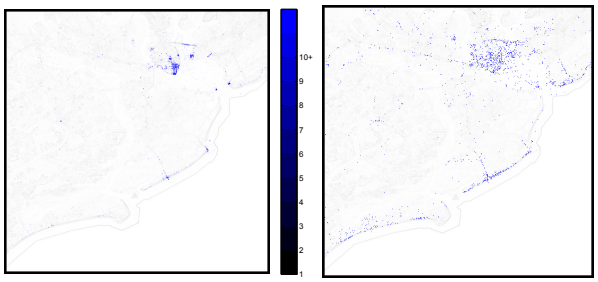
[1] http://dish.andrewsullivan.com/vfyw-contest/

The im2gps approach of [9] was the first to predict image geolocation by matching visual appearance. Following this thread, [13, 5, 27] posed the geolocalization task as an image retrieval problem, focusing on landmark images on the Internet. [22, 20, 4] use the images captured from street view cars in order to handle more general query images. [1] developed an approach to build 3D point cloud models of famous buildings automatically from publicly available images. [10, 15, 21, 14] leverage such 3D information for more efficient and accurate localization. Note that all the above mentioned approaches assume at least one training image is taken from a similar vantage points as that of the query. A query image in an unsampled location has no hope of being accurately geolocated. Recently, [2] presented an algorithm that does not require ground-level training images. They leverage a digital elevation model to synthesize 3D surfaces in mountainous terrain and match the contour of mountains extracted from ground-level images to the 3D model. The paper extends the solution space of existing geolocalization methods from popular landmarks and big cities to mountainous regions. However, most photographs do not contain mountains. For our region of interest in this work reasoning about terrain shape would not be informative.

Matching ground-level photos to aerial imagery and geographic survey data has remained unexplored despite these features being easily available and densely distributed on the surface of the earth. One major issue preventing the use of these features is that the mapping between them is very complex. For instance, the visual appearance of a building in ground-level vs. an aerial view is very different due the extremely wide baseline, varying focal lengths, non planarity of the scene, different lighting conditions and mismatched image quality.

Our approach takes inspiration from recent works in cross-view data retrieval that tackle problems such as static cameras localization with satellite imagery [11], cross-view action recognition [16] and image-text retrieval [19, 23]. These approaches achieve cross-view retrieval by learning the co-occurrence of features in different views. The training data consists of corresponding features that describe the same content in multiple views and these techniques often involve solving a generalized eigenvalue problem to find the projection basis that maximizes the cross-view feature correlation. One example is Kernelized Canonical Correlation Analysis [7] which requires solving the eigenvalue problem of training kernel matrix. However, this approach is not scalable to very large datasets.

Inspired by [8, 18, 26], we propose a data-driven framework for cross-view image matching. To this end, we collected a new dataset that consists of ground-level images, aerial images, and land cover attribute images as the training data. At the test time, we leverage the scene matches at ground-level to find possible matches to features in aerial and attribute images. The contributions in this paper include (1) we are able to geolocalize images without corresponding ground-level images in the database and therefore extend the current solution space; (2) We build a new dataset for cross-view matching and leverage aerial imagery and land cover attributes; (3) produce a dense prediction score over a large scale (40km $\times$ 40km) search space.



(a) All database Images (b) Isolated images

Figure 2: The distribution of ground level images from Panoramio within the region indicated in Figure 1. (a) Most of the ground level images are concentrated in a few highly populated areas for which ground-to-ground level matching will suffice. (b) Each point marks an "isolated" image for which no nearby ground level image is available in the database. Such images must be localized using cross-view (ground-to-aerial) image matching.

2. Dataset

For the experiments in this paper we examine a 40km $\times$ 40km region around Charleston, SC. This region exhibits great scene variety (urban, agricultural, forest, marsh, beach, etc.) as shown in Figure 1. We take one ground-level image as query and match it to the map database of aerial images and land cover attributes. Our training data consists of triplets of ground-level images, aerial images, and attribute maps; see Figure 4. To justify the need for training data from different views, Figure 2a shows the ground-level image distribution in our training data. Because the ground-level images are sparsely distributed over the region of interest, ground-level retrieval methods in the vein of im2gps will fail when no nearby training images are close to query images as shown in Figure 2b. In our database, 98.76% of the space is not covered by any ground-level image if we consider each image occupies 180m x 180m field of view. In such cases, the proposed method can leverage co-occurrence information in training triplets to match "isolated images" to aerial and attribute images in the map database. Note that the proposed method can be generalized nationwide or globally because the training data we use is widely available.

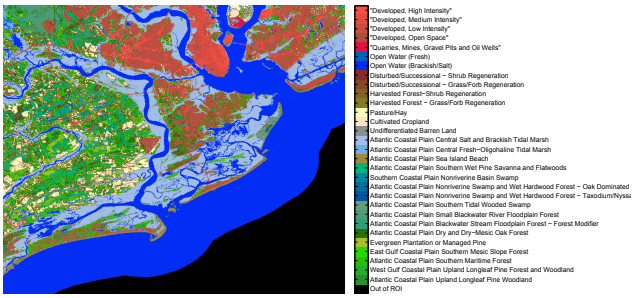


Figure 3: Snapshot of the USGS GAP land cover attributes (available nationwide) for Charleston, SC.

2.1. Ground-level Imagery

We downloaded 6756 ground-level images from Panoramio. We do not apply any filters to the resulting images, even though many are likely impossible to geolocate (e.g., close up views of flowers). The great scene variety of photographs "in the wild" makes the dataset extremely challenging for geolocalization.

2.2. Aerial Imagery

The satellite imagery is downloaded from Bing Maps. Each 256×256 image covers a 180m × 180m region. The image resolution is 0.6 m/pixel. Training images are cropped from the center of the labeled geolocation. The map database images are densely sampled by a sliding window with half stride on the region of interest. We collect a total of 182,988 images for the map database. We rotate each aerial image to its dominant orientation for rotation invariant matching.

2.3. Land Cover Attribute Imagery

The National Gap Analysis Program (GAP) conducted a national wide geology survey and produced the USGS GAP Land Cover Data Set.[2] Figure 3 shows a snapshot of the dataset at our region of interest. The dataset contains two hierarchical levels of classes including 8 general classes (e.g., Developed & Other Human Use, Shrubland & Grassland, etc.) and 590 land use classes as the subclasses (e.g., Developed/Medium Intensity, Salt and Brackish Tidal Marsh, etc.) Each pixel in the attribute map is assigned to one general class and one subclass. The attribute image resolution is 30 m/pixel. We crop 5×5 images for the training data and the map database.

3. Geolocalization via Cross-View Matching

In this section we discuss feature representation for ground, aerial, and attribute images and introduce several methods for cross-view image geolocation.

[2]http://gapanalysis.usgs.gov/gaplandcover/

Figure 4: Example "training triplet": ground level image, its corresponding aerial image and land cover attribute map.

3.1. Image Features and Kernels

Ground and Aerial Image: We represent each ground image and aerial tile using four features: HoG [6], self-similarity [24], gist [17], and color histograms. The use of these features to represent ground level photographs is motivated by their good performance in scene classification tasks [25]. Even though these features were not designed to represent overhead imagery, we use them for that purpose because (1) it permits a simple "direct matching" baseline for cross-view matching described below and (2) aerial image representations have not been well explored in the literature, so there is no "standard" set of features to use for this domain. HoG and self-similarity local features are densely sampled, quantized, and represented with a spatial pyramid [12]. When computing image similarity we use a histogram intersection kernel for HoG pyramids, χ^2 kernel for self-similarity pyramids and color histograms, and RBF kernel for gist descriptors. We combine these four feature kernels with equal weights to form the aerial feature kernel we use for learning and prediction in the following sections. All features and kernels are computed by code from [25].

Land Cover Attribute Image: Each ground and aerial image has a corresponding 5x5 attribute image. From this attribute image we build histograms for the general classes and subclasses and then concatenate them to form a multinomial attribute feature. We compare attribute features with a χ^2 kernel.

Our geolocation approach relies on translating from ground level features to aerial and/or land cover features. The aerial and attribute features have distinct feature dimension, sparsity, and discriminative power. In the Experiments section, we compare geolocation predictions based on aerial features, attribute features, and combinations of both.

3.2. Matching

In this section, We introduce two noval data-driven approaches, Data-driven Feature Averaging and Discriminative Translation, for cross-view geolocalization. We introduce three baseline algorithms for comparison. We first introduce im2gps [9] which can only match images within the ground-level view. We then introduce Direct Match and Kernelized Canonical Correlation Analysis as the baseline

algorithms for cross-view geolocalization. In the following section, x denotes the query image. The bold-faced $\mathbf{x}$ denotes ground-level images and $\mathbf{y}$ denotes aerial/attribute images of corresponding training triplets. $\mathbf{y}_{map}$ denotes the aerial/attributes images in the map database. $k(.,.)$ denotes the kernel function.

im2gps: im2gps geolocalizes a query image by guessing the locations of the top k scene matches. im2gps or any image retrieval based method makes no use of aerial and attribute information and can only geolocate query images in locations with ground-level training imagery. We compute $k(x, \mathbf{x})$ to measure the similarity of the query image and the training images.

Direct Match (DM): In order to geolocate spatially isolated photographs, we need to match across views from ground level to overhead. The simplest method to match ground level images to aerial images is just to match the same features with no translation, i.e., to assume that ground level appearance and overhead appearance are correlated. This is not universally true, but a beach from ground level and overhead do share some rough texture similarities, as do other scene types, so this baseline performs slightly better than chance. We compute similarity with:

$$sim_{dm} = k(x, \mathbf{y}_{map}) \tag{1}$$

The direct match baseline cannot be used for ground to attribute matching, because those features sets are distinct.

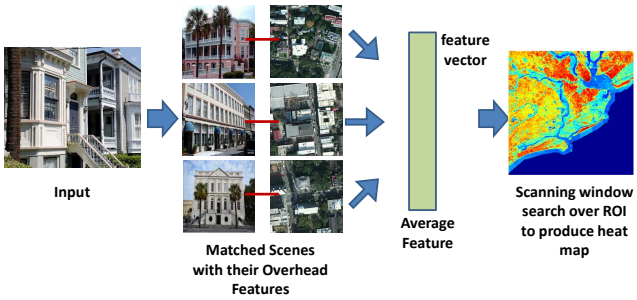
Kernelized Canonical Correlation Analysis (KCCA): KCCA is a tool to learn the basis along the direction where features in different views are maximally correlated:

$$\max_{w_x \neq 0, w_y \neq 0} \frac{w_x^\top \Sigma_{xy} w_y}{\sqrt{w_x^\top \Sigma_{xx} w_x} \sqrt{w_y^\top \Sigma_{yy} w_y}} \tag{2}$$

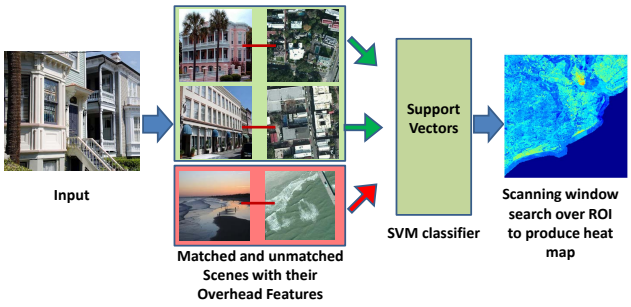
where Σ_{xx} and Σ_{yy} represent the covariance matrices for ground-level feature and aerial/attribute feature in training triplets. Σ_{xy} represent the cross-covariance matrix between them. The optimization can be posed as a generalized eigenvalue problem and solved as a regular eigenvalue problem. In KCCA, we use the "kernel trick" to represent a basis as a linear combination of training examples by $w_{xi} = \alpha_i^\top \phi(\mathbf{x})$ and $w_{yi} = \beta_i^\top \phi(\mathbf{y})$. We compute the cross-view correlation score of query and training images by summing over the correlation scores on the top d bases:

$$sim_{kcca} = \sum_{i=1}^{d} \alpha_i^\top k(x, \mathbf{x}) \beta_i^\top k(\mathbf{y}, \mathbf{y}_{map}) \tag{3}$$

We use the cross-view correlation as the matching score. KCCA has several disadvantages for large scale cross-view matching. First, we need to compute the singular value decomposition for a non-sparse kernel matrix to solve the eigenvalue problem. The dimension of kernel matrix grows with the number of training samples. This makes the solution infeasible as training data increases. Second, KCCA assumes one-to-one correspondence between two views. But in our problem, it is common to have multiple ground-level images taken at the same geolocation. In this case, we need to throw away some training data to enforce the one-to-one correspondence.



(a) Feature Averaging.



(b) Discriminative translation.

Figure 5: Our proposed data-driven pipelines for cross view matching. (a) The input ground-level image is matched to available ground-level images in the database, the features of the corresponding aerial imagery are averaged, and this averaged representation is used to find potential matches in the region of interest (ROI). (b) We train an SVM using batches of highly similar and dissimilar aerial imagery and apply it to sliding windows over the ROI.

Data-driven Feature Averaging (AVG): When images match well in the ground-level view they also tend to have similar overhead views and land cover attributes. Based on this observation, we propose a simple method to translate ground-level to aerial and attribute features by averaging the features of good scene matches. Figure 5a shows the pipeline of AVG. We first find the k most similar training scenes as we do for im2gps and then average their corresponding aerial or land cover features to form our predictions of the unknown features. Finally, we use the averaged features to match features in the map database:

$$sim_{avg} = k(\sum_i y_i, \mathbf{y}_{map}), i \in knn(x, \mathbf{x}) \tag{4}$$

Discriminative Translation (DT): In the AVG approach, we only utilize the best scene matches to predict the ground truth aerial or land cover feature. But the *dissimilar* ground level training scenes can also be informative – scenes with very different ground level appearance tend to have distinct overhead appearance and ground cover attributes. Note that this can be violated when two photos at the same location (and thus with the same overhead appearance and attributes) have very different appearance (e.g., a macro flower photograph vs. a landscape shot). In order to capitalize on the benefits of negative scene matches, we propose a discriminative cross-view translation method. The pipeline of "DT" is shown in Figure 5b. The positive set of DT is the same as AVG and we add to that a relatively large set of negative training samples from the bottom 50% ranked scene matches. We train a support vector machine (SVM) on the fly with the aerial/attribute features of the positive and negative examples. Finally, we apply the trained classifier on the map database:

$$sim_{dt} = \sum_i \alpha_i k(y_i, \mathbf{y}_{map}) \tag{5}$$

where the α_i is the weight of support vectors. Note that unlike KCCA, DT allows many-to-one correspondence for training triplet and results in a more compact representation (the support vectors) for prediction. As a result, DT is more applicable to large training databases and more efficient for testing.

4. Experiments

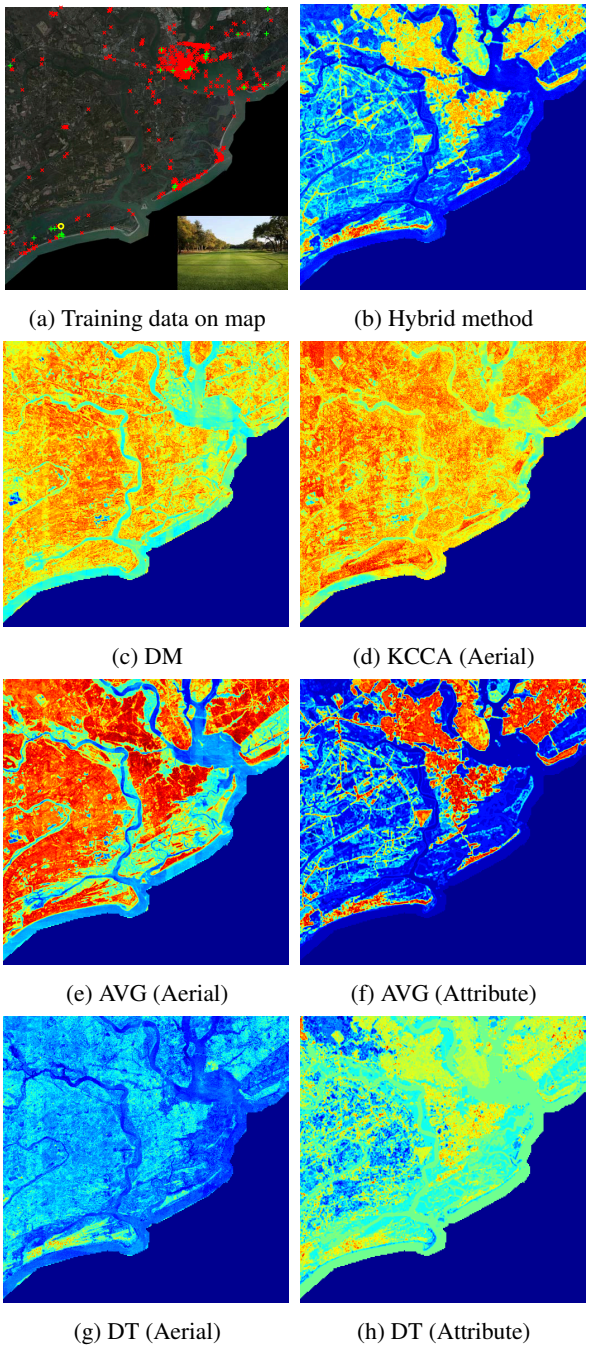
4.1. Test Sets and Parameter Settings

To evaluate the performance, we construct two hold-out test sets **Random** and **Isolated** from the ground-level image database.

Random: we randomly draw 1000 images from the ground-level image database. Some images in this set come from frequently photographed locations and thus the training images could contain instance-level matches.

Isolated: We use 737 isolated images with no training images in the same 180m x 180m block. We are most interested in geolocalizing these isolated images because existing methods (e.g., im2gps) fail for all images in this set.

For discriminative translation (DT), for each query we select the top 30 scene matches and 1500 random bad matches for training. We train the kernel SVM classifier by libsvm package [3]. For KCCA, we enforce one-to-one correspondence by only choosing one ground-level image from multiple images at the same location in the training data. We project the data onto first 500 learned bases to compute the cross-view correlation.



(a) Training data on map (b) Hybrid method

(c) DM (d) KCCA (Aerial)

(e) AVG (Aerial) (f) AVG (Attribute)

(g) DT (Aerial) (h) DT (Attribute)

Figure 6: (a) shows the geolocation of query image (yellow), the best 30 scene matches (green) and 1500 bad scene matches (red). (b) The "hybrid" approach has the best performance in the experiment. (c-d) Direct matching (DM) and KCCA as baseline algorithm. (e-h) Feature averaging (AVG) and discriminative translation (DT) on aerial and attribute feature. AVG is more reliable with attribute features while DT is more accurate with aerial features.

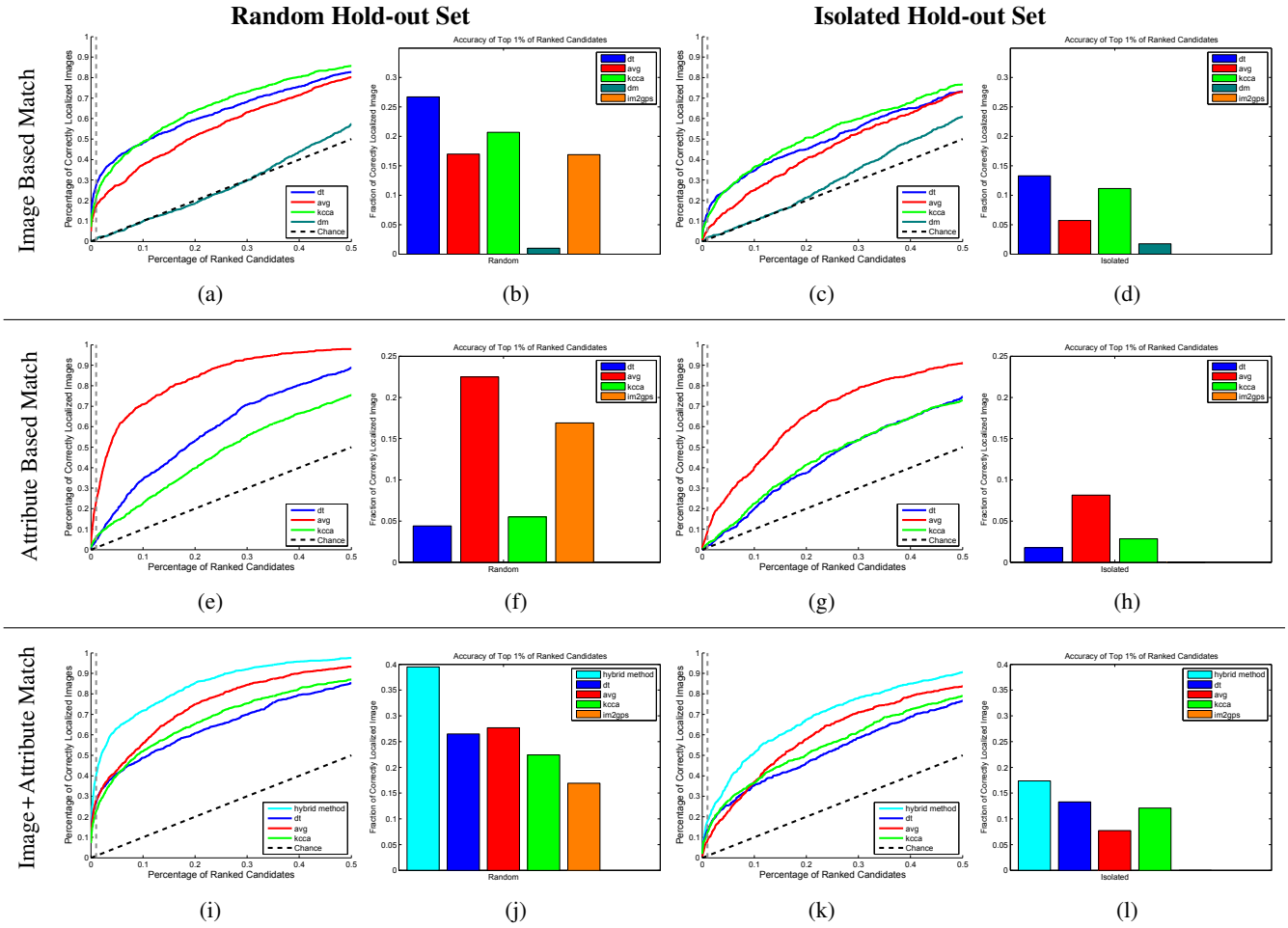


Figure 7: Each row shows the accuracy of localization as a function of the fraction of retrieved candidate locations for random locations and isolated locations, respectively, followed by a bar plot slice through the curves at the 1% mark (rank 1830). (a-d) Image based matching only, (e-h) land cover attribute based matching only, (i-l) combined image and land cover attribute based matching. Note in particular that the accuracy of im2gps is exactly zero for the isolated cases across all three bar plots, since in such cases ground-level reference imagery is not available. In contrast, im2gps enjoys strong performance in the random case, for which ground level reference imagery is often available. We also note that the SVM performance is poor in row 2 (land cover attributes only) since many regions with similar and dissimilar visual appearance share the same attribute distribution. (l) With our hybrid approach, we can determine the correct geolocation for 17.37% of our query images when we consider the top 1% best matching candidates.

4.2. Performance Metric

Each cross-view matching algorithm returns a score for every map location (we visualize these as heat maps). For each query, we sort the scores and the evaluate the performance by the rank at the ground truth location. The second and fourth column of Figure 7 shows the fraction of query images where the ground truth location is ranked in the top 1% of map locations. For im2gps, we look at the top 30 scene matches for evaluation. Note that we use the same 30 matched images as our positive training examples for AVG and DT. That means the accuracy of im2gps indicates the fraction of query images that have the training data at the ground truth location for AVG and DT.

4.3. Matching Performance

In this section, we evaluate matching ground-level features to aerial and/or attribute features. First we show the heat map of DM and KCCA as our baseline algorithms. Figure 6c shows DM does not correctly indicate the ground truth location. Figure 6d shows KCCA can produce higher probability density around the ground truth region but its probability density also spreads over to some regions that are impossible for the ground truth location.

The top row of Figure 7 shows the performance of

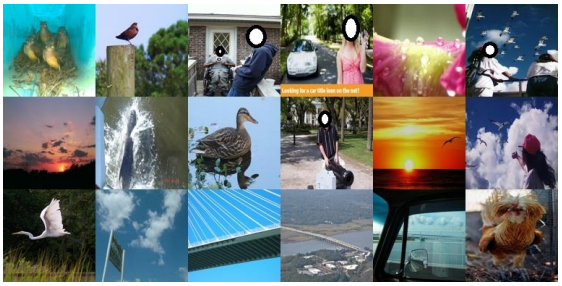
matching ground to aerial features. The performance of DT is better than AVG especially at the low rank region. Figure 6e shows that the heat map of AVG is too uniform while Figure 6g shows that the heat map of DT concentrates around the ground truth location. This suggests that the predicted features from AVG fail to learn a discriminative representation for each query. This is unsurprising, as discriminative methods have been successful across many visual domains. For instance, in object detection, averaging descriptors from objects in the same class may produce a new feature that is also similar to many different classes. On the other hand, DT can leverage the negative training data to predict the aerial feature that is only similar to the given positive examples in a discriminative way.

The middle row of Figure 7 shows the performance of matching ground to attribute features. This approach is less accurate than matching to overhead appearance because attribute features are not discriminative enough by themselves. For instance, an attribute feature may correspond to many locations. Perhaps because the attribute representation is low dimensional and there are many duplicate instances in the database, DT does not perform as well as the simple AVG approach.

The bottom row of Figure 7 shows the performance of matching based on predicting both aerial and attribute features. This approach performs better than each individual feature which suggests that the aerial and attribute features complement each other. In particular, we find that the best performance comes by combining the attribute-based rankings from the AVG method and the overhead appearance based rankings from the DT method. This "hybrid" method achieves 39.5% accuracy for **Random** and 17.37% for **Isolated**.

Figure 8 shows a random sample of successfully and unsuccessfully geolocated query images. Our system handles outdoor scenes better because they are more correlated to the aerial and attribute images. Photos focused on objects may fail because the correlation between objects and overhead appearance and/or attributes is weak. Figure 9 visualizes query images, corresponding scene matches, and heat maps. We demontrate that our proposed algorithm can handle the wide range of scene variety by showing query examples from three very different scenes.

(a)



(b)

Figure 8: Gallery of (a) successfully and (b) unsuccessfully matched isolated query images in our experiments. Photos that prominently feature objects tend to fail because the aerial and land cover features are not well constrained by the scene features.

5. Conclusion

In this paper we propose a cross-domain matching framework that greatly extends the domain of image geolocation. Most of the Earth's surface has no ground level geotagged photos publicly (98% of the area of our experimental region, even though it is part of the densely populated Eastern coast of the US) and traditional image geolocation methods based on ground level image to image matching will fail for queries in such locations. Using our new dataset of ground-level, aerial, and attribute images we quantified the performance of several baseline and novel approaches for "cross-view" geolocation. In particular, our "discriminitive translation" approach in which an aerial image classifier is trained based on ground level scene matches can roughly geolocate 17% of isolated query images, compared to 0% for existing methods. While the experiments in this paper are at a modest scale (1600km^2 region of interest), the approach scales up easily both in terms of data availability and computational complexity. Our approach is the first to use overhead imagery and land cover survey data to geolocate photographs, and it is complementary with the impressive mountain-based geolocation method of [2] which also uses a widely available geographic feature (digital elevation maps).

Acknowledgments: Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building Rome in a day. In *ICCV*, 2009. 2

[2] G. Baatz, O. Saurer, K. Köser, and M. Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *ECCV*, 2012. 2, 7

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011. 5

[4] D. Chen, G. Baatz, Köser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. In *CVPR*, 2011. 2

[5] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW*, 2009. 2

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, 2005. 3

[7] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664, Dec. 2004. 2

[8] J. Hays. *Large Scale Scene Matching for Graphics and Vision*. PhD thesis, Carnegie Mellon University, 2009. 2

[9] J. Hays and A. Efros. IM2GPS: estimating geographic information from a single image. In *CVPR*, 2008. 1, 2, 3

[10] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR*, 2009. 2

[11] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. Geolocating static cameras. In *ICCV*, Oct. 2007. 2

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 3

[13] X. Li, C. Wu, C. Zach, S. Lazebnik, and J. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV*, 2008. 2

[14] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3D point clouds. In *ECCV*, 2012. 2

[15] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *ECCV*, 2010. 2

[16] J. Liu, M. Shah, B. Kuipers, and S. Savarese. Cross-view action recognition via view knowledge transfer. In *CVPR*, 2011. 2

[17] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 2001. 3

[18] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 2

[19] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM International Conference on Multimedia*, 2010. 2

[20] A. Roshan Zamir and M. Shah. Accurate image localization based on Google maps street view. In *ECCV*, 2010. 2

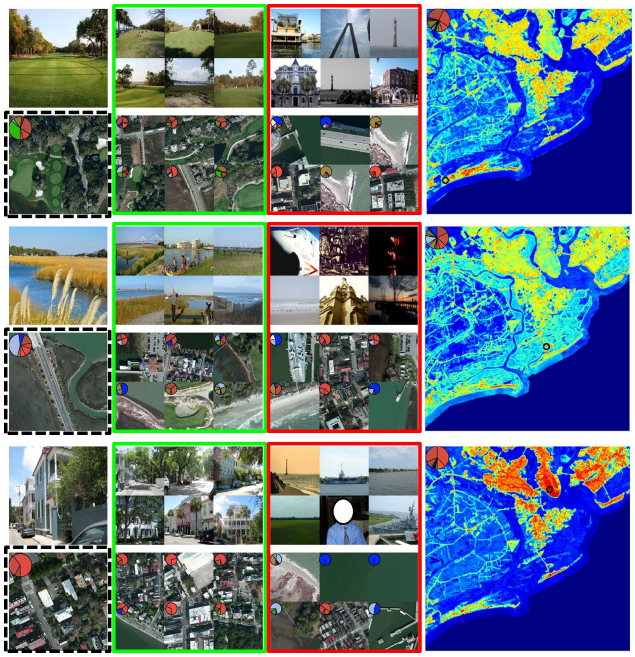


Figure 9: Left: input ground level image (shown above) and corresponding satellite image and pie chart of attribute distribution (shown below, but not known at query time). Middle: similar (in green) and dissimilar (in red) ground-level and satellite image pairs used for training the SVM in our discriminative translation approach. Right: geolocation match score shown as a heat map. The ground truth location is marked with a black circle.

[21] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching. In *ICCV*, 2011. 2

[22] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *CVPR*, 2007. 2

[23] A. Sharma, A. Kumar, H. Daumé III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012. 2

[24] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007. 3

[25] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 3

[26] H. Zhang, A. C. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006. 2

[27] Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T. Chua, H. Neven, and J. Yagnik. Tour the world: building a web-scale landmark recognition engine. In *CVPR*, 2009. 2