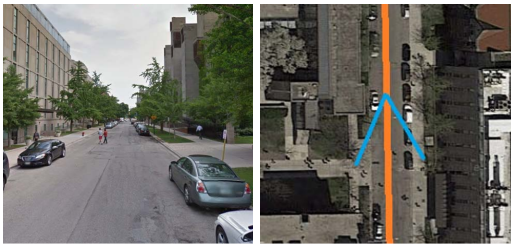
Accurate Vision-based Localization by Transferring Between Ground and Satellite Images

Hang Chu **Hongyuan Mei** **Mohit Bansal** **Matthew R. Walter**
Toyota Technological Institute at Chicago, Chicago, IL 60637, USA
{hchu,hongyuan,mbansal,mwalter}@ttic.edu

Abstract

We present a method for accurately localizing a ground image with the aid of satellite imagery by transferring between the two image modalities. We perform visual localization by estimating the co-occurrence probabilities between the ground and satellite images. This method allows us to estimate location probabilities at arbitrary locations, thus enabling more information for accurate localization without expanding the ground-image database. We also propose a ranking-based algorithm to learn location-discriminative feature projection matrices that result in further improvements in accuracy. We evaluate our method on the Malaga [2] and KITTI [9] datasets and demonstrate significant improvements.

1 Introduction



(a) ground image (b) satellite image

Figure 1: Given a ground image , our method outputs the vehicle location (blue) on the satellite image, along the known vehicle path (orange).

Autonomous vehicles have recently received a lot of attention in the research community. Accurate estimation of a vehicle's location is a key capability to realizing autonomous operation. Satellite imagery provides an alternative, readily-available source of information that can be employed as a reference for vehicle localization. In this paper, we are interested in transferring between ground and satellite image modalities. We present a system that takes as input a stereo ground image acquired by a vehicle, and returns its location in a geo-referenced satellite image (Fig. 1), assuming access to a database of ground (stereo) and satellite images of the environment.

Previous work has achieved good results on visual place recognition and large scale geo-localization [10, 5, 12, 14, 4]. However, the problem of determining the location of a ground image in a satellite image, with the focus in precision, hasn't been intensively explored yet. Cummins and Newman [5] and Sünderhauf et al. [14] describe methods for ground-to-ground visual place recognition. Their methods are limited to locations associated with the geo-tagged database. Hays and Efros [10] and Lin et al. [12] address the problem of identifying the location of a ground image over impressively large areas using a satellite image, where the focus is on scalability across different regions. Viswanathan et al. [15] and Chu and Vu [4] localize the ground image by matching its orthographic texture pattern with the satellite image. These latter approaches perform well, but rely on the existence of clear, non-occluded orthographic information.

Our approach learns to transfer location information between ground and satellite images, and estimates the co-occurrence probability of a query ground image and a local satellite image at a particular location. In this way, our approach uses readily available satellite images for localization, which improves accuracy without requiring a dense database of ground images. We also propose a ranking-

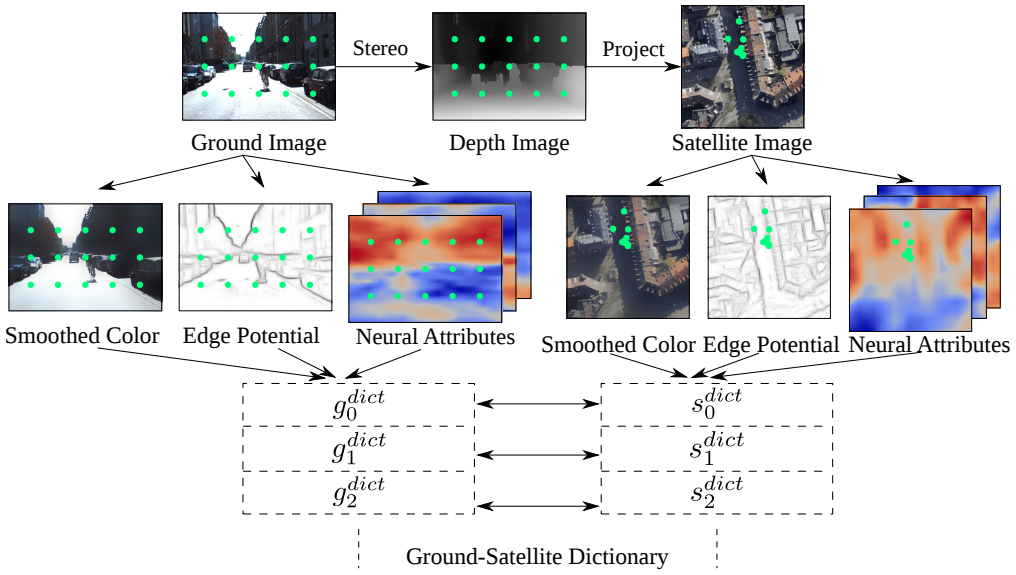


Figure 2: A depiction of the ground and satellite feature dictionary learning process.

based algorithm to learn feature projection matrices that increase the feature's discriminating power in terms of location, thus further improving the localization accuracy. The novel contributions of this paper are: 1) We propose a strategy for localizing a camera based upon an estimate of the ground image-satellite image co-occurrence, which improves localization accuracy without ground image database expansion. 2) We describe a ranking-based method that learns general feature projection matrices that further improve the accuracy by increasing the location-discrimination of features.

2 Approach

2.1 Ground-Satellite Image Feature Dictionary Construction

The ground-satellite dictionary captures the relationship that exists between feature-based representations of ground images and their corresponding overhead view. Specifically, we use three types of features: 1) Pixel-level RGB intensity, which we smooth using bilateral filtering to preserve color transitions [3]. 2) Edge potentials for which we use a structured forest-based method [7] that is robust to non-semantic noise and can be computed in real-time. 3) Neural semantic attributes that are also pixel-wise dense. For these, we use fully-convolutional neural networks [13] trained on ImageNet [6] and fine-tuned on PASCAL VOC [8].

For each ground image in the database, we identify the corresponding satellite image centered on and oriented with the ground image pose. We then compute pixel-wise features for both images. Next, we compute the ground image features $\{g_i^{dict}\}$ on a fixed-interval 2D grid, and project them onto the satellite image using the depth obtained via image stereo [16]. Points that are outside the satellite image are rejected. We record the satellite features corresponding to the remaining projected points ($\{s_i^{dict}\}$). We repeat this to form our one-to-one ground-satellite feature dictionary. We store the dictionary with two k-d trees for fast retrieval. Figure 2 illustrates this process.

2.2 Location-Discriminative Projection Learning

The goal of learning projection matrices is to identify two linear projections W_g and W_s that transform the features into a space such that features that are close to each other in this projected space also correspond to physical locations that are nearby. We formulate this as optimization over a loss function that expresses the sum of all location distances between each feature point and its nearest neighbor in the projected feature space. For ground images, for example, we have

$$W_g = \underset{W}{\operatorname{argmin}} \sum_i \Delta L(i, k_i^*) = \underset{W}{\operatorname{argmin}} \sum_i \Delta L(i, \underset{k \in \mathcal{N}(i)}{\operatorname{argmin}} f_g(i, k, W)) \quad (1)$$

Algorithm 1: Learning a projection matrix

Input: $\{g_i^{dict}\}$, $\{L(i)\}$
Output: W

Initialize $W = \mathbb{I}$ and $t = 0$
for epi=1:MAXITER **do**
  for each i **do**
    $t = epi \times i_{max} + i$
    if $f_{i,k^*} - \min_{k \in \mathcal{N}(i)} \Big(f_{i,k} - m_{i,k} \Big) > 0$ **then**
      $\partial l_t \leftarrow \partial \Big(f_{i,k^*} - \min_{k \in \mathcal{N}(i)} f_{i,k} \Big) / \partial W$
      $\Delta W_t \leftarrow \text{ADAM}(\{\partial l_0, \ldots, \partial l_t\})$ [11]
      $W \leftarrow W - \Delta W_t$
    end if
  end for
end for

where $\Delta L(i,k)$ is the location distance between two feature points, $\mathcal{N}(i)$ is the neighborhood around the feature i in feature space, and $f_g(i,k,W) = \|W g_i^{dict} - W g_k^{dict}\|_2$ (short as $f_{i,k}$ for simplicity). A similar definition is used for W_s. The objective of projection is that feature pairs that are closest in embedding space are also closest in location. This leads us to solving a ranking problem that is equivalent to Eqn. 1, defined by hinge-loss ℓ as below:

$$\ell = \sum_i \Big(f_{i,k_i^*} - \min_{k \in \mathcal{N}(i)} \Big(f_{i,k} - m_{i,k} \Big) \Big)_+ \quad (2)$$

where $m_{i,k} = \Delta L(i,k) - \Delta L(i,k_i^*)$, and $(x)_+ = \max(0,x)$. Intuitively, we would like f_{i,k_i^*} to be smaller than any other $f_{i,k}$ by a margin $m_{i,k}$. We minimize the loss function (2) using stochastic gradient decent with Adam [11] as the weight update algorithm. Algorithm 1 describes the process, it is repeated twice for both W_g and W_s.

2.3 Localization

In localization, we compute the probability $P(L|I^q)$ that a given query ground image I^q was taken at a particular position and orientation L, where we interpolate the database locations on the vehicle path to get a larger number of candidates for L (Figure 3). In order to compute this probability, we first extract features for the query image I^q and then retrieve the precomputed dense features for the satellite image associated with L. Next, we sample the query ground image features with a 2D grid, and their corresponding satellite image features using the query stereo image. After rejecting points that lie outside the satellite image, we obtain a set of ground-satellite feature pairs, where the n^{th} pair is denoted as (g_n^q, s_n^L).

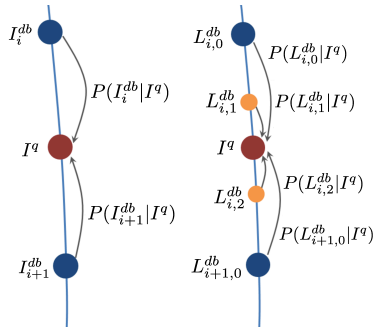
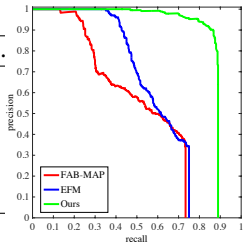


Figure 3: Left: localization by image-to-image matching. Right: by estimating the ground-satellite co-occurrence, our method yields a more fine grained distribution over the camera's location. I_q and I_i^{db} denote the query and database ground images. $L_{i,0}^{db}$ denotes database image location. $L_{i,1}^{db}$ and $L_{i,2}^{db}$ are interpolated locations along the vehicle path. Our method is able to evaluate localization possibilities of $L_{i,1}^{db}$ and $L_{i,2}^{db}$ without knowing their ground images.

For each feature pair, we evaluate their co-occurrence score according to the size of the intersection of their respective database neighbor sets in the projected feature space. To do this, we first transform the features as $W_g g_n^q$ and $W_s s_n^L$. Next, we retrieve the M-nearest neighbors in the transformed dictionary, each for the transformed ground and satellite images, using Approximate Nearest Neighbor [1]. The retrieved neighbor index sets are denoted as $\{id_g^m(W_g g_n^q)\}$ and $\{id_s^m(W_s s_n^L)\}$, and the Euclidean feature distances are denoted as $\{d_g^m(W_g g_n^q)\}$ and $\{d_s^m(W_s s_n^L)\}$. A single pair co-occurrence score $S(s_n^L|g_n^q)$ is computed as the consistency between the two retrieved sets $\sum_{(m_1,m_2)\in I} \Big(d_g^{m_1}(W_g g_n^q) \cdot d_s^{m_2}(W_s s_n^L) \Big)^{-1}$ where $I = \{id_g^m(W_g g_n^q)\} \cap \{id_s^m(W_s s_n^L)\}$ denotes all the (m_1, m_2) pairs that are in the intersection of the two sets. We then compute the desired probability over the location L for the query image I^q as $P(L|I^q) \propto \sum_n \text{score}(s_n^L|g_n^q)$. We determine the final location as that where $P(L|I^q)$ is maximized.

Figure 4: Example images and data split. Left to right: KITTI-City, KITTI-Residential, and Malaga. Yellow, cyan, and purple denote the database, revisit query, and the outside query set, respectively.

Method	KITTI-City	KITTI-Resi.
FAB-MAP	1.24 (0.69)	2.29 (1.55)
EFM	0.87 (0.15)	1.18 (0.91)
Ours-GO	0.81 (0.07)	1.13 (0.81)
Ours-NP	0.41 (0.20)	0.62 (0.33)
Ours-full	**0.39** (0.22)	**0.42** (0.20)



Method	Precision	Recall
FAB-MAP	49.3%	61.6%
EFM	89.4%	43.6%
Ours	**90.2%**	**86.6%**

Figure 5: Left to right: KITTI error in meters; Malaga precision-recall curve; Malaga optimal precision-recall.

3 Experimental Results

We evaluate our method on the widely-used KITTI [9] and Malaga-Urban [2] datasets. We conduct two experiments on the KITTI dataset. First, we use five raw data sequences from KITTI-City, which include 822.9 m driving distance. We randomly select 40% of the images as the database image set, and use the rest as the query image set. Second, we consider the scenario in which the vehicle initially passes an environment and later uses the resulting database for localization upon a subsequent revisiting. We use a long sequence from KITTI-Residential, which has 3080 m driving distance (233 m for the query revisiting). For Malaga-Urban we adopt the setup similar to KITTI-Residential. In addition, we also set aside images taken from a path outside the database to evaluate the ability to handle negative queries. We use Malaga-10, which includes 6.08 km driving distance (583.5 m and 534.3 m as the inside and outside query set). Figure 4 shows example images and the data split. Unlike the KITTI datasets, the quality of the ground-truth location tags in the Malaga dataset is relatively poor. Thus, we evaluate the ability to localize the camera within 10 m of the ground-truth location as opposed to localization error.

We compare two previous methods and three variations of our method: 1. The well-known FAB-MAP [5] with SURF features and cluster-size of 0.45. 2. Exhaustive Feature Matching (EFM) that exhaustively matches all SURF features for image retrieval, with RANSAC-based geometric check. 3. Ours-Ground-Only (Ours-GO) that doesn't consider satellite images and instead performs ground image retrieval using our image features (as opposed to SURF). 4. Ours-No-Projection (Ours-NP) that uses two identity matrices instead of the learned matrices. 5. Lastly, our full method.

The KITTI results are shown in Figure 5. For Malaga we define true positives as images that are identified as inliers and localized within 10 m of their ground-truth locations, Figure 5 shows the precision-recall curve. We pick optimal thresholds based on optimal square area under curves, Figure 5 shows the resulting statistics. Ours-GO achieves lower error than the two SURF-based methods, which shows the effectiveness of our proposed features at discriminating between ground images that has significant overlapping. Ours-NP further reduces the error by interpolating the trajectory (as in Figure 3) and evaluating ground-satellite co-occurrence probabilities, which brings in more localization information. Ours-full achieves the lowest error, which demonstrates the effectiveness of the learned location-discriminative projection matrices. Note that on Malaga, the average (std.) location errors in meters when rough localization succeeds for FAB-MAP, EFM, and ours are $3.45\,(2.16)$, $3.65\,(2.21)$, and $3.33\,(2.08)$, respectively. Although our method achieves better accuracy, it is difficult to draw strong conclusions due to the deficiency in the ground-truth locations. We believe the improvement in accuracy of our method will be more significant if accurate ground-truth location tags are available, similar to what we have observed in our KITTI experiments.

References

[1] Arya, S., Mount, D. M., Netanyahu, N. S., Silverman, R., and Wu, A. Y. (1998). An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923.

[2] Blanco, J.-L., Moreno, F.-A., and González-Jiménez, J. (2014). The málaga urban dataset: high-rate stereo and lidars in a realistic urban scenario. *Int'l J. of Robotics Research*, 33(2):207–214.

[3] Chaudhury, K. N. (2013). Acceleration of the shiftable algorithm for bilateral filtering and nonlocal means. *IEEE Trans. on Image Processing*, 22(4):1291–1300.

[4] Chu, H. and Vu, A. (2015). Consistent ground-plane mapping: A case study utilizing low-cost sensor measurements and a satellite image. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*.

[5] Cummins, M. and Newman, P. (2011). Appearance-only slam at large scale with fab-map 2.0. *Int'l J. of Robotics Research*, 30(9):1100–1123.

[6] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[7] Dollár, P. and Zitnick, C. L. (2013). Structured forests for fast edge detection. In *Proc. Int'l Conf. on Computer Vision (ICCV)*, pages 1841–1848.

[8] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2009). The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

[9] Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: the KITTI dataset. *Int'l J. of Robotics Research*.

[10] Hays, J. and Efros, A. A. (2008). IM2GPS: Estimating geographic information from a single image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK.

[11] Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proc. Int'l Conf. on Learning Representations (ICLR)*.

[12] Lin, T.-Y., Belongie, S., and Hays, J. (2013). Cross-view image geolocalization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 891–898, Portland, OR.

[13] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[14] Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., and Milford, M. (2015). Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free. In *Proc. Robotics: Science and Systems (RSS)*, Rome, Italy.

[15] Viswanathan, A., Pires, B. R., and Huber, D. (2014). Vision based robot localization by ground to satellite matching in gps-denied situations. In *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS)*, pages 192–198.

[16] Yamaguchi, K., McAllester, D., and Urtasun, R. (2014). Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 756–771.