Accurate Vision-based Vehicle Localization using Satellite Imagery

Hang Chu Hongyuan Mei Mohit Bansal Matthew R. Walter

***Abstract*— In this paper, we present a method for accurately localizing a vehicle with the aid of satellite imagery. Our approach takes a ground image as input, and outputs the vehicle's corresponding location on a geo-referenced satellite image. We localize the vehicle by estimating the co-occurrence probabilities between the ground and satellite images, based on a ground-satellite feature dictionary. This method allows us to estimate location probability at an arbitrary location, thus enabling more information for accurate localization without expanding the ground image database. We also propose a ranking-based algorithm that learns a location-discriminative feature projection matrix that results in further improvements in accuracy. We evaluate our method on the Malaga [1] and KITTI [2] public datasets and demonstrate significant improvements over an exhaustive baseline.**

I. Introduction

Autonomous vehicles have recently received a lot of attention in the robotics, intelligent transportation, and artificial intelligence research communities. Accurate estimation of vehicle location is one of the core problems that need to be solved in order to make autonomous driving a reality. Currently, may vehicles employ Global Positioning System (GPS) receivers to estimate their absolute, geo-referenced pose. However, most commercial GPS systems suffer from limited precision and are sensitive to multipath effects (e.g., in the so-called "urban canyons" between tall buildings), which can introduce significant biases that are difficult to identify. Visual place recognition, i.e., the ability to recognize where the vehicle is in a known environment using vision, seeks to overcome this (typically in combination with map-based localization, which uses visual recognition for loop-closure.) Visual recognition, however, is a challenging task due to the appearance variations that result from environment and perspective changes (e.g., parked cars no longer present, illumination changes, weather variations) and the perceptual ambiguity that comes with different areas having similar appearance. A number of techniques have been proposed of late that have made significant progress towards overcoming these challenges [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15].

Satellite imagery is an alternative, useful information source for vehicle localization. With the improved sophistication of online map services, such as Google and Bing Maps, high-resolution satellite image databases have become increasingly accessible and well-maintained. Due to this recent ease of obtaining satellite imagery, multiple approaches have been proposed to utilize satellite imagery information for accurate localization [16, 17].

H. Chu, H. Mei, M. Bansal, and M.R. Walter are with the Toyota Technological Institute at Chicago, Chicago, IL 60637, USA {hchu,hongyuan,mbansal,mwalter}@ttic.edu

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

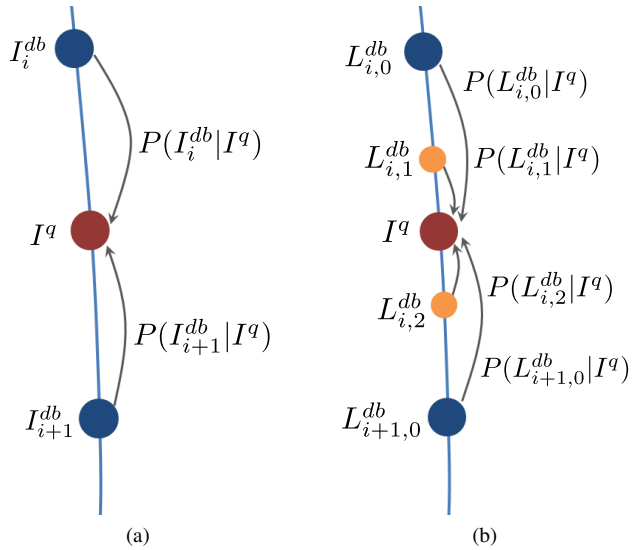
In this paper, we present a method that performs accurate vision-based vehicle localization with the aid of satellite imagery. Our system takes as input a stereo ground image acquired outdoors and returns the location of the stereo pair in a geo-referenced satellite image, assuming access to database of ground (stereo) and satellite images of the environment (e.g., such as those acquired during a previous environment visit). This is illustrated in Figure 1. Instead of matching the query ground image against the database of ground images, as is typically done for visual place recognition, we estimate the co-occurrence probability of the query ground image and the local satellite image of a certain location, i.e., the probability of a certain location given the query ground image being observed. This allows us to evaluate location probability on more locations by interpolating the vehicle path with sampled local satellite images. Our approach is able to use readily available satellite images for localization, which improves accuracy without requiring a dense database of ground images. We also propose a listwise ranking algorithm to learn effective feature projection matrices that in turn learn location-discriminative feature patterns, and thus further improve the localization accuracy.

The novel contributions of this paper are:

- We propose to localize the vehicle by estimating the ground image-satellite image co-occurrence, which improves localization accuracy without ground image database expansion.
- We propose to further improve the localization accuracy by learning feature space projections, which can be solved effectively by a ranking-based algorithm.

II. Related Work

The problem of visual place recognition has received a great deal of attention in the robotics and vision communities [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. The biggest challenges to visual place recognition arise due to variations



(a) (b)

Fig. 2. (a) Localization by image-to-image matching, where two information sources are available for estimating the location. (b) Localization by estimating ground-satellite image co-occurrence, which is able to estimate possibilities at arbitrary location, thus enables more information sources without expanding the ground image database. I_q denotes the query ground image. I_i^{db} denotes a database ground image, and its location $L_{i,0}^{db}$. $L_{i,1}^{db}$ and $L_{i,2}^{db}$ are the 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.

in image appearance that result from changes in viewpoint, environment structure, and illumination, as well as perceptual aliasing, which are both typical of real-world environments.

Much of the work seeks to mitigate some of these changes by using interest point detectors and descriptors like SIFT [18] and SURF [19], which are robust to transformations in scale and rotation, as well as slight variations in illumination. Place recognition then follows as image retrieval, i.e., image-to-image matching search against this database (with various methods to improve efficiency) [20, 21, 22, 23]. While these methods have demonstrated reasonable performance under appearance variations, they are prone to failure when the environment is perceptually aliased. Under these conditions, feature descriptors are no longer sufficiently discriminative, which results in false matches (notably, when the query image corresponds to a environment location that is not in the map). When used for data association in a downstream SLAM framework, these erroneous loop closures can result in estimator divergence.

The FAB-MAP algorithm by Cummins and Newman [3, 5] is designed to address challenges that arise as a result of perceptual aliasing. To do so, FAB-MAP learns a generative model of region appearance using a bag-of-words representation that expresses the commonality of certain features. By effectively modeling this perceptual ambiguity, the authors are able to reject invalid loop closures despite significant aliasing, while correctly recognizing valid loop closures. Alternatively, other methods achieve greater robustness to perceptual aliasing and appearance variations by treating visual place recognition as problem of matching image sequences [24, 7, 8, 9, 10], whereby joint consistency helps to reduce the likelihood of false matches. The robustness of image retrieval methods can be further improved by increasing the space of appearance variations spanned by the database [25, 26]. However, approaches that achieve invariance proportional to the richness of their training necessarily require larger databases to achieve robustness. Our method is similar to several of these methods [3, 5] in that we learn a probabilistic model for matching, though ours is over the location of the query in the geo-referenced satellite image. Rather than treating localization as retrieval whereby we find the nearest map image for a given query, we instead leverage the availability of satellite images to estimate the interpolated position. This provides additional robustness to viewpoint variation, particularly that which results from increased separation between the database images (e.g., on the order of 10 m).

Meanwhile, recent attention in visual place recognition has focused on the particularly challenging problem of identifying matches in the event that there are significant appearance variations due to large illumination changes (e.g., matching a query image taken at night to a database image taken during the day) and seasonal changes (e.g., matching a query image with snow to one taken during summer) [25, 12, 27, 28, 14, 15]. While some of these variations can be captured with a sufficiently rich database [25], this comes at the cost of requiring a great deal of training data and its ability to generalize is not clear [14]. McManus et al. [26] seek to overcome the brittleness of point-based features to environment variation [29, 30, 7, 10] by learning a region-based detector to directly achieve improved invariance to appearance changes. Alternatively, Sünderhauf et al. [14] build on the recent success of deep convolutional networks and eschew traditional features in favor of features that can be learned from large corpora of images [31]. Their framework first detects candidate landmarks in an image using state-of-the-art proposal methods, and then employs a convolutional network to generate features for each landmark. These features provide a robustness to appearance and viewpoint variations that enables accurate place recognition under challenging environmental conditions. We also take advantage of this recent development in convolutional neural networks for image segmentation, which provides us with effective pixel-wise semantic feature extractors.

Related to our approach of projection matrix learning with a ranking loss is a long history of machine learning research with ranking loss objective functions [32]. Our approach and task is also related to the field of multi-view learning [33], since we work with the ground image and satellite image views and try to project them to a shared space.

In similar fashion to our work, previous methods have investigated vehicle localization with an input ground image and a satellite image reference[16, 17]. These methods focus on extracting orthographical texture patterns, typically road lane markings on the ground plane, and then matching these observed patterns with the satellite image. Although these methods show encouraging localization performance, their

methods rely on the existence of clear, non-occluded lane mark information.

III. Approach

Our framework has three primary components: constructing the ground-satellite image feature dictionary, learning location-discriminative projections for both ground and satellite image features, and estimating the probability of a arbitrary location given a query ground image. We next describe these three steps in detail.

A. Ground-Satellite Image Feature Dictionary

We use three types of features. The first type is smoothed color intensities. We use bilateral filtering to smooth the image while preserving sharp color transitions, and employ a publicly available implementation [34] that operates $O(1)$ time. The second feature type is edge potentials. We use the structured forest-based implementation [35], which is not only more robust to non-semantic noise (as compared to classic gradient-based edge detectors), but can also be computed in real-time. The third feature type we use is neural, pixel-wise, dense, semantic attributes. For these, we use fully-convolutional neural networks [36] trained on ImageNet [37] and fine-tuned on PASCAL VOC [38].

For each ground image in the database, we identify the corresponding satellite image whose center position and orientation matches that of the ground image. We then compute pixel-wise features for both images. Next, we sample the ground image features with a fixed-interval 2D grid, and project these sampled points onto the satellite image using the image depth obtained via image stereo [39]. Sampled points that are outside the satellite image region are rejected. We repeat this process for all ground images in the database and the ground-satellite feature pairs of all accepted points to form our one-to-one ground-satellite image feature dictionary. We store the dictionary with two k-d trees (multidimensional binary search trees) for fast retrieval. Figure 3 illustrates our dictionary construction process. We denote the i^{th} entry of the dictionary as (g_i^{dict}, s_i^{dict}).

B. Location-Discriminative Projection Learning

The goal of location-discriminative projection learning 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 are also close to each other in actual location. We can express this learning task 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, i.e.,

$$W_g = \arg\min_{W} \sum_i \Delta L \left(i, \arg\min_{k \in \mathcal{N}(i)} f_g(i,k,W) \right) \quad (1)$$

where $\Delta L(i,k)$ denotes the location difference between two feature points, $\mathcal{N}(i)$ denotes the neighbourhood of the i^{th} feature, and $f_g(i,k,W) = \|W g_i^{dict} - W g_k^{dict}\|_2$. We have empirically found that a neighborhood size of $\mathcal{N}(i) = 20$ to be effective. A similar definition is used for W_s.

Algorithm 1: Learning location-discriminative projection via ranking

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

Initialize $W = \mathbb{I}$ and $t = 0$
for epi=1:MaxIter **do**
  for each i **do**
    $t = epi \times i + i$
    if $f_g(i,k^*,W) - \min_{k \in \mathcal{N}(i)} \Big(f_g(i,k,W) - m(i,k) \Big) > 0$ **then**
      Compute ∂l_t as $\partial \Big(f_g(i,k^*,W) - \min_{k \in \mathcal{N}(i)} f_g(i,k,W) \Big) / \partial W$
      Compute ΔW_t as $Adam(\{\partial l_0, \partial l_1, \ldots, \partial l_t\})$ [40]
      Update $W \leftarrow W - \Delta W_t$
    end if
  end for
end for

We want to select the location $k^* = \arg\min_k \Delta L(i,k)$ that is equivalent to solving a ranking problem defined by hinge-loss l as below:

$$\sum_i \left(f_g(i,k^*,W) - \min_{k \in \mathcal{N}(i)} \Big(f_g(i,k,W) - m(i,k) \Big) \right)_+ \quad (2)$$

where $m(i,k) = \Delta L(i,k) - \Delta L(i,k^*)$, $(x)_+$ gives x if $x > 0$, and 0 otherwise. Intuitively, we would like $f_g(i,k^*,W)$ to be smaller than any other $f_g(i,k,W)$ by a margin $m(i,k)$. We minimize the loss function in Equation (2) using stochastic gradient decent with Adam [40] as the weight update algorithm. The complete projection learning algorithm is described in Algorithm 1.

C. Localization

In our final localization step, we compute the probability of a certain locaiton-orientation given an observed ground image. More specifically, we estimate $P(L|I^q)$, where I^q and L denote the query ground image and the location-orientation that is being evaluated, respectively. To do this, we first obtain the satellite image associated with location L, and compute dense features for both this satellite image and the query ground image I^q. Next, we sample the query ground image features with a 2-d grid, and find their corresponding satellite image features by projecting using the query stereo image, as if I^q was centered and oriented at L. After rejecting samples that are projected outside the satellite image range, we obtain a set of ground-satellite image feature pairs, where the n^{th} pair is denoted as (g_n^q, s_n^L).

For each pair of ground-satellite image features, 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 compute the

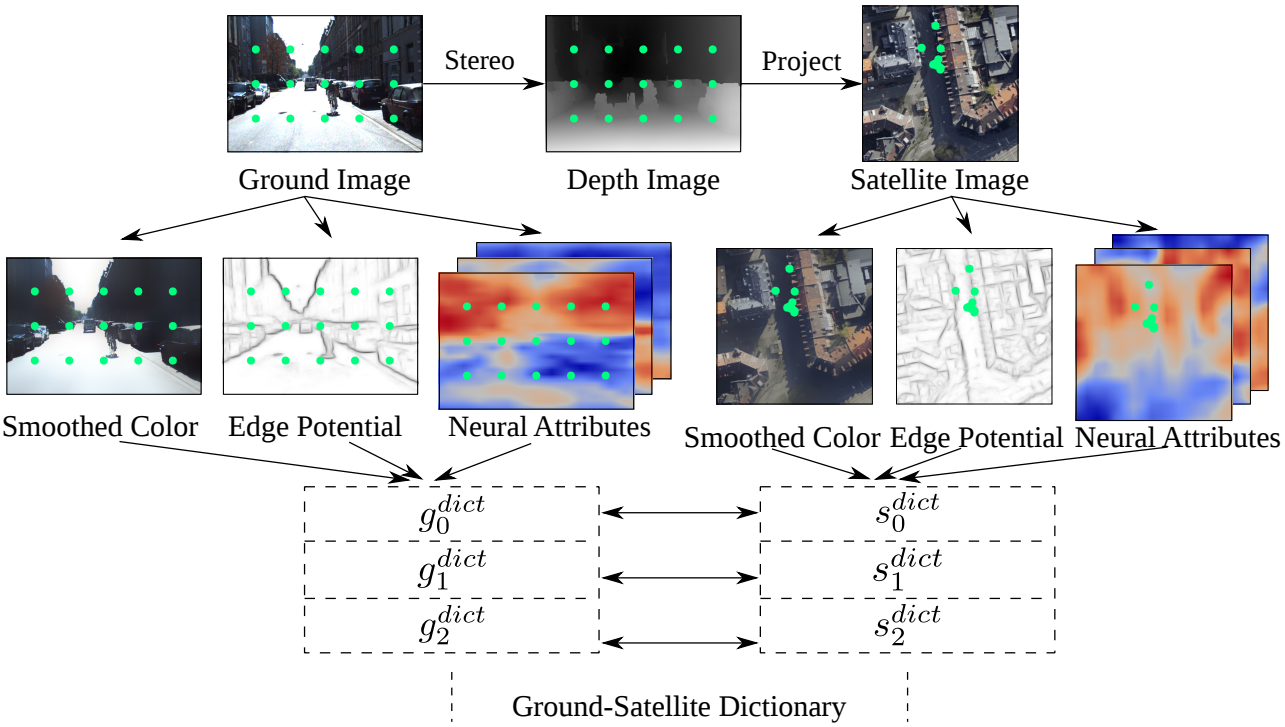


Fig. 3. Construction process for our ground-satellite image feature dictionary.

projected features $W_g g_n^q$ and $W_s s_n^L$. Next, we retrieve their M-nearest neighbors in a similarly projected feature database with the Approximate Nearest Neighbor algorithm [41]. 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 is computed as the consistency between the two retrieved sets:

$$score(s_n^L|g_n^q) = \sum_{(m_1,m_2)\in int} \left(d_g^{m_1}(W_g g_n^q) d_s^{m_2}(W_s s_n^L)\right)^{-1} \tag{3}$$

where int denotes all the (m_1, m_2) pairs that are in the intersection $\{id_g^m(W_g g_n^q)\} \cap \{id_s^m(W_s s_n^L)\}$. Finally, the probability of L given I^q is observed is

$$P(L|I^q) = \frac{1}{C} \sum_n score(s_n^L|g_n^q) \tag{4}$$

where C is the normalizing factor. We interpolate the database vehicle path with $L_{i,j}^{db}$ (as we showed in Figure 2) and determine the final location as the location where $P(L|I^q)$ is maximized.

IV. Experimental Results

We evaluate our method on two widely-used publicly available datasets: the KITTI dataset [2], and the Malaga-Urban dataset [1].

A. KITTI Dataset

We conduct two experiments on the KITTI dataset. In the first experiment, we use 5 raw data sequences from the KITTI-City category. The total number of images is 1067, each with a resolution of 1242×375. The total driving distance for these sequences is 822.9m. We randomly select 40% of these images as the database image set, and the rest as the query image set. Examples of the used KITTI-City data are shown in Figure 4.

In the second KITTI experiment, we consider a more practical scenario: a probe vehicle goes through an area and builds the ground image database and then this database is used to localize other vehicles. We simulate this scenario using a long sequence under the KITTI-Residential category, where we use the ground images of the vehicle passing a certain location for the first time as the database set, and ground images of that vehicle passing for the second time as the query set. We sample the database and query set by approximately one image per second. The total number of images is 3654, each with resolution 1241×376. The total driving distance of the used data is $3.08km$. Figure 5 shows example ground images and the database query data split.

We compare five different methods, including two baselines (previous work) and three variations of our model, the latter as a means of ablating our framework. One baseline that we consider is a publicly available implementation of FAB-MAP [5][1], where we used the SURF feature detector and descriptor with a cluster-size of 0.45, which yields 2611 and 3800 bag-of-words for the two experiments, respectively. We consider a second baseline that performs exhaustive matching over a dense set of SURF features for image retrieval, which we refer to as Exhaustive Feature Matching

[1]https://code.google.com/p/openfabmap/

Fig. 4. Examples images (top) from the KITTI-City dataset, with a satellite image (bottom) annotated with the vehicle path and arrows denoting the location of the images.

Fig. 5. Examples images (top) from the KITTI-Residential dataset with a satellite image (bottom) annotated with the vehicle path. Green arrows denote the location of the images, yellow indicates the database trajectory, and cyan denotes the second visit, which is used as the query set.

(EFM). We further refine these matches using RANSAC-based [42] image-to-image homography estimation to identify a set of geometrically consistent inlier features. We use the average feature distance over these inliers as the final measurement of image-to-image similarity. The third method is based on using our proposed image features instead of SURF (without using the satellite image), which we refer to as Ours-Ground-Only (Ours-GO). The fourth method consists of our proposed framework with satellite information, but without learning the location-discriminative feature projection, which we refer to as Ours-No-Projection (Ours-NP). The last method is our full proposed framework, with satellite information and where we use all images on the database path for projection learning. The experimental results are shown in Table I.

From Table I, we see that our method and its variations

TABLE I

LOCATION ERROR AND STANDARD DEVIATION IN METERS

Method	KITTI-City	KITTI-Residential
FAB-MAP [5]	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)

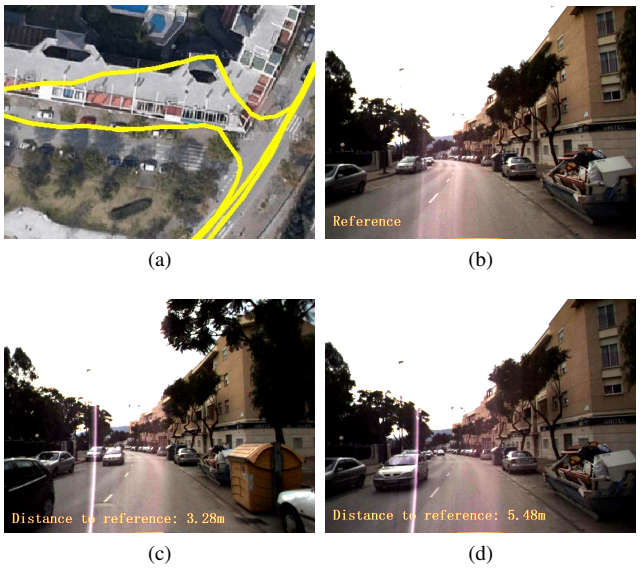
Fig. 6. Data split for the Malaga sequence where yellow, cyan, and purple denote the database, revisit query, and the outside query set, respectively.

outperform the FAB-MAP [5] and EFM methods. Ours-GO outperforms these two SURF-based methods, which shows the effectiveness of discriminating ground images using our proposed features. Ours-NP achieves further improvements by interpolating the trajectory between two adjacent ground database images and evaluating ground-satellite co-occurrence probabilities, which brings in more localization information. Ours-full achieves the best accuracy, which demonstrates the effectiveness of learning the location-discriminative projection matrix learned by our ranking-based algorithm. Note that Ours-NP and Ours-full use stereo information which is not used by FAB-MAP, though we have found that the improvement is largely due to our method's use of satellite imagery.

B. Malaga-Urban Dataset

We also evaluate our framework on the Malaga-Urban dataset [1], where we adopt the setup similar to KITTI-Residential: using the first vehicle pass of an area as the database set, and the second visit as the query set. In addition, we also set aside one part of the vehicle path as neither part of the database nor the query, which is treated as "outside" images and should be classified as such by the evaluated method. We used the longest sequence Malaga-10, which contains 18203 images, each with a resolution of 1024×768. We down-sample the database and query sets at approximately one frame per second. The total driving distance is 6.08km, with 4.96km as the database set, 583.5m as the inside query set, and 534.3m as the outside query set. The dataset split is shown in Figure 6.

Unlike the KITTI datasets, the quality of ground-truth



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

Fig. 7. Deficiency of ground-truth location tags in the Malaga dataset.

TABLE II
OPTIMAL PRECISION-RECALL OF DIFFERENT METHODS.

Method	Precision	Recall
FAB-MAP [5]	50.0%	55.6%
EFM	92.3%	66.7%
Ours	**92.9%**	**72.2%**

location tags in the Malaga dataset is relatively poor. Examples of ground-truth location tag deficiencies are shown in Figure 7. Figure 7(a) shows a part of the ground-truth trajectory, where some location tags are clearly marked in a non-drivable area. Figure 7(b)-(d) further demonstrate issues with the ground-truth locations. Treating Figure 7(b) as the reference, the image in Figure 7(d) looks more similar to Figure 7(b), than that in Figure 7(c) (note the white trash-can on the bottom right). However, Figure 7(d) is actually farther from the reference (5.48m) than the image in Figure 7(c) (3.28m) according to the ground-truth location tags. Due to this issue in the dataset, it is more suitable for place recognition performance evaluation (i.e., rough localization of within 10m to the ground-truth) rather than precise accuracy evaluation.

We again compare our algorithm to the same FAB-MAP [5] and Exhaustive Feature Matching (EFM) methods as with the KITTI experiments. We set the bag-of-words size for FAB-MAP to 2589. We define true positives as inside images that not only have been correctly identified as revisits, but have also been localized to within 10m of their ground-truth locations. We picked optimal thresholds for all the methods based on optimal square area under the roc curve; the precision and recall using this optimal threshold are shown in Table II.

It can be seen from Table II that our method is more effective than other methods both at determining inside versus outside images, and at localizing the inside images at the correct location. The EFM method achieves a comparable precision score, however the computational expense of doing exhaustive feature matching makes it intractable for real-time use in all but trivially small environments. Note that using only inside images, the average location error (standard deviation) 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 aforementioned 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.

V. FUTURE WORK

The usage of neural features in our current method is still rather simplistic. We believe with careful feature learning, we can obtain more effective semantic segmentors that are specific to outdoor driving environments, which enables learning the feature projection matrices with fewer amount of data, and learning general ground-to-satellite knowledges across different places. This is the problem where we will focus our future work on.

VI. CONCLUSION

We presented a method for accurately localizing a vehicle with the aid of satellite imagery. Our approach takes a ground image as input, and outputs the vehicle's corresponding location on a geo-referenced satellite image. We proposed to estimate the co-occurrence probabilities between the ground and satellite images, and also a ranking-based algorithm that learns a location-discriminative feature projection matrix that results in further improvements in accuracy. We evaluated our method on multiple public datasets.

REFERENCES

[1] J.-L. Blanco, F.-A. Moreno, and J. González-Jiménez, "The málaga urban dataset: high-rate stereo and lidars in a realistic urban scenario," *Int'l J. of Robotics Research*, vol. 33, no. 2, pp. 207–214, 2014.

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

[3] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, June 2008.

[4] M. Park, J. Luo, R. T. Collins, and Y. Liu, "Beyond GPS: Determining the camera viewing direction of a geotagged image," in *Proc. Int'l Conf. on Multimedia, (ACM MM)*, 2010, pp. 631–634.

[5] M. Cummins and P. Newman, "Appearance-only slam at large scale with fab-map 2.0," *Int'l J. of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[6] W. Churchill and P. Newman, "Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation," in *Proc. IEEE Int'l Conf. on Robotics*

and Automation (ICRA), Saint Paul, MN, May 2012, pp. 4525–4532.

[7] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, Saint Paul, MN, May 2012, pp. 1643–1649.

[8] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: Appearance-based localisation throughout the day," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, Karsruhe, Germany, May 2013, pp. 3212–3218.

[9] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons," in *Proc. Work. on Long-Term Autonomy at ICRA*, Karsruhe, Germany, May 2013.

[10] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Proc. Nat'l Conf. on Artificial Intelligence (AAAI)*, 2014.

[11] S. Lynen, M. Bosse, P. Furgale, and R. Siegwart, "Placeless place-recognition," in *Proc. Int'l Conf. on 3D Vision (3DV)*, Tokyo, Japan, December 2014, pp. 303–310.

[12] C. McManus, W. Churchill, W. Maddern, A. D. Stewart, and P. Newman, "Shady dealings: Robust, long-term visual localisation using illumination invariance," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, Hong Kong, May 2014, pp. 901–906.

[13] P. Hansen and B. Browning, "Visual place recognition using HMM sequence matching," in *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS)*, Chicago, IL, September 2014, pp. 4549–4555.

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

[15] N. Sünderhauf, F. Dayoub, S. Shirazi, B. Upcroft, and M. Milford, "On the performance of ConvNet features for place recognition," in *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS)*, 2015.

[16] H. Chu and A. Vu, "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)*, 2015.

[17] T. Senlet and A. Elgammal, "A framework for global vehicle localization using stereo images and satellite and road maps," in *Proc. Int'l Conf. on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 2034–2041.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. on Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[19] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. European Conf. on Computer Vision (ECCV)*, 2006, pp. 404–417.

[20] J. Wolf, W. Burgard, and H. Burkhardt, "Robust vision-based localization by combining an image-retrieval system with monte carlo localization," *Trans. on Robotics*, vol. 21, no. 2, pp. 208–216, 2005.

[21] F. Li and J. Kosecka, "Probabilistic location recognition using reduced feature set," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, Orlando, FL, May 2006, pp. 3405–3410.

[22] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, 2007, pp. 3921–3926.

[23] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[24] O. Koch, M. R. Walter, A. Huang, and S. Teller, "Ground robot navigation using uncalibrated cameras," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, Anchorage, AK, May 2010, pp. 2423–2430.

[25] P. Neubert, N. Sünderhauf, and P. Protzel, "Appearance change prediction for long-term navigation across seasons," in *Proc. European Conf. on Mobile Robotics (ECMR)*, Barcelona, Spain, September 2013, pp. 198–203.

[26] C. McManus, B. Upcroft, and P. Newman, "Scene signatures: Localised and point-less features for localisation," in *Proc. Robotics: Science and Systems (RSS)*, Berkeley, CA, July 2014.

[27] W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman, "Transforming morning to afternoon using linear regression techniques," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, Hong Kong, May 2014.

[28] S. M. Lowry, M. J. Milford, and G. F. Wyeth, "Transforming morning to afternoon using linear regression techniques," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, Hong Kong, May 2014, pp. 3950–3955.

[29] C. Valgren and A. J. Lilienthal, "SIFT, SURF and seasons: Long-term outdoor localization using local features," in *Proc. European Conf. on Mobile Robotics (ECMR)*, Freiburg, Germany, September 2007.

[30] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "FAB-MAP + RatSLAM: Appearance-based SLAM for multiple times of day," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, Anchorage, AK, May 2010, pp. 3507–3512.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int'l J. on Computer Vision*, pp. 1–42, 2014.

[32] L. Hang, "A short introduction to learning to rank," *IEICE TRANSACTIONS on Information and Systems*, vol. 94, no. 10, pp. 1854–1862, 2011.

[33] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[34] K. N. Chaudhury, "Acceleration of the shiftable algorithm for bilateral filtering and nonlocal means," *IEEE*

Trans. on Image Processing, vol. 22, no. 4, pp. 1291–1300, 2013.

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

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

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

[38] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

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

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

[41] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *Journal of the ACM (JACM)*, vol. 45, no. 6, pp. 891–923, 1998.

[42] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, June 1981.