Semantically Guided Location Recognition for Outdoors Scenes

Arsalan Mousavian and Jana Košecká and Jyh-Ming Lien

Abstract— **The problem of image based localization has a long history both in robotics and computer vision and shares many similarities with image based retrieval problem. Existing techniques use either local features or (semi)-global image signatures and consider the retrieval either in the context of topological mapping or loop closure detection. The difficulty of the location recognition problem is affected by large appearance and viewpoint variation between the query view and reference dataset and presence of large number of non-discriminative structures due to vegetation, sky and road. In this work we show that semantic segmentation of images leading to labeling of man-made structures can inform the traditional bag-of-visual words models to obtain proper feature weighting and improve the overall location recognition accuracy. We also demonstrate additional capability of identifying individual buildings and estimating their extent in images, providing the essential building block for sub-sequent geo-location. Towards this end we introduce a new challenging outdoors urban dataset exhibiting large variations in appearance and viewpoint.**

I. INTRODUCTION

The problem of image based localization and place recognition entails for each query view retrieving the nearest view from the reference dataset of views. Many variations of this problem have been considered in the past and the existing methods typically differ in the acquisition mode of the query and reference views, the choice of image and reference dataset representation and the associated matching strategy. This task is challenging due to often large changes in appearance due to illumination, viewpoint or season between the query views and the reference dataset and the size and the nature of the environment and the dataset. The two commonly used representations for the problem are either local feature based methods or (semi) global signatures. In computer vision literature, many approaches use the bag-of-visual-words representations, followed by spatial verification. In this setting local features are weighted by a TF-IDF [7] scheme which considers the discriminative nature of the features. The evaluations are typically done on structured datasets as Google street-view with dense sampling of canonical viewpoints or in less structures settings utilizing geo-tagged images from photo-sharing sites such as Panoramio or Flickr. In the context of autonomous navigation, typical evaluation of these methods assumed rather small variations in the viewpoint and moderate variation in appearance as the sequences were acquired from very similar viewpoints along the same driving routes and under similar imaging conditions [3]. Alternative representations have been proposed more recently, which are robust with respect to more significant changes in imaging conditions, such as day and night [15] and seasonal and weather changes [13], but in all these settings the viewpoint variations are very small as the reference dataset model and the test images were were acquired by the same vehicle along the same route. This proposed work aims to extend the resources used to create the reference data for location recognition and geo-location and considers unstructured datasets with large variations in viewpoint and appearance due to seasonal changes. In order to be able to match views with smaller overlap we adopt bag-of-visual-words approach followed by spatial verification. The traditional weighting schemes for local features are however often not adequate, especially in the environment with many repetitive image structures which are not discriminative for locations.

Contribution: We demonstrate that the availability of semantic information about the presence of man-made landmark structures such as buildings, can enhance the traditional local feature based methods when the viewpoint change and appearance changes is more significant. Focusing the matching on man-made structures, helps to discard irrelevant features from the scene which often act as confusers in various voting strategies, such as trees, vegetation and road features are often not discriminative of location. We show (1) that semantic segmentation of images into commonly encountered semantic categories and explicit labeling of man-made structures provides an improvement in retrieval accuracy of traditional bag-of-words methods with spatial verification despite large viewpoint variation between the query views and reference views; (2) we further demonstrate the capability of these local features to categorize individual building instances and estimate their extent in query views. This obtained semantic representation can be exploited further for semantic localization and mapping as well as determining geographic location of the query views.

II. RELATED WORK

The problem of visual place recognition has been studied extensively by many and the existing approaches vary in the proposed image representation and associated matching strategy as well as datasets used for evaluation. In the computer vision literature the approaches often bear similarity with image based retrieval techniques, where the considered baseline method is often the bag-of-visual-words representation, followed by spatial verification of top retrieved images using geometric constraints [17]. Various improvements of this methods include learning better vocabularies, developing bet-

Arsalan Mousavian, Jana Košecká , and Jyh-Ming Lien are with the Computer Science Department, Volgenau School of Engineering at George Mason University, Fairfax, VA 22030, USA. amousavi@gmu.edu, kosecka@cs.gmu.edu, jmlien@cs.gmu.edu.

ter quantization and spatial verification methods [14], [24], [25], employing more sophisticated models of dependence of local features [3] or improving the scalability [19].

In the context of visual place recognition, in contrast to image based retrieval, there is often additional structural information available which can be exploited towards the task. In case the locations and/or landmarks are covered by many views 3D reconstruction techniques can help in quatiftying the 'matchability' of individual features. Authors in [1] chose the stable keypoints which are detectable across many different views and trained a randomized decision tree classifier per keypoint and used these classifiers to classify the new images. In [18] authors built adjacency matrix between reference images and then find the clusters within training images. During the recall, they first classify the image by assigning it to most likely image cluster and then retrieve the score using TF-IDF weighting scheme. In [10] learn the confusion weight for each feature exploiting the geographic location of the unweighted retrieved views and in [8] the authors trained exemplar SVM per image in the training set where the negative set are the images which have similar cosine distance but they are far away.

In robotics setting, the existing approaches use both local feature based methods as well as alternative image signatures. In traditional loop-closure detections problem variations of local features have been used effectively by many [11], [3], typically followed by spatial verification and/or global registration of the entire trajectories. To tackle the challenged posed by large appearance variations due to seasonal or time of the day changes alternative image representations have been deployed. In [15] authors used low resolution patches discriminative for locations. In [13] authors adopt the strategy proposed by [5] and use mid-level patches endowed by HOG descriptors are used accompanied by per-location training procedure which determines which patches are discriminative for the location assuming that large variety of images under different imaging conditions are acquired per location. This approach can effectively select the HOG scene signatures which are then used for pose estimation, but it does require many images of the location under different imaging conditions to learn what aspects of the location are discriminative.

The retrieval approach of [10] is the most similar to our method in selecting the features which are informative for localization and suppressing the ones which are confusing. We instead of learning the weight for each feature using the GPS coordinates, we propose to use general purpose semantic segmentation techniques [20] to classify features to different semantic categories and filter those which do not belong to man-made landmark structures.

While in the current work we focus only on the use of semantic information to guide the location recognition, we also take it one further step to detect the location and extent of man-made landmark structures in query views. This is motivated by recent approaches to semantic localization [2], which use the traditional object detection pipelines [4] to generate hypotheses about presence of objects and their extent and bearing in images, followed by the particle filter based localization. The current work naturally extends the type of semantic information which can be considered for this task. Our final goal is to aid localization and geo-localization using the meta-data associated with the maps along with the images on an autonomous agent engaged in navigation in outdoors environment using visual sensing.

III. PROPOSED METHOD

In the following section we describe the baseline bag-of-words approach, the semantic segmentation component which is used to filter out the confusing features and the strategy for detecting the individual buildings and their extent in the query views. The dataset used in our experiments and additional details of the algorithms can be found in the experiments section.

We use the bag-of-words BoW representation as the core of our method. The dataset of images we consider, is a sparse set of views capturing the majority of landmarks man-made structures on our campus, with small or no viewpoint overlap between the views. Local features are more suitable for these conditions, due to their smaller extent and capability of tolerating larger viewpoint variations. Matching larger signatures for the views which have a small overlap would pose difficulties. Given the reference set of images, the standard BoW method clusters the SIFT features to build vocabulary of visual words. One of the weakness of the BoW representation is that it treats all the features being present in the image with the same weight. TF-IDF weighting [7] tackles this problem by defining two terms. The first one is term frequency which is the frequency of occurrence of that word and the second term is inverse document frequency which is inversely proportional to the number of documents in which each word occurs. Even though TF-IDF improves the performance considerably, it still considers all the visual words present in the image. Therefore, if there are many frequently occurring words with small IDF weights in the query image, the product of the frequency of that word and the corresponding IDF weight would be large. Consequently, it changes the meaning of the the query image representation. We propose to rectify this problem by exploiting the semantic information available in the image.

A. Semantic Segmentation

The semantic labels we consider are five commonly occurring semantic categories in street scenes - *ground, sky, building, car, tree*. Our approach for semantic labeling is based on using a single bottom-up segmentation of the image where the superpixels are characterized with a variety of features including color, texture, location and perspective cues. The labeling is performed using boosting classifiers which automatically compute feature relevance. The proposed semantic segmentation approach is closely related to [9] and [22]. The labeling is done on superpixels obtained by the color based over segmentation scheme proposed in [6]. The evaluation of the performance of the boosting classifier,

Fig. 1. Semantic Segmentation. Left: original image; Right: color coded semantic categories. Detected man-made structure are colored red.

details of the training procedure as well as comparison to the state of the art systems, is described in more detail in [21]. An example of the obtained semantic layout is shown in Fig. 1.

B. Semantically Aware Image Retrieval

In the absence of semantic information, the BoW approach with traditional TF-IDF weighting of visual words often retrieves incorrect images from the reference dataset. For example, Fig 2(a) depicts an example of such situation. As it is shown, the query image contains considerable amount of vegetation and if we use the TF-IDF score to find the top matches, only one of the retrieved images contains the building of interest and the rest is dominated by vegetation. This is the case that in spite of the fact that IDF weight of the visual words capturing the vegetation is small due to frequent occurrences in most of the scenes but since there are many of them in the image, the TF component becomes large as well as the TF-IDF weight. In other words, these features contaminate the TF-IDF weighted BoW representation. Vegetation is not the only category which causes confusion. Other semantic categories such as road, sidewalks can also mislead the feature weighting. Another aspect which has drawn recent attention [13] is the robustness to changes in appearance due to the season change. Since the seasonal changes mostly affect the appearance of vegetation, not considering these features We first build vocabulary of all of the SIFT features using k-means algorithm and then we compute the IDF term for each of the visual words in the vocabulary (Algorithm 1, lines 1 and 2). We compute the semantic segmentation [21] for all images in the reference set labelling each pixel as belonging to one of the 5 semantic categories *man-made structure, sky, grass, road, trees, vegetation*. Given this information we discard the local SIFT features which do not belong to man-made structure regions (Algorithm 1, line 3). In order to compensate for the semantic segmentation errors, we determine whether a local feature is belongs to man-made structure region by considering 7×7 pixel patch around each feature and checking whether at least 50 percent of the pixels in this patch are labeled as man-made structure/building. Other semantic categories can be chosen as well. After semantically-pruning the features which do not belong to buildings, we recompute the TF component for each of the training images and compute new TF-IDF BoW representation (Algorithm 1, line 4). It is worth noting that there might be a visual word which occurs both on a man-made structure and other semantic classes. In this case, we do not eliminate all the occurrences of that visual words. We only eliminate the ones which are not in our desired semantic regions. This is different from [10] where they compute the weight of each visual word for all of its occurrences regardless of the semantic category each feature belongs to and we do not need to have GPS coordinate for each of the images.

In the test phase, we extract the SIFT features from query image and assign each of the SIFT features to one of the visual words in the learned vocabulary. We then compute TF-IDF representation using the IDF and the frequency of visual words in the query image (Algorithm 2, line 1). We retrieve top N matches by finding K-nearest neighbor using the cosine distance between the query feature vector and semantically-pruned feature vectors of training images (Algorithm 2, line 2). Note that we do not discard any of the SIFT features which are not inside of the buildings in the test phase. That is because we have already eliminated their counter part features from the training images which decreases the effect of such features on the final cosine distance significantly. In order to build the dictionary of the visual words, we use all the features regardless of the semantic categories. If we only use features belonging to man-made structures, our dictionary overfits to the features on man-made structures and it does not generalize well to the features present in the query image. Examples of comparison of the nearest view retrieval with and without semantic pruning can be found in figure 2, where left side are the nearest views retrieved with the baseline BoW methods and right column are the nearest views retrieved with semantically weighted BoW methods. Note that the semantic weighting for TF-IDF notable improves the number of correctly retrieved views in top k matches. This superior performance can be also observed after spatial geometric verification stage. The quantitative comparison can be found in Table I in the experimental section.

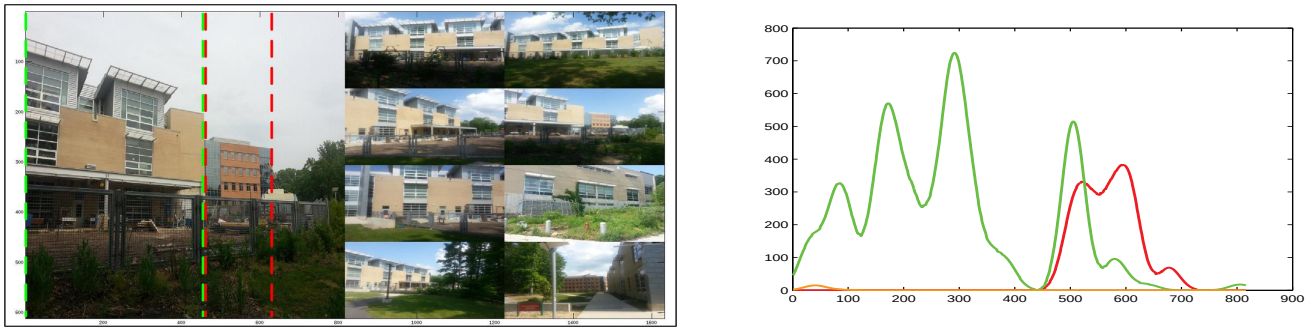
C. Building categorization and detection

The approach described in the previous section enables us to retrieve relevant images from the reference dataset. The retrieval results could be further refined using spatial verification using RANSAC with motion model of the choice (homography, essential matrix, trifocal tensor etc). This would yield possible refinement of the results. In the case the reference images have GPS coordinates one could then proceed with geometric verification and pose estimation between the query view and retrieved images provided that at least two geometrically consistent images were available. This strategy was used in the past by [24], [25] as means of computing geographic location of the query view. In many practical settings, it is often difficult to find the two views with sufficient overlap and even if we have enough overlap between images, there is no guarantee that the resulting pose estimates would be correct due to the repetitive

(a) Baseline (b) Our method

(c) Baseline (d) Our method

Fig. 2. Qualitative comparison between baseline TF-IDF BoW and our method. At each cell, the left image is the query image and the right images are the top 8 retrieved matches. The matches are shown from left to right in row order. Correct retrieved images are indicated by red rectangle. (a) and (b) illustrate the effect of vegetation on the performance of image retrieval. (c) and (d) illustrate the effect of season change on the retrieval. Note that our method not only retrieves more relevant images but also the retrieved set contains all the building which are present in the query image and it has more visual overlap with the query image.



(a) (b)

Fig. 3. Illustration of horizontal extent finding procedure. (a) contains all the retrieved images and the horizontal extents of the buildings are shown with the dashed vertical lines. The color of the lines corresponds to the color of diagrams in the right column. (b) contains the overall unnormalized probability for the top 8 matched images.

structures on different sides of buildings exhibit the same visual appearance. Fig. 4 illustrates such a situation where one of the retrieved images have the same visual features but it is taken from the other side of the building.

In our setting we seek a representation which could be deployed in the semantic localization setting given a map represented by landmarks [2] and does does not rely on the constraints such as similarity of robot's trajectory in the training and test phase or retrieving two images with sufficient overlap and accurate GPS coordinates. Semantic localization can be formulated as other traditional localization problems in the particle filtering framework. Without GPS or trajectory constraints, each particle can represent feasible locations and is characterized by the observation likelihood of the image given a 2D map of the presence of landmarks and buildings. The observation likelihood of the image at particular location requires a method for identifying and localizing buildings in the image. Similarly to the localization methods using range sensors, we want to compute the relative bearing of each of the buildings in an image. We assume that that tilt and roll of camera is small and the problem of bearing estimation is equal to finding *horizontal extent* of the buildings in the image field of view. Horizontal extent of a building is a bounding box that has a height of image and it contains a specific building. Fig 4 visualizes the horizontal extent.

We address the problem of computing horizontal extent of a building in three parts: 1) semantic segmentation 2) retrieving similar images to the query image using the semantic information 3) computing the horizontal extent of the building in the scene from the retrieved images.

Finding Horizontal extent: In this section, we want to transfer the knowledge from the training images to the test images. Let $B = \{b_1, b_2, ..., b_{n_b}\}$ be the set of buildings in the training dataset. Let P^i be the set of features in the ith retrieved image and Q be the set of features in the query image. p^i_j represents the jth feature in the ith retrieved image

Fig. 4. Illustration of horizontal extent: The query image is the one which has circle around its balloon and the other two images are retrieved images. The yellow and red dashed lines represent the horizontal bearing of the building in the image and its meaning in the real word is shown with the same color.

and q_k denotes the kth feature in query image. If p_j^i and q_k correspond to each other and p_j^i is on building b, then it is likely that feature q_k will be on building b too. In order to find feature correspondence we use the standard method by [12] followed the symmetric matching to keep only the mutual matches. For the geometric verification stage we use 5-point algorithm [16] (Algorithm 3, line 1) to find the inliers.

We next proceed with estimation of horizontal extents in images. We discretize the horizontal field of view into columns $c \in C$ and for each column in the query image, we compute the likelihood of column c containing building $b \in B$ using the top N retrieved images. The idea is to transfer the labeled information from the inlier features in the retrieved set to the query image and then accumulate all the likelihoods to get the overall probability of each building b given each column c. Therefore, it is important to make sure that the ith retrieved image is actually correct. That is it contains at least one common building with query image I^q. $p(I^q|I_i^t)$ is the likelihood of ith retrieved image having common building with query image I^q and it is equal to

$$p(I^q|I_i^t) = \alpha \times |M_i| \times S_i \tag{1}$$

where $|M_i|$ is the number of inlier matches between query image I^q and the retrieved image I_i^t. S_i is the cosine distance between the feature vector of query image and the retrieved image and α is the normalization factor. The more inlier features there are, the more likely that we retrieved correct image. In addition, having higher cosine distance means that the feature vectors are more similar and again the likelihood increases (Algorithm 3 line 2). We also define $p(q_k|p_j^i)$ which is equal to likelihood of feature p_j^i be the correct match for q_k and is equal to $\mathcal{N}(||p_j^i - q_k||_2, \sigma_s^2)$ (Algorithm 3, line 3). The larger the euclidean distance between features, the less likely it would be that this feature is a correct match. We also need to determine which building each feature p_j^i belongs to. To do this, we use the building label information and building identity available in the dataset. $p(b|p_j^i)$ is the probability of feature p_j^i be on building b and it is equal to the frequency of pixels in the 7×7 pixels neighborhood of the feature point and being labeled as b over the number of pixels in the patch, which is 49 in our case (Algorithm 3, line 4). Last but not the least, the probability of identity of each column is not independent of its neighboring columns because buildings are continuous. We enforce this by propagating the probability of each column to its neighbors by a weight of $\mathcal{N}\left(x_{q_k} - c, \sigma_x^2\right)$ (Algorithm 3, lines 5 and 6). The following equation consolidates all of the probabilities we discuss in this section.

$$p(b|c) = \beta_c \times \sum_{f, I_i^t, q_k, p_j^i} \left[p(I^q|I_i^t) \times p(q_k|p_j^i) \times p(b|p_j^i) \times \mathcal{N}\left(x_{q_k} - c, \sigma_x^2\right)\right] \tag{2}$$

where β_c is the normalization factor for column c. Fig. 3 illustrates the procedure more clearly.

IV. Experiments

A. Dataset

Majority of the datasets are either images of commonly photographed landmarks, with single landmark being the dominant central part of the image or structured datasets like StreetView dataset, where images of urban streets are taken at regularly sampled intervals from canonical viewpoints. In our case the goal was to collect a dataset which would have good coverage in terms of visibility of landmarks on our campus, but the landmarks themselves would not be the central part of each image. We selected a subset of buildings on the campus and collected images of the buildings. Some of the images contain only some part of the buildings. There are also images which have the close-up of the buildings. The dataset consist of 849 images which we partitioned it into almost equally sized sets for test and train sets. The images are taken by cell-phone camera. For all the images we labeled the identity of the buildings which are present in addition to the horizontal extent of each of the buildings in each image. For the images in the training set, we applied semantic segmentation and label each of the segments as background if they do not contain any part with label building, or with the identity of the building if they lie on any of the buildings.

B. Semantic Based Retrieval

In order to build the dictionary of visual words, we gathered all the sift features of the images with peak threshold for key point detection equal to 0. We used vlfeat [23] implementation of approximate k-means algorithm to build our vocabulary. We increased the number of clusters until the idf histogram of the dictionary has larger entropy indicating discriminative properties of certain visual words. We use the vocabulary of 7000 visual words. For the semantic segmentation, we used the implementation of our previous work [21]. Since we are interested in retrieving images which has the overlapping set of the buildings with the query image, we define the precision and recall as follows. Recall is the number of correctly retrieved buildings in the top N matches over the number of buildings which are present

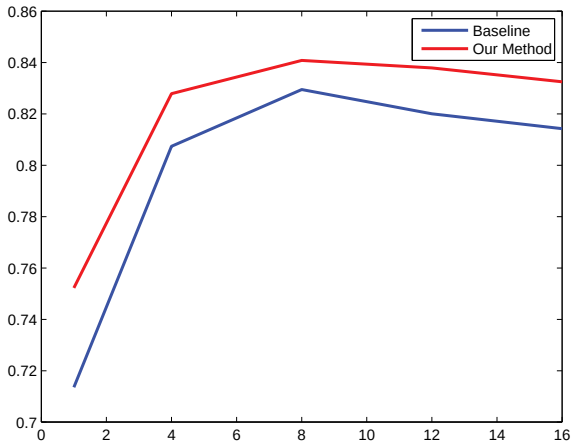


Fig. 5. The F-1 score with various number of retrieved images. F-1 scores are computed using top 1, 4, 8, 12, and 16 matches.

TABLE I
COMPARISON OF IMAGE RETRIEVAL PRECISION USING TOP N MATCHES

	N=1	N=4	N=8	N=12	N=16
Baseline	0.7596	0.7579	0.7406	0.7164	0.6990
Our Method	**0.8050**	**0.7857**	**0.7664**	**0.7466**	**0.7326**

in the query image. Correctly retrieved buildings are those being present in the query image. Precision is the number of relevant images over the number of retrieved images. An image is considered to be relevant, if the set of the buildings in that image has overlap with the set of the buildings present in the query image. Precision and recall for our method and baseline is included in table I and II. Our method improves the overall precision. However, when the number of retrieved images grow, the recall for our method become slightly less than the baseline. The reason for that is in the scenes which buildings take small portion of the scene, the baseline method retrieves images with the similar structure to the query image which does not necessarily mean that they contain the correct building. When the number of images being retrieved increases, the probability of retrieving images that contain the actual building by chance will increase. Fig 2 is an example of such situation. This phenomenon also shows itself in the top 1 retrieval. To be more concrete, we have also computed the F-1 score with various number of retrieved images. Fig. 5 shows the diagram of F-1 score with/without using semantic information. As it is shown, semantic information improves the result without exerting extra computational cost in the test phase which is appealing in robotic applications. Note that no spatial verification applied on the retrieved images to refine the retrieved short list.

C. Estimating Horizontal Extent

In order to estimate the horizontal extent of the building, we used 5-points algorithm implementation of the [16] in RANSAC framework. Having found the inlier matches between the query image and semantically-pruned features

TABLE II
COMPARISON OF IMAGE RETRIEVAL RECALL USING TOP N MATCHES

	N=1	N=4	N=8	N=12	N=16
Baseline	0.6727	0.8638	**0.9426**	**0.9588**	**0.9751**
Our Method	**0.7060**	**0.8749**	0.9312	0.9546	0.9641

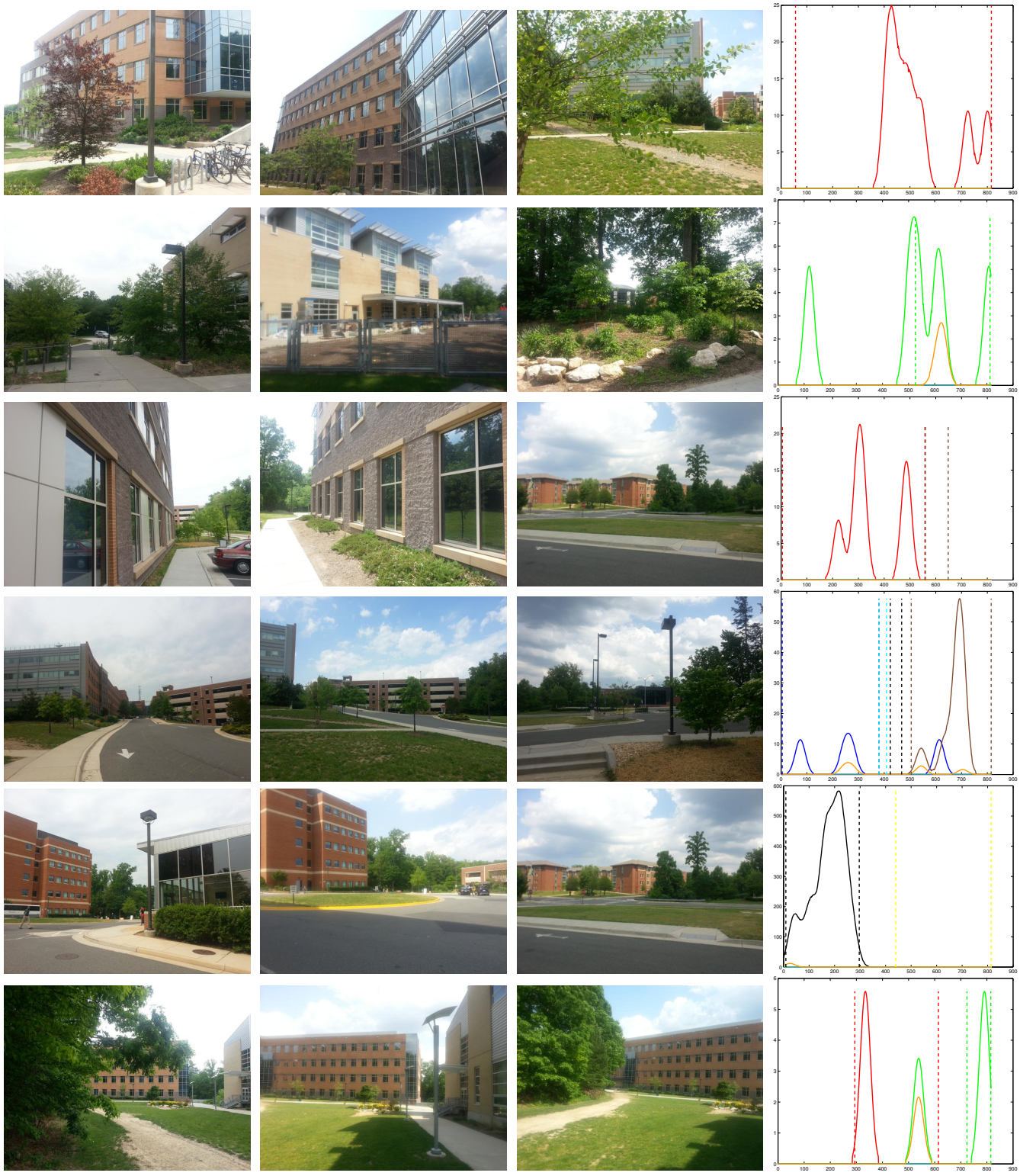
TABLE III
QUANTITATIVE EVALUATION OF THE HORIZONTAL EXTENT ESTIMATION ACCURACY

	N=1	N=4	N=8	N=12	N=16
Baseline	0.4458	0.6046	0.6384	0.6370	0.6308
Our Method	**0.5331**	**0.7514**	**0.8025**	**0.8145**	**0.8218**

in the training images, we calculate the weight of the features using (2) for all the buildings and one extra category which represent background. we used $\sigma_x = 16$. We classify each column c to b_c^* such that $b_c^* = argmax_b p(b|c)$. We compared our performance with the baseline method for estimating horizontal extent. In the baseline algorithm, matching process is done between all the features in the query image and all the features in the training images. For evaluation, we labeled the horizontal extent of each image manually. We defined the accuracy of horizontal extent estimation as ratio of the number of columns which are correctly classified as one of the buildings over the total number of columns having building. Table III shows the quantitative comparison of using semantic or not in estimation of the horizontal extent. As it is shown, our method improves by retrieving more and more images. On the other hand, the performance plateaus with the increase in the number of retrieved images. The main reason is the accuracy advantage of our method over the baseline method in retrieving correct images. Another phenomenon which occurs by the increase the number of retrieved images is that the accuracy of the baseline starts decreasing with the grow in the size of the retrieval set. This is due to the fact that that the images which are being retrieved is not accurate and they cause confusion and as a result it leads to confusion of the method in estimating horizontal extent of the building.

V. CONCLUSIONS

We have demonstrated that the capability of detecting man-made structures can enhance the traditional local feature based methods when the viewpoint change and appearance changes are more significant. Focusing the matching on man-made structures, helps to discard irrelevant features from the scene which often act as confusers in various voting strategies, such as trees, vegetation and road features are often not discriminative of location. Towards this end we have introduced a new dataset for building localization and recognition. In addition to the building detection problem, we have shown that the local features can be used to categorize the individual building instances and their extent in images. At the moment we evaluate the building localization approach using commonly used measure for measure performance of object detectors, but in the future work we plan to tie it together with the semantic geo-localization and



. (a) Query Image (b) Semantic Aware Retrieval (c) Baseline (d) Horizontal Extents

Fig. 6. Qualitative evaluation of our method with and without semantic information. The first column is the query image and column (b) and (c) show the top retrieved image. Column (d) illustrate the estimated horizontal extent using only the top retrieved images in column (b). The vertical dash line represent the ground truth of the horizontal extents of buildings in the query image and the solid lines represent unnormalized probabilities for each column.

evaluate the accuracy of determining the geographic location of the agent.

VI. Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, con-

tract FA8650-12-C-7212. 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] L. Torresani A. Bergamo, S. N. Sinha. Leveraging structure from motion to learn discriminative codebooks for scalable landmark classification. In *CVPR*, 2013.

[2] Nikolay Atanasov, Menglong Zhu, Kostas Daniilidis, and George Pappas. Semantic Localization Via the Matrix Permanent. In *RSS*, 2014.

[3] Mark Joseph Cummins and Paul M. Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *International Journal of Robotics Research*, 27(6):647–665, 2008.

[4] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. What makes Paris look like Paris? *ACM Transactions on Graphics*, 31(4):101, 2012.

[6] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[7] C. Buckley G. Salton. Term-weighting approaches in automatic text retrieval. In *In Information Processing and Management*, 1988.

[8] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla. Learning per-location classifiers for visual place recognition. In *CVPR*, 2013.

[9] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007.

[10] Jan Knopp, Josef Sivic, and Tomás Pajdla. Avoiding confusing features in place recognition. In *ECCV*, pages 748–761, 2010.

[11] K. Konolige, J. Bowman, J. Chen, P. Michelich, M. Calonder, V.Lepetit, and P. Fua. View-based maps. In *RSS*, 2009.

[12] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[13] Colin McManus, Ben Upcroft, and Paul Newman. Scene signatures: Localised and point-less features for localisation. In *Proceedings of Robotics Science and Systems (RSS)*, Berkley, CA, USA, July 2014.

[14] A. Mikulik, M. Perdoch, O. Chum, and J. Matas. Learning a fine vocabulary. In *ECCV*, 2010.

[15] M. J. Milford and G. F. Wyeth. Mapping a suburb with a single camera using a biologically inspired slam system. *IEEE Transactions on Robotics*, 29(1), 2008.

[16] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, 2006.

[17] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[18] N. Snavely S. Cao. raph-based discriminative learning for location recognition. In *CVPR*, 2013.

[19] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. In *CVPR*, 2007.

[20] G. Singh and J. Kosecka. Nonparametric scene parsing with adaptive feature relevance and semantic context. In *CVPR*, 2013.

[21] Gautam Singh and Jana Košecká. Acquiring semantics induced topology in urban environments. In *ICRA*, pages 3509–3514, 2012.

[22] Joseph Tighe and Svetlana Lazebnik. SuperParsing: Scalable Nonparametric Image Parsing with Superpixels. In *ECCV (5)*, pages 352–365, 2010.

[23] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

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

[25] Wei Zhang and Jana Košecká. Image based localization in urban environments. In *3DPVT06*, pages 33–40, 2006.