Semantically Aware Bag-of-Words for Localization

Arsalan Mousavian
George Mason University
4400 University Dr, Fairfax, VA, USA
amousavi@gmu.edu

Jana Košecka
George Mason University
4400 University Dr, Fairfax, VA, USA
kosecka@gmu.edu

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 in the context of topological mapping or loop closure detection. Difficulties of the location recognition problem are often affected by large appearance and viewpoint variation between the query view and reference dataset and presence of non-discriminative features due to vegetation, sky and road. We demonstrate that the availability of semantic information about the presence of man-made landmark structures such as buildings, can enhance the traditional BoW local features methods. Focusing the matching on man-made structures, helps to discard irrelevant features from the scene that often act as confusers in various voting strategies (e.g. trees, vegetation and road features are often not discriminative of location).

1. Introduction

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 (BoW) representation, followed by spatial verification of the top k retrieved images using geometric constraints [13]. The spatial verification helps to refine the matched using RANSAC with an appropriate geometric model effectively eliminating outliers. Various improvements of BoW methods include learning better vocabularies, developing better quantization and spatial verification methods [10, 19], employing more sophisticated models of dependence of local features [3] or improving the scalability [15]. In the context of visual place recognition, in contrast to image based retrieval, there is often additional structural information available that can be exploited towards the task. In case the locations and/or landmarks are visible in many views 3D reconstruction techniques can help quatifty the 'matchability' of individual features. Authors in [1] chose the stable keypoints detectable across many views and trained a randomized decision tree classifier per keypoint to classify the new images. In [14] 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 image using TF-IDF weighting scheme. The confusion weight for each features is learned in [8] exploiting the geographic location of the unweighted retrieved views. In [6] they train exemplar SVM per image in the training set where the negative examples are the images that have similar cosine distance but they are far away. The retrieval approach of [8] is similar to our method in selecting the features that are informative for localization and suppressing the confusing ones. We instead of learning the weight for each feature using the GPS coordinates, propose to use general purpose semantic segmentation technique [16] to classify features to different semantic categories and filter those that do not belong to man-made structures. Arandjelovic and Zisserman [2] augmented the dictionary with semantic categories surrounding each SIFT feature and they retrieve using the augmented vocabulary. The difference of their method and the proposed method is that we use semantic information to refine the unrelated SIFT features, while they use semantic information to add extra information to the SIFT features in order to disambiguate the SIFT representation.

Figure 1. Overview of our semantically aware BoW: Left is query image. Middle is the semantic segmentation of the image. Right shows horizontal extent of the building with dashed lines.

We show (1) that semantic segmentation of man-made

structures provides an improvement in retrieval accuracy of traditional bag-of-words methods in the presence of large variation in imaging conditions 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. The obtained semantic representation can be exploited further for semantic localization and mapping as well as determining geographic location of the query views.

2. Proposed Method

Given the reference set of images, the standard BoW method clusters the SIFT features to build vocabulary of visual words, assigning each visual word with the same weight.TF-IDF weighting [5] tackles this problem by defining two terms; The term frequency $tf(w, I)$ which is equal to the number of times words w appears in image I and inverse document (image) frequency $idf(w)$ is equal to

$$idf(w) = \log \frac{N}{|\{I_i : w \in I_i\}|}$$

where N is the number of images in the dataset and $|\{I_i : w \in I_i\}|$ is the number of images I_i containing visual word w. The more frequent a visual word is, the smaller $idf(w)$ gets. TF-IDF representation of image I will be the $|W|$ dimensional vector where $|W|$ is the vocabulary size and each dimension is equal to the product of $tf(w, I)$ and $idf(w)$. Even though TF-IDF improves the performance considerably, it still considers all the visual words of 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 will be large. We propose to rectify this problem by exploiting the semantic information available in the image, which will be used to change the weighting in more informative way.

2.1. Semantic Segmenatation

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 proposed semantic segmentation approach is closely related to [7], [18], and [17]. An example of semantic segmentation is shown in Figure 1.

2.2. Semantically Aware Image Retrieval

We first build vocabulary of all of the SIFT features using k-means algorithm and then compute the IDF term for each of the visual words in the vocabulary (Algorithm 1, lines 1 and 2). Semantic segmentation [17] for all images in the reference set labels 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 that do not belong to man-made structure regions (Algorithm 1, line 3). After pruning the local features that do not belong to buildings, we recompute the TF component for each of the training images and compute a new TF-IDF BoW representation (Algorithm 1, line 4). This is different from [8] where the weight is computed for each visual word for all of its occurrences regardless of the semantic category by using GPS coordinates of the images.

Algorithm 1 Training Procedure

1: Build vocabulary using approximate k-means.
2: Compute IDF for each visual word.
3: Prune SIFT features that are not in the man-made structure regions.
4: Compute TF-IDF BoW representation for each training image.

Algorithm 2 Retrieval Procedure

1: Compute the TF-IDF vector using all the extracted SIFT features in the query image.
2: Retrieve top K images based on the TF-IDF representation using cosine distance.

In the test phase, we extract the SIFT features from the query image and assign each feature 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 K matches by finding K-nearest neighbors using the cosine distance between the query BoW and semantically pruned feature vectors of training images (Algorithm 2, line 2). More details can be found in [11]. Note that we compute TF-IDF representation of query image without any semantic pruning. In order to build the vocabulary of the visual words, we use all features regardless of the semantic categories. If we only use features belonging to man-made structures, our vocabulary overfits to the features on man-made structures and it does not generalize well to the features present in the query image.

2.2.1 Finding Horizontal extent

In this section, we describe a method for localization and identification of buildings in query views, assuming that our reference training set has $B = \{b_1, b_2, ..., b_{n_b}\}$ set of building identities. We have in the training stage labelled each image in the reference set by the building identity present

Algorithm 3 Horizontal Extent Estimation

1: Find corresponding feature pairs between query image and each of the retrieved images.
2: Compute retrieval score for each of the retrieved images using Eq 2.
3: Compute feature match probability $p(q_i|p_j^k)$.
4: Compute feature identity probability $p(b|p_j^k)$.
5: Accumulate $p(q_i|p_j^k) \times p(b|p_j^k, i) \times p(I^q|I_k^t)$ for all the K retrieved images in the corresponding column x_{q_k}.
6: Convolve it with Gaussian filter with $\sigma = \sigma_x$.

in the image and their horizontal extent. Given the matches between query view and reference view, we will show how the matched features provide evidence about building identity and its horizontal extent. Let P^k be the set of features in the kth retrieved image and Q be the set of features in the query image. p_j^k represents the jth feature in the kth retrieved image and q_i denotes the ith feature in query image. If p_j^k and q_i correspond to each other and p_j^k is on building b, then it is likely that feature q_i will be on building b too. In order to find feature correspondence we use the standard method by [9] followed the symmetric matching to keep only the mutual matches. For the geometric verification stage we use 5-point algorithm [12] (Algorithm 3, line 1) to find the inliers.

We next proceed with the estimation of horizontal extents of buildings in images. Horizontal extent of a building is the horizontal interval that contains a building. An example of horizontal extent is shown in Figure 1. 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 K retrieved images. For each image in the reference set, we label each building pixel with the identity of the matched building. The idea is to transfer the building identity from the inlier features in the retrieved set to the query image and then accumulate all the likelihoods to get the overall probability building b given each column c. The probability of column c having identity b is computed using the following equation:

$$p(b|c) = \beta_c \times \sum_{f, I_i^t, q_i, p_j^k} \big[p(I^q|I_k^t) \times p(q_i|p_j^k) \times p(b|p_j^k) \times \mathcal{N}\left(x_{q_i} - c, \sigma_x^2\right)\big] \quad (1)$$

$p(I^q|I_k^t)$ is the likelihood of kth retrieved image having at least one common building with query image I^q (Algorithm 3, line 2). Let $p(q_i|p_j^k)$ be the likelihood that feature q_i is indeed a correct match for p_j^k and is equal to $\mathcal{N}(||p_j^k - q_i||_2, \sigma_s^2)$ (Algorithm 3, line 3). $p(b|p_j^k)$ is the probability of feature p_j^k belongs to 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 (Algorithm 3, line 4). 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 $\mathcal{N}\left(x_{q_i} - c, \sigma_x^2\right)$ (Algorithm 3, lines 5 and 6). β_c is the normalization parameter for column c.

Table 1. Comparison of Image Retrieval F1-score using Top K Matches

	K=1	K=4	K=8	K=12
Baseline	**0.7766**	0.8475	0.8511	0.8460
Method of [2]	0.6322	0.7001	0.7289	0.7414
Our Method	0.7621	**0.8647**	**0.8735**	**0.8672**

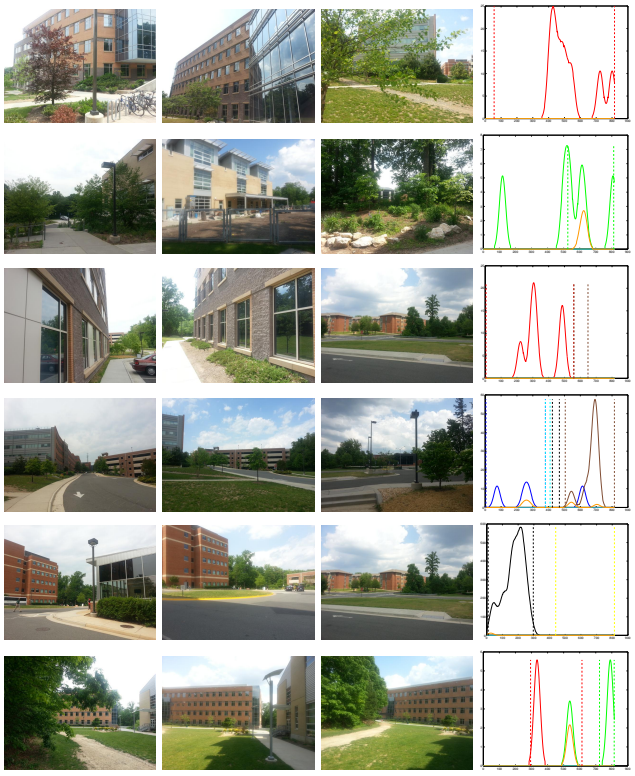
Two images that contain at least one common building, should have 1) similar TF-IDF representation and 2) large number of inlier corresponding features. Therefore, we define $p(I^q|I_k^t)$ as following:

$$p(I^q|I_k^t) = \alpha \times |M_k| \times S_k \quad (2)$$

where $|M_k|$ is the number of inlier matches between query image I^q and the retrieved image I_k^t. S_k is the cosine distance between the feature vector of query image and the retrieved image. If the TF-IDF representation of the images I^q and I_k^t are similar to each other, their dot product becomes larger. In addition, the more inlier features there are, the more likely that we retrieved correct image (Algorithm 3 line 2). Algorithm 3 illustrates the procedure more clearly. Figure 2 shows the qualitative comparison between BoW and semantically aware BoW (SBoW). As it is shown the top image using the proposed method is more suitable for retrieving images containing the same buildings.

3. Experiments

The dataset consists of 816 sparsely sampled images which are taken from 11 buildings in a university campus. The images are labeled with the pixel level identity of the buildings. Table1 shows the F-1 measure for the image retrieval. The precision is defined as the number of images containing at least a building in common with the query image and recall is total number of buildings in the query image for which at least one image is retrieved. We compared our approach with the method [2] where they augmented visual words with semantic categories. As Table2 shows the accuracy of horizontal extent identification, SBoW can retrieve images with more overlap comparing to BoW and [2]. Therefore, our method is more suitable for identifying the landmarks, buildings in our case, for localization.



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

Figure 2. 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.

Table 2. Quantitative Evaluation of the Horizontal Extent Estimation Accuracy

	K=1	K=4	K=8	K=12
Baseline	0.4458	0.6046	0.6384	0.6370
Method of [2]	**0.5433**	0.6586	0.6947	0.7451
Our Method	0.5331	**0.7514**	**0.8025**	**0.8145**

4. Discussion

We used semantic segmentation to eliminate features which do not belong to man-made structures. Our results show that semantically aware bag-of-words not only improves the retrieval, but also retrieves more images with a sufficient overlap, which is desirable for image-based geolocation and increases the matchability of the related instances. In the future work we plan to use the semantically aware bag-of-words and building identification in a localization setting.

References

[1] L. T. A. Bergamo, S. N. Sinha. Leveraging structure from motion to learn discriminative codebooks for scalable landmark classification. In *CVPR*, 2013.

[2] R. Arandjelovic and A. Zisserman. Visual vocabulary with a semantic twist. *Asian Conference on Computer Vision*, 2014.

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

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

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

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

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

[8] J. Knopp, J. Sivic, and T. Pajdla. Avoiding confusing features in place recognition. In *ECCV*, pages 748–761, 2010.

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

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

[11] A. Mousavian, J. Košecká, and J. M. Lien. Semantically guided location recognition for outdoors scenes. In *ICRA*, 2015.

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

[13] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

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

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

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

[17] G. Singh and J. Košecká. Acquiring semantics induced topology in urban environments. In *ICRA*, pages 3509–3514, 2012.

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

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