Image Retrieval for Image-Based Localization Revisited

Torsten Sattler[1]
tsattler@cs.rwth-aachen.de
Tobias Weyand[2]
weyand@vision.rwth-aachen.de
Bastian Leibe[2]
leibe@vision.rwth-aachen.de
Leif Kobbelt[1]
kobbelt@cs.rwth-aachen.de

[1] Computer Graphics Group
RWTH Aachen University, Germany

[2] Computer Vision Group
RWTH Aachen University, Germany

Image-based localization is the task of determining the exact location from which a query photo was taken. In this paper, we consider image-based localization relative to a 3D point cloud of a scene, obtained from Structure-from-Motion, which allows an accurate estimate of the full camera pose from correspondences between 2D features and 3D points. To quickly establish the required 2D-to-3D matches, Irschara *et al*. use image retrieval methods [5] to find database images (used for the reconstruction) similar to the query image [1]. Since the relation between 2D features and 3D points is known for the database images, the correspondences for the query image can be computed by feature matching between images. Recent work has demonstrated that directly matching the features against the points outperforms retrieval-based methods in terms of the number of images that can be localized successfully [4]. Yet, direct matching is inherently less scalable than retrieval-based approaches since it needs to keep SIFT descriptors [3] in memory at all times.

In this paper, we therefore analyze the algorithmic factors that cause the gap in registration performance. We show that using *selective voting* schemes enable retrieval methods to outperform state-of-the-art direct matching methods and explore how both selective voting and correspondence search can be accelerated by using a Hamming embedding [2].

Selective Voting

The main cause for the performance gap are the *incorrect votes* that are cast by image retrieval-based approaches such as [1]. Fig. 1 illustrates this problem. Although the query feature (pink) corresponds to only a single 3D point (red), inverted file scoring also casts a vote for every image that has a feature (black) matched to the same visual word. Dealing with these incorrect votes is challenging even for advanced re-ranking schemes such as *tf∗idf* weighting [5] or *probabilistic ranking* [1]. Since pose estimation is only attempted for the top-*k* images, failure to rank any of the relevant images among the top-*k* negatively impacts localization performance.

Two *selective voting* schemes can be used to avoid incorrect votes. *Correspondence voting* finds the two nearest neighbors among all descriptors of 3D points having the same visual word and votes for the image that contains the nearest neighbor if the SIFT ratio test [3] is passed. This scheme essentially uses the correspondences found by the direct matching approach from Sattler *et al*. [4] to vote for database images. The camera pose is then estimated from correspondences found with pairwise image matching. Since *correspondence voting* requires that SIFT descriptors are kept in memory at all times, a *selective voting* scheme using Hamming embedding [2] can be used to the reduce memory requirements. The resulting *Hamming voting* only casts a vote for an image containing a point if the Hamming distance between the binary embeddings of the query feature and the point is below a certain threshold (*cf*. Fig. 1(right)). Using 64-bit for the embedding requires only little memory overhead to store the embeddings in the words, while 10^6 Hamming distance computations can be done in about 2ms on a modern CPU. Thus, *Hamming voting* preserves the scalability of retrieval-based methods.

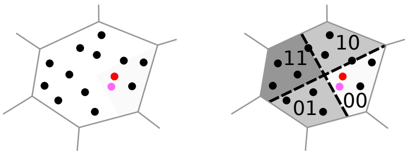

Figure 1: (Left) The query feature (pink) corresponds to a single 3D point (red), yet unrelated inverted file entries (black) cause *false positive votes*. (Right) By thresholding Hamming distances of a Hamming embedding, *Hamming voting* can avoid casting many of the incorrect votes.

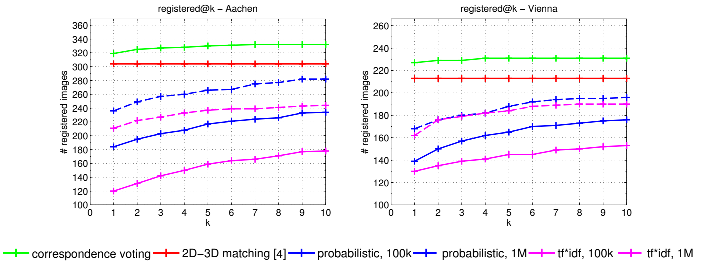

Figure 2: The *correspondence voting* scheme is able to achieve significantly better results than standard ranking schemes due to its ability to discard incorrect votes. It also outperforms the direct matching approach.

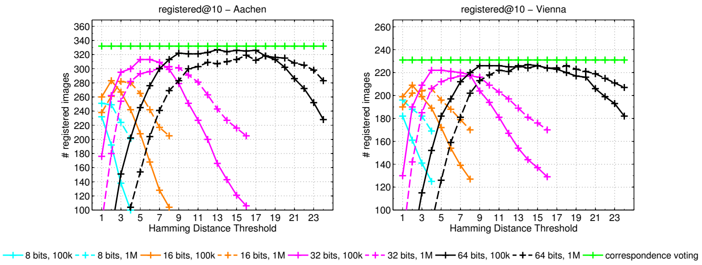

Figure 3: *Hamming voting* using 32- and 64-bit vectors achieves nearly the same performance as *correspondence voting* with SIFT descriptors.

Results

We compare *selective voting*-based localization to classical image retrieval-based methods and the state-of-the-art direct matching approach from [4]. We measure the performance of the methods in the number of images for which a pose can be estimated successfully. Two large-scale datasets are used for the evaluation, including our novel Aachen dataset consisting of 1.5M 3D points and 369 query images [1].

Fig. 2 compares *correspondence voting* to retrieval methods using different ranking schemes with different visual vocabulary sizes and the method from [4]. Using this *selective voting* scheme significantly improves image retrieval-based localization and enables us to outperform the state-of-the-art direct matching method [4].

The evaluation of *Hamming voting* with different sizes for the resulting binary descriptors and different vocabulary sizes in Fig. 3 shows that a performance similar to *correspondence voting* can be achieved using nearly one order of magnitude less memory. Thereby, using a coarser vocabulary yields better results due to less quantization errors.

Further results on accelerating the matching between images required for correspondence search can be found in the paper.

[1] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From Structure-from-Motion Point Clouds to Fast Location Recognition. In *CVPR*, 2009.

[2] H. Jégou, M. Douze, and C. Schmid. Packing bag-of-features. In *ICCV*, 2009.

[3] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2), 2004.

[4] T. Sattler, B. Leibe, and L. Kobbelt. Fast Image-Based Localization using Direct 2D-to-3D Matching. In *ICCV*, 2011.

[5] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, 2003.

[1] Available at `http://www.graphics.rwth-aachen.de/localization`.