Combining monocular and stereo cues for mobile robot localization using visual words

Friedrich Fraundorfer*, Changchang Wu† and Marc Pollefeys*
*Department of Computer Science
ETH Zürich
{fraundorfer,marc.pollefeys}@inf.ethz.ch
†Department of Computer Science
University of North Carolina at Chapel Hill
ccwu@cs.unc.edu

***Abstract*—This paper describes an approach for mobile robot localization using a visual word based place recognition approach. In our approach we exploit the benefits of a stereo camera system for place recognition. Visual words computed from SIFT features are combined with VIP (viewpoint invariant patches) features that use depth information from the stereo setup. The approach was evaluated under the ImageCLEF@ICPR 2010 competition[1]. The results achieved on the competition datasets are published in this paper.**

I. Introduction

The ImageCLEF@ICPR 2010 competition was established to provided a common testbed for vision based mobile robot localization, to be able to evaluate different approaches against each other. For the competition image datasets of a realistic indoor scenario were created and manually labeled to get ground truth data. The mobile robot was equipped with a stereo vision system, that generates an image pair for each location instead of a single image only. This availability of image pairs from a stereo vision system allowed us to design an approach that combines monocular and stereo vision cues. The approach we designed is based on a place recognition system using visual words [1], [2]. For one part visual words are computed from SIFT features [3] as the monocular cue. As the other cue we use visual words computed from viewpoint invariant patches (VIP) [4]. For the extraction of VIP features the local geometry of the scene needs to be known. In our case we compute dense stereo from the stereo system and use the depthmap for VIP feature extraction. Both sets of features are combined and used as visual description of the location. The approach has already been evaluated and the scores achieved in the competition are given in this paper. Additionally, recognition rates on the *validation* competition datasets (which was used to prepare for the competition) are given, which demonstrate the excellent performance of our method by achieving 98% correct localization.

II. Related work

Our visual word based place recognition system is related to [1], [2] where a similar technique was used for image retrieval. It is also related to FABMAP [5], a visual word based approach to robot localization. However, FABMAP focuses on a probabilistic framework to identify matching locations, whereas we do a two-stage approach of visual ranking and geometric verification. The proposed geometric verification takes the planar motion constraints of a mobile robot into account. It is also related to [6], however they use a different technique of quantizing local features into visual words.

VIP features were firstly described in [4] and used for registration of 3D models. The local geometry was computed using a monocular structure-from-motion algorithm. In our approach we compute VIP features from a stereo system and use them for mobile robot localization the first time.

III. SIFT and VIP features for place recognition

The availability of depthmaps for every image pair of a stereo video sequence makes it possible to use viewpoint invariant patches (VIP). VIP's are image features extracted from images that are rectified with respect to the local geometry of the scene. The rectified texture can be seen as an ortho-texture of the 3D model which is viewpoint independent. This ortho-texture is computed using the depthmap from the stereo system. This rectification step, which is the essential part of this concept delivers robustness to changes of viewpoint. We then determine the salient feature points of the ortho-textures and extract the feature description. For this the well known SIFT-features and their associated descriptor [3] is used. The SIFT-features are then transformed to a set of VIPs, made up of the features 3D position, patch scale, surface normal, local gradient orientation in the patch plane, in addition to the SIFT descriptor. Fig. 1 illustrates this concept. The original feature patches are the lower left and right patches, as seen from the gray and green camera. Rectification is performed by changing the viewpoints to the red cameras, so that the camera's image plane is parallel to the features scene plane. This results in the rectified VIP patches which are the center patches. It can be seen, that because of the rectification the VIP patches overlap perfectly.

Because of their viewpoint invariance, VIP features are a perfect choice for place recognition. For place recognition VIP features can be used instead of SIFT features from the original

[1]This approach was ranked 1st in the ImageCLEF@ICPR 2010 RobotVision competition.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.955

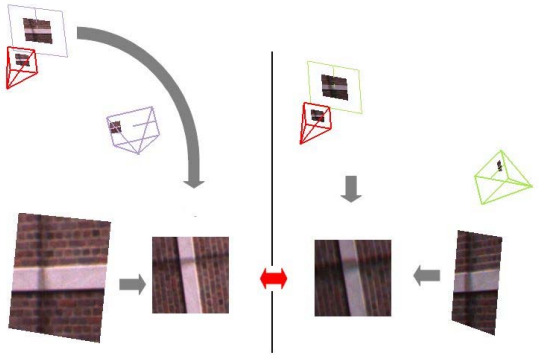

Fig. 1. Two corresponding viewpoint invariant patches (VIPs). The lower left and right patches are the original feature patches, while the center patches are the rectified VIP patches (see text for details).

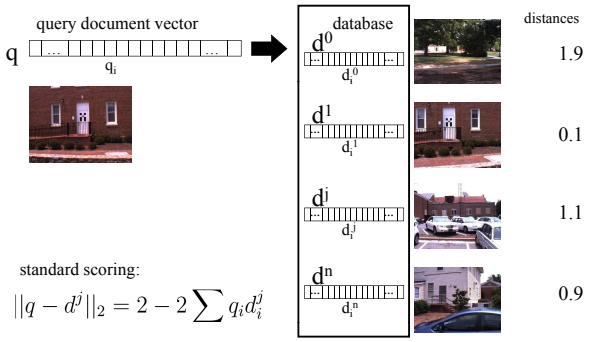

Fig. 2. Illustration of the visual place recognition system. A query image is represented by a query document vector (a set of visual words). A query works by computing the L_2-distance between the query document vector and all the document vectors in a database which represent the stored images, i.e. places. The L_2-distance is used as similarity score and is in the range of $[0, 2]$. A similarity score close to zero stands for a very similar image, while a similarity score close to 2 stands for a different image. The computation of the similarity score is using an inverted file structure for efficiency.

images or in addition to SIFT features (this is beneficial if the local geometry cannot be computed for the whole sequence). With view point invariant features place recognition will be possible with even large view point changes.

A. *Extraction of VIP features*

The first step of VIP feature extraction is the computation of a dense depth map from an image pair. This is done by scanline based stereo using dynamic programming [7]. As similarity measure the sum-of-absolute-differences (SAD) within a 9×9 pixel window is used. The next step is the detection of scene planes in the 3D point data from the depthmap. The final step is to transform the detected scene planes into an orthographic view, extract SIFT features from the rectified planes and transform them to VIP's by adding additional information about the 3D scene.

IV. Place recognition and verification

Robot localization can be phrased as a place recognition problem as described in [8]. The camera path is split up into distinct locations and the visual appearance of each location is described by visual features. A database of the environment is created holding the visual appearance of each location together with the actual coordinates of the location, and a label is assigned to each location. On performing global localization the current view of the robot is compared to all views in the database. The location with the most similar appearance is returned and the robot now knows its location up to the accuracy of the stored locations. For an efficient database search a visual word based approach is used. The approach quantizes a high-dimensional feature vector (in our case SIFT and VIP) by means of hierarchical k-means clustering, resulting in a so called hierarchical vocabulary tree. The quantization assigns a single integer value, called a visual word, to the originally high-dimensional feature vector. This results in a very compact image representation, where each location is represented by a list of visual words, each only of integer size. The list of visual words from one location forms a document vector which is a v-dimensional vector where v is the number of possible visual words (a typical choice would be $v = 10^6$). The document vector is a histogram of visual words normalized to 1. To compute the similarity matrix the L_2 distance between all document vectors is calculated. The document vectors are naturally very sparse and the organization of the database as an inverted file structure makes this very efficient. This scheme is illustrated in Fig. 2.

In our case robot localization is a 2-stage approach. First a similarity ranking is performed using the visual words, afterwards a geometric verification step tests the top-n results. Geometric verification is a very powerful cue. For each visual word the 2D image coordinates are stored, too. This makes it possible to compute the epipolar geometry between each database image and the query image. Only results that fulfill this epipolar constraint are considered. The geometric verification is passed if the number of inliers that fulfill the epipolar geometry is higher than a threshold t. To compute the epipolar geometry we use the planar 3-pt algorithm [9]. This algorithm assumes that the robot is moving in a plane and is therefore more efficient than an unconstrained motion estimation algorithm.

V. Evaluation

The evaluation of the algorithm was done within the framework of the ImageCLEF@ICPR 2010 Robotvision competition. The Robotvision competition consisted of two independent tasks for place recognition, *task1* and *task2*. *Task1* was conducted in a large office environment, which was divided into 13 distinct areas, which got different labels assigned. The goal for the robot was, to answer the question in which area it is given an input image, by assigning the correct label to

the input image. For the competition, labeled image data was provided for training and validation and an unlabeled data set (*testing*) on which the competition was carried out. After the competition the ground truth labels for the *testing* set were released to the participants. The data sets were acquired with a stereo set consisting of two Prosilica GC1380C cameras mounted on a MobileRobots PowerBot robot platform. For the two training sets *training_easy* and *training_hard* the robot was driven through the environment and the captured images were labeled with the area code. The *training_easy* set consists of more images and more viewpoints than the *training_hard* set. The *validation* set comes from a run through the environment at a different time and was also labeled. Both training sets and also the validation set include only 9 of the 13 areas. With the use of a stereo system an image pair is available for each location which allows the use of depth information. However this was not required in the competition. The goal of *task1* was to label each image pair of the *testing* set with the correct area code. The *testing* set consists of 2551 image pairs taken at a different time and includes all 13 areas. This means that the 4 areas not included in the training sets needed to be labeled as "Unknown area". For *task1* each image pair had to be labeled independently without using knowledge from the labeling of the previous image pair. *Task2*, an optional task, was very similar to *task1*, however here it was possible to include sequential information into the labeling. Thus this *task2* is considerably easier than *task1*.

In the following we will present and discuss the recognition rates on the *validation* set and compare it to the ground truth and give the score achieved on the *testing* set in the competition. Table I shows the results on the *validation* set (2392 images) where ground truth data as labels is available. We measured recognition scores using standard scoring and standard scoring with geometric verification on both training sets. For the recognition score we compare the label of the top-ranked image (denoted as standard scoring) with the ground truth label and compute the number of correctly labeled images. For geometric verification the top-50 images from standard scoring get re-ranked according to the number of inliers to the epipolar constraint. Using the *training_easy* set the recognition rate with standard scoring was 96% and this number increased to 98% with geometric scoring. For the *training_hard* set the recognition rate with standard scoring was 87%. This number increased to 92% with subsequent geometric verification. Interestingly the recognition rates with standard scoring are already very high, geometric verification seems to give only a small improvement. However, only after geometric verification can one be sure that the database image really matches the query image. Standard scoring provides a ranking but the similarity measure does not guarantee that the top ranked image is really the matching one, e.g. if the query image is not in the database at all.

Table II shows the competition scores achieved on the *testing* set. The score is computed as follows:

- +1.0 points for a correctly classified image (includes the correct detection of an unknown location).

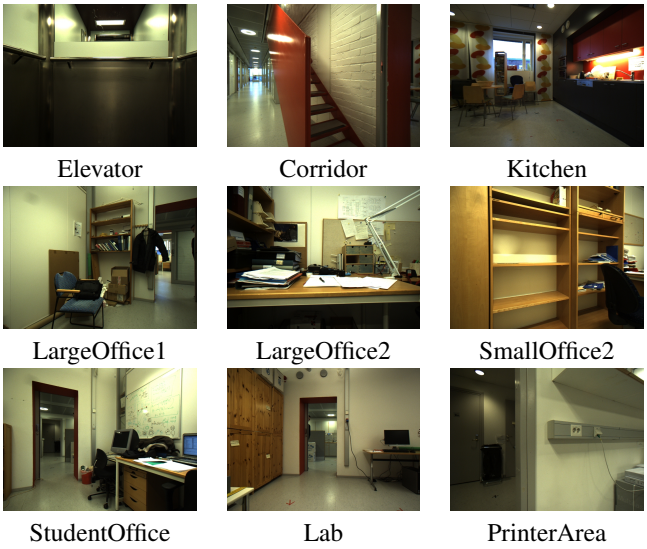

Fig. 3. Example images of the 9 classes in the training sets.

- -0.5 points for a misclassified image
- 0 points for an image that was not classified

The maximal achievable score would be 5102 points as the sum of the two 2551 points for each of the individual training sets. The top-10 ranked images are used for geometric verification. Image matches with less than 50 inlier matches were classified as "Unknown" location. We denote this parameterization as the "Competition method". Every image of the database was classified to either an area or the "Unknown" class, the option of refraining from a decision was not used. Table III shows the recognition rates for each individual class of the *testing* set. The table shows that some classes seem to be easier (100% recognition rate for the "Lab" class) while others seem to be harder. The table also confirms that the "Unknown" class is troublesome for our method, which has low recognition rates.

Finally we would like to give some runtime measurements from a 2.4GHz Intel Quadcore. An individual localization using standard scoring will take 26.4ms (including feature quantization into visual words). "Competition scoring" with geometric verification is currently taking 1.5s. Here the feature matching is not optimized and takes most of the time. Excluding the runtime for feature matching leaves 53.5ms for localization with epipolar geometry verification. Feature matching can easily be speeded up by using proper datastructures, e.g. a kd-tree [3], so that realtime speed can be achieved with this approach.

Method	*training_easy*	*training_hard*
standard scoring	0.96	0.87
geometric verification	0.98	0.92

TABLE I
RECOGNITION SCORES FOR THE VALIDATION SET WITH THE DIFFERENT TRAINING SETS AND METHODS.

	training_easy	*training_hard*	combined score
task1	2047	1777	3824

TABLE II
COMPETITION SCORES (WITH "COMPETITION METHOD") FOR TASK1.

Class	*training_easy*	*training_hard*
Full set	0.80	0.70
Elevator	0.97	0.98
Corridor	0.96	0.86
Kitchen	0.99	0.96
LargeOffice1	0.93	0.63
LargeOffice2	0.96	0.83
SmallOffice2	0.99	0.94
StudentOffice	0.73	0.62
Lab	1.00	1.00
PrinterArea	0.99	0.64
Unknown	0.56	0.65

TABLE III
RECOGNITION SCORES (WITH "COMPETITION METHOD") FOR THE INDIVIDUAL CLASSES AND THE FULL SET ON THE TESTING SET.

VI. CONCLUSION

The ImageCLEF@ICPR 2010 competition provides a challenging dataset to evaluate different methods for robot localization. The use of a stereo camera as imaging system for the robot allowed us to combine monocular and stereo cues for robot localization.

Our approach achieves high recognition rates (e.g. 98% on the *validation*), which signals that the proposed approach is reliable enough to be used in practice. To foster the use of this approach we made the code for the visual word based similarity ranking publicly available[2].

REFERENCES

[1] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition, New York City, New York*, 2006, pp. 2161–2168.

[2] F. Fraundorfer, C. Wu, J.-M. Frahm, and M. Pollefeys, "Visual word based location recognition in 3d models using distance augmented weighting," in *Fourth International Symposium on 3D Data Processing, Visualization and Transmission*, 2008.

[3] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys, "3d model matching with viewpoint invariant patches (vips)," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[5] M. Cummins and P. Newman, "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008. [Online]. Available: http://ijr.sagepub.com/cgi/content/abstract/27/6/647

[6] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "Fast and incremental method for loop-closure detection using bags of visual words," *Robotics, IEEE Transactions on*, vol. 24, no. 5, pp. 1027–1037, Oct. 2008.

[7] Y. Ohta and T. Kanade, "Stereo by intra- and inter-scanline search using dynamic programming," vol. PAMI-7, no. 1, pp. 139–154, March 1985.

[8] D. Scaramuzza, F. Fraundorfer, M. Pollefeys, and R. Siegwart, "Closing the loop in appearance-guided structure-from-motion for omnidirectional cameras," in *The Eight Workshop on Omnidirectional Vision , ECCV 2008*, 2008, pp. 1–14.

[9] D. Ortín and J. M. M. Montiel, "Indoor robot motion based on monocular images," *Robotica*, vol. 19, no. 3, pp. 331–342, 2001.

[2]http://www.cvg.ethz.ch/people/postgraduates/fraundof/vocsearch