Visual localization using global visual features and vanishing points

Olivier Saurer, Friedrich Fraundorfer, and Marc Pollefeys

Computer Vision and Geometry Group,
ETH Zürich, Switzerland
{saurero,fraundorfer,marc.pollefeys}@inf.ethz.ch

Abstract. This paper describes a visual localization approach for mobile robots. Robot localization is performed as location recognition. The approach uses global visual features (e.g. GIST) for image similarity and a geometric verification step using vanishing points. Location recognition is an image search to find the most similar image in the database. To deal with partial occlusions, which lower image similarity and lead to ambiguity, vanishing points are used to ensure that a matching database image was taken from the same viewpoint as the query image from the robot. Our approach will assign a query image to a location learned from a training dataset, to an "Unknown" location or in case of too much uncertainty the algorithm would refrain from a decision. The algorithm was evaluated under the ImageCLEF 2010 RobotVision competition[1]. The results on the datasets of this competition are published in this paper.

Keywords: visual place recognition, semantic annotation of space, visual localization, vanishing points

1 Introduction

Recent approaches to visual robot localization using local image features and visual words proved to work very well [2, 8, 1, 3]. An underlying assumption for these methods however is, that one already collected images for all possible locations in a database. A scenario, where a database was created using images of one floor of a building and having the robot localize itself on a different floor of the building would be beyond the capabilities of these methods. This is exactly the scenario that was created for the ImageCLEF 2010 RobotVision competition [7]. The goal was to train the robot with locations (e.g. office, kitchen, printer room) from one floor, so that it can identify the corresponding locations on the other floor, where the locations differ in details like different chairs, different desks, different posters, different curtains, etc. In this paper we describe an approach that is targeted towards resolving this scenario. The approach works by using a global image descriptor that captures the large scale features of the location, but not the fine details. This would allow to match up two locations that share

[1] This approach was ranked 1^{st} in the ImageCLEF 2010 RobotVision competition.

Fig. 1. Two images from the location class 'Meetingroom' on different floors. To identify these two images as matching an image descriptor needs to identify the large scale similarities (room configuration, table position) despite the obvious differences on the small scale.

the similar overall structure but differ on the fine details. Fig. 1 illustrates this concept. The two images show two meeting rooms from the different floors. The table, chairs and pictures on the wall are different but the overall structure is similar. There is a table in the center of the room, which creates a strong horizontal edge feature. The outline of the room walls itself creates also strong edge features converging in a similar manner. These are the features that we would like to capture. To achieve this our approach uses GIST [6] as a global visual descriptor. In addition to visual similarity we propose a subsequent geometric verification check. For geometric verification we compare the vanishing points of matching images, which are computed from line features in the images. This geometric check ensures, that images are matched up only, if they are taken in the same geometric setting (e.g. a similar sized room) and from the same viewpoint.

In the experiments using the dataset of the ImageCLEF 2010 RobotVision competition [7] we demonstrate that using GIST it is possible to capture these larger scale similarities and that it is possible to match up the locations like the one depicted in Fig. 1. We also show that the vanishing points are useful for geometric verification and improve the localization results. Finally we report the scores achieved in the ImageCLEF 2010 RobotVision competition.

2 Related Work

The GIST descriptor used in our approach was first introduced by Oliva *et al.* in [6]. It was used in [9] for place and object recognition. They showed that it is possible to distinguish between different places or rather scenes using the GIST descriptor. In particular they presented classification results on the following scenes: building, street, tree, sky, car, streetlight, person. In our current work we show that it is possible to use GIST for place recognition in typical indoor environments. In addition we added a geometric verification step targeted to indoor environments. GIST was also used in [5] for place recognition using

panoramic images. There the GIST descriptor was adapted to the properties of panoramic images.

3 GIST descriptor and Vanishing Points

Before presenting our pipeline for semantic labeling of space, we first discuss the GIST descriptor which was introduced by Oliva *et al.* in [6]. The GIST descriptor represents scenes from the encoding of the global configuration, ignoring most of the details and object information present in the scene. We then further discuss the concept of vanishing points, which are projections of points laying at infinity. They provide information on the relative camera orientation with respect to the scene and are used as a geometric verification after image retrieval using the GIST descriptor.

3.1 GIST descriptor

The GIST descriptor was proposed by Oliva *et al.* in [6] for scene categorization without the need for segmentation and processing of objects. The structure of the scene is estimated using a few perceptual dimensions such as naturalness, openness, roughness, expansion, ruggedness which describe the spatial properties of the scene. The dimensions are reliably estimated using spectral and coarsely localized information, where membership in semantic categories such as streets, highways, etc. are projected close together in a multidimensional space. The low dimensional representation of a scene is represented by a 960 dimensional descriptor, which allows quick retrieval of similar images from a large database. In the following, image search consists in finding the set of images with the smallest L2 distance.

3.2 Vanishing points

The premise to find vanishing points are man-made environments containing parallel straight lines. When looking at the perspective projection of three dimensional parallel lines, they intersect in one common point in the image plane, the so called vanishing point (VP) [4]. Vanishing points therefor represent the projections of 3D points laying at infinity, since parallel lines intersect at infinity.

To estimate the vanishing points, we first detect edges using canny edge detection and extract long straight lines from the edge segments. The straight lines are used as input for our RANSAC (random sample consensus) algorithm, which estimates multiple vanishing points in a single image. The algorithm first randomly selects two lines and computes their intersection point P. If at least 20 of the lines passes through the intersection point P, the point is re-estimated using a non-linear optimization, where all supporting lines are included in the optimization process. The supporting lines are then removed from the input set and the procedure is repeated until either no further lines are available or no further vanishing point is found.

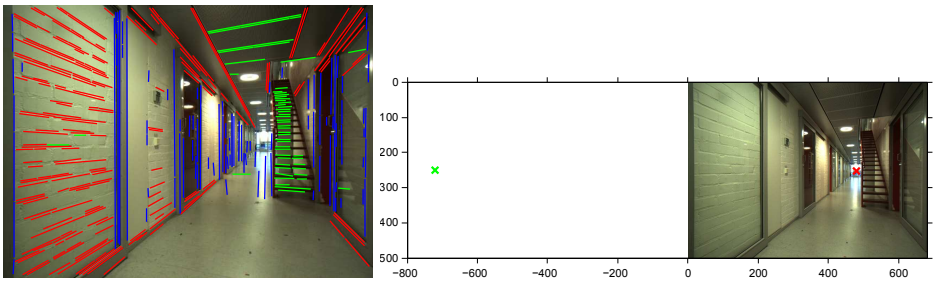


Fig. 2. Left, lines are classified to the according VPs. On the right the VPs are shown, except the one located far from the image center (infinity).

Fig. 2 shows line sets supporting different VPs. Each color represents the support of a different VP.

4 Place recognition

The proposed pipeline for semantic labeling of space is illustrated in Fig. 3. The method classifies an image into one of the following three categories, which is either a label learned from a training dataset, the "Unknown" label or in some cases the algorithm would make no decision.

In a first step a database of GIST descriptors is build from the training dataset. Our database consists of 4780 images and is represented by a kd-tree for fast k-nearest neighbors search, we chose k to be 10 in our experiments. In a first step we query the database with the query image q, for its 10 nearest neighbors stored in the result set r. Images in the result set r with a L2 distance to the query image q, greater than a given threshold (0.6 in our experiments) are removed from r. If r is empty, the image q is labeled as "Unknown". Otherwise the set r is further matched to a set of ambiguous images, which were previously learned from the training dataset, see Fig. 4. If the set of ambiguous images in the set r is greater than the set of non-ambiguous images, the algorithm refrains a decision on the image q, due to lack of confidence. Otherwise, a geometric verification is applied to the remaining set of non-ambiguous images. The geometric verification compares the angular distance of vanishing points between the query image and the non-ambiguous images. Images with a large angular distance (0.34 in our experiments) are removed from the set r. Finally, the query image is assigned the label of the image with the smallest angular distance or is assigned the label "Unknown" if the set r is empty.

To find vanishing point matches between two images, we first normalize the VP vector to unit length, such that the VP lays on the surface of a Gaussian sphere. Then, for each VP in one image we do an exhaustive search for the closest VP in the other image i.e., the VP with smallest angular distance. We assume that two similar scenes match, if their appearance is similar i.e., similar GIST

descriptor and similar vanishing points, meaning the camera has a similar point of view of the 3D scene being observed.

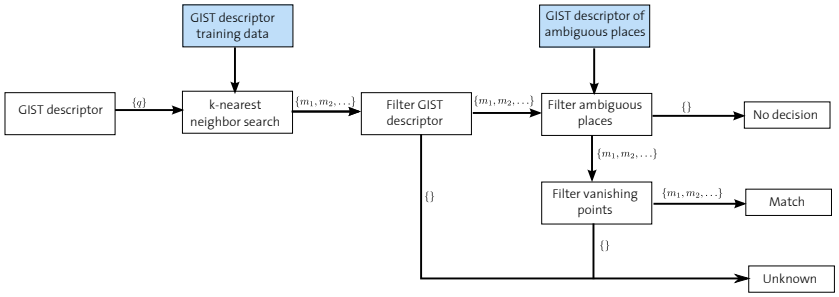


Fig. 3. Overview of the proposed pipeline.

5 Evaluation

Our algorithm was evaluated at the 3rd edition of the Robot Vision challenge, held in conjunction with IROS 2010. The challenge addresses the problem of classifying rooms and functional areas based on a pair of stereo images. Three image sets were provided, one training set for learning, one validation set for the participants to validate their algorithm and one testing set used for the competition. All three sets were captured in the same building, but on different floors. All three floors have a set of common rooms, such as *Offices, Toilet, Printer Area, Corridor, etc.* and rooms which are only present in one of the dataset such as *Kitchen, Lab, Elevator, etc.*. Sample images of the training sets are provided in Fig. 5.

Task 1 of the competition asked to build a system which can answer the question "Where am I?", given one pair of stereo images. The answer can either be a previously learned label, the "Unknown" label if the system is presented with a new location not contained in the training set or it can refrain a decision by leaving the image unclassified. The performance of the algorithm was evaluated using the following scoring scheme:

- +1.0 point for each correctly classified image.
- -1.0 point for each misclassified image.
- 0.0 point for each image that was not classified.
- +2.0 points for a correct detection of unknown category.
- -2.0 points for an incorrect detection unknown category.

Our system ranked first, with 677 points in the 3rd edition of the Robot Vision challenge. The winning run used the following configuration: a search window size of 10 images, a minimum GIST distance threshold of 0.6, and

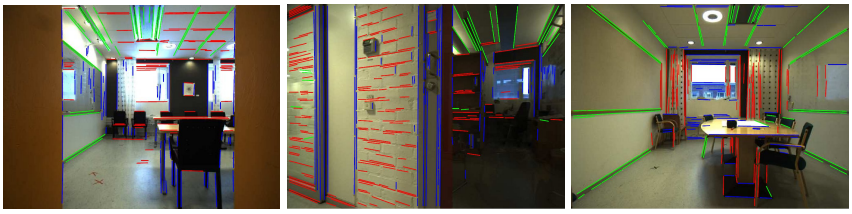
Fig. 4. Ambiguous places are represented by similar GIST descriptors with different image labels. We have trained two classes of ambiguous images, door frames (left image) and walls, whiteboards (right image).

Fig. 5. Sample images from the training dataset.

a minimum mean angular distance threshold of 0.34. Door frames, walls and whiteboards were learned and added to the ambiguous location set as well as the following four rooms *Kitchen, Small Office* and *Large Office*.

Bellow we further discuss the benefit of the geometric verification. The evaluation is based on the validation set, which contains 2069 images, where 14.4% of the image labels are unknown to the training set. Without geometric verification an image match is obtained by searching the training set for the image with the smallest L2 GIST distance. Using the geometric verification an image match is obtained by choosing the image with the smallest mean angular distance between the query image and the images obtained from the k-nearest neighbors, with $k = 30$. Table 1 lists the recognition rate of each category known to the training set. Overall, the geometric verification performed slightly superior (recognition rate of 43.15%) to the pure GIST based method (recognition rate of 42.03%). The *Meeting Room* category achieved an improvement of over 8%.

For the label *Corridor* and *Large Office* the pure GIST method performs better. The reason our method provides a lower performance on the *Corridor*



Corridor *a*) | Corridor *b*) | Meeting Room *c*)

Fig. 6. False image match due to ambiguous labeling of the training set. *a*) shows the original query image. *b*) shows the image with the smallest GIST distance 0.41 and a mean angular distance of 0.28. *c*) shows the image with the smallest angular distance, GIST distance 0.46 and a mean angular distance of 0.01.

Corridor *a*) | Small Office *b*) | Corridor *c*)

Fig. 7. Correct image match after geometric verification. *a*) shows the original query image. *b*) shows the image with the smallest GIST distance, 0.39 and a mean angular distance of 0.37. *c*) shows the image with the smallest angular distance, GIST distance 0.47 and a mean angular distance of 0.01.

category is that many images are misclassified at transitional places, where the robot moves from the corridor into a room. Fig. 6 illustrates such a misclassification. In the *Large Office* category the misclassified images are mainly classified as *Kitchen* or as *Small Office*. Fig. 7 illustrated a misclassification based on the GIST method, which is corrected by the geometric verification.

Our unoptimized Matlab implementation takes 51.21 seconds on a 2.66GHz Core2 Quad CPU, to classify 2069 images using precomputed GIST descriptors and precomputed vanishing points. Extracting GIST descriptors takes in average 1.91 seconds on a 487×487 pixel image. We make use of the freely available Matlab code provided by *Antonio Torralba*[2]. We use our own Matlab implementation to extract vanishing points. In average it takes 0.65 seconds to extract the vanishing points of one image.

[2] http://people.csail.mit.edu/torralba/code/spatialenvelope/

	Without Geometric Verification	With Geometric Verification
Corridor	74.31%	72.85%
Kitchen	0.00%	0.00%
Large Office	23.41%	19.93%
Meeting Room	44.44%	52.77%
Printer Area	40.21%	40.21%
Recycle Area	47.94%	52.05%
Small Office	14.50%	20.54%
Toilet	84.61%	91.20%

Table 1. Recognition rate obtained from the validation dataset. Note that the validation set does not hold the label *Kitchen*, therefor the recognition rate for that label is 0.00%. See text for more details.

6 Conclusion

We have presented a system for visual localization using global visual features (GIST) and a geometric verification based on vanishing points. We have shown that the geometric verification can indeed improve the recognition rate when used together with global visual features. The evaluation on the ImageCLEF 2010 RobotVision dataset showed that the approach manages to recognize similar locations despite of differences on the small scale. The evaluation however also revealed that the approach has difficulties in handling 'Unknown' locations. 'Unknown' locations are sometimes matched with locations from the training set and known locations are sometimes classified as 'Unknown' locations.

Acknowledgments. We would like to thank Georges Baatz for sharing his vanishing point detection code.

References

1. Angeli, A., Filliat, D., Doncieux, S., Meyer, J.A.: Fast and incremental method for loop-closure detection using bags of visual words. Robotics, IEEE Transactions on 24(5), 1027–1037 (Oct 2008)
2. Cummins, M., Newman, P.: FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. The International Journal of Robotics Research 27(6), 647–665 (2008), http://ijr.sagepub.com/cgi/content/abstract/27/6/647
3. Fraundorfer, F., Wu, C., Pollefeys, M.: Combining monocular and stereo cues for mobile robot localization using visual words. In: Proc. International Conference on Pattern Recognition. pp. 1–4 (2010)
4. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge (2000)
5. Murillo, A.C., Kosecka, J.: Experiments in place recognition using gist panoramas. In: IEEE Workshop on Omnidirectional Vision, Camera Netwoks and Non-Classical Cameras, ICCV 2009. pp. 1–8 (2009)

6. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (May 2001)
7. Pronobis, A., Fornoni, M., Christensen, H.I., Caputo, B.: The robot vision task at imageclef 2010. In: In the Working Notes of CLEF 2010, Padova, Italy (2010)
8. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, Minnesota. pp. 1–7 (2007)
9. Torralba, A.B., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: ICCV. pp. 273–280 (2003)