Robust Place Recognition with Combined Image Descriptors

Martin Dörfler[1(✉)] and Libor Přeučil[2]

[1] Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
martin.dorfler@fel.cvut.cz
[2] Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Prague, Czech Republic
libor.preucil@ciirc.cvut.cz

Abstract. In this paper, a method of place recognition is presented. The method is generally classified under the bag-of-visual-words approach. Information from several global image descriptors is incorporated. The data fusion is performed at the feature level.

The efficacy of the combined descriptor is investigated on the dataset recorded from a real robot. To measure the composition effect, all component descriptors are compared along with their combinations. Information on computational complexity of the method is also detailed, although the algorithms used did not undergo a big amount of optimization. The combined descriptor exhibits greater discriminative power, at the cost of increased computational time.

Keywords: Visual place recognition · Robust image features · Bag of visual words

1 Introduction

In the task of visual navigation of mobile robots, the first problem encountered is often the initialization. Determining the initial location of the robot in the environment is necessary precondition for further navigation, as well as a possible recovery method in cases where the navigation experiences a failure. For the robot equipped with a camera, a possible approach consists of calculating a similarity of observed environment appearance with the appearance of known locations, and selecting the closest match as a probable robot location. This task is called place recognition.

In preceding years, good results were obtained using robust features detected in the image. Variety of image features were investigated in the literature. The

L. Přeučil—The presented research was supported by the Czech Technical University in Prague under grant SGS16/160/OHK3/2T/13, by the Technology Agency of the Czech Republic under the project No. TE01020197 Centre for Applied Cybernetics, and by Horizon 2020 program under the project No. 688117 "Safe human-robot interaction in logistic applications for highly flexible warehouses".

J. Hodicky (Ed.): MESAS 2016, LNCS 9991, pp. 196–203, 2016.
DOI: 10.1007/978-3-319-47605-6_16

different types of local image features contain different information. By employing complemental image features, robustness of the place recognition algorithm can be increased and uneven density of features in the environment compensated.

Furthermore, in a feature-starved or highly self-similar environment, the presence of local features is insufficient to distinguish the locations from one another. A global descriptor might be able to characterize high-level differences not present at the level of singular details.

One possible approach would be switching descriptors used in the case of lower localization quality. A possible problem is that an abrupt switching of localizing method would introduce discontinuities in the localization result. Instead, we decided to use a combined descriptor, containing a BoW (bag of visual words) representation and global image descriptors. A distance in the combined descriptor space is used as a similarity measure.

As a work in progress, this paper presents the combined descriptor approach. Several global features were investigated, and effects of their inclusion experimentally evaluated. The approach was compared to plain BoW visual place recognition. Comparisons were performed on outdoor data collected by a real robot.

2 Previous Work

The topic of appearance-based localization has been extensively studied and much work on this topic is available. Many results have been obtained by using local image features. These approaches identify salient regions of the image that can be identified despite changes in illumination, scale and point of view. Furthermore, given their locality, such methods are not thwarted by occlusion of the part of the image, as the local features in the remaining part are not disturbed.

There are many variants of local image descriptor. SIFT [1], SURF [2] and similar methods [3] detect salient points as a peak in the image gradient. Works [4] or [5] detect image regions stable to change in the scale.

A common weakness of the local image features is their quantity. With several hundred features per image, matching against large image sets gets prohibitively expensive. Bag-of-words techniques (BoW) circumvent this problem by building a vocabulary of common image features [6]. Presence or absence of these features is a descriptor of the image. While this descriptor is of a global character, advantageous properties of local descriptors (such as invariance, or resistance to occlusion) are partially preserved.

A successful example of such approach is the method FabMap, presented in [7]. In addition to the use of a visual dictionary, it achieves additional performance gains by modeling feature probability by Chow-Liu tree. Realtime recognition is reported even on large datasets [8].

Another approach for appearance-based localization consists of method to characterize the whole image. The so-called *global descriptors* use features such as lies, edges, or gradient of image function to make a compressed representation

of the image frame. Various properties of the image have been used for this task. For example, in [9] the balance of color components of the image is used.

A GIST method proposed in [10] divides the frame in 16 parts and extracts prevailing gradient direction in each by applying Gabor filters. The approach was later refined in [11] by using PCA to reduce the dimensionality of the descriptor, as well as in [12] by exchanging the gradient direction by BRIEF descriptor.

Some of these methods are of deeper relevance to this paper. These are covered in more detail in the Sect. 3.1 and following.

There have been previous attempts to combine local and global descriptors. Most cited employ a multi-step approach, using one method to pre-select likely candidates for further processing by a more computationally costly steps. For example, in [13] a global color descriptor was used to select candidates for subsequent matching based on line features described by their line support regions. The authors of [14] use saliency measure to select interest regions of the image to save the costly computation of the gist descriptors.

In the following section, our approach employing combination of local and global features is presented.

3 Proposed Approach

To perform a place recognition in locations where local image features may be scarce, a different kind of information needs to be incorporated in the decision-making. Our aim is to extract such information in form of global image descriptors, and integrate it seamlessly into the place recognition algorithm. The objective is to perform the fusion at the earliest possible point in the pipeline. This way, it is possible to take advantage of later parts of the pipeline which handle the dependency between the features. Also, fitting in the common framework means the proposed method can be without much difficulty meaningfully compared to its predecessor.

For this reason, we propose using a pipeline similar to the one used in the place recognition by a bag-of-words approach. In the first steps of image processing, the local image features are extracted. During the training, a set of commonly appearing features are selected. Their presence or absence will henceforth be used as features identifying respective image frames. During the recognition, a vector identifying the presence of all the relevant features is passed to the algorithm and the most similar of the learned images is selected. Optionally, some weighting or feature distribution model is used to balance the fact that the presence of features is not independent event and some normalization can improve the results.

To extend this model, more features are introduced in the descriptor. In the image processing stage, a set of global image descriptors is extracted from the image. These descriptors are converted in a vector of features, which are then included in the descriptor constructed in the bag-of-words calculating stage. Proper normalization is performed to maintain similar range of values for all parts of the descriptor.

One advantage of so constructed descriptor: the use is straightforward. The distance in the combined descriptor space can be used as a (dis)similarity measure for the purposes of place recognition.

Furthermore, when considering the combined descriptor, the additional members play a very similar role to the already existing ones. They are simply another features, describing the scene. It is thus equally straightforward to integrate the combined descriptor into any system based on a set of features. In this paper, such approach was demonstrated with FabMAP [7,8].

In the original implementation, FabMAP is using a bag-of-visual-words approach to place recognition. Improvements are obtained by modeling the conditional dependence of the features in the descriptor, which correspond to the prevalence of vocabulary landmarks in the target scene. This approach can also be applied to the global features, when properly normalized.

Following sections detail the global image descriptors selected for making the combination descriptor. At this stage, the main criteria of selection were the diversity of underlying principle, ease of implementation, and sensitivity to the rotation. As stated previously, many global image descriptors are influenced by the precise position and orientation of the robot at the time of taking the picture. Only a minority exhibits rotational invariance. As this property is desirable and possible with the rest of the algorithm, global descriptors were selected that can provide it also.

3.1 Color Histogram

One of the first global image descriptors in use is a color histogram [9]. It describes the image by its most straightforward characteristic - the pixel values.

A simple variant of the color histogram shows the relative prevalence of the three color components, quantized to the fixed scale. A more involved approach consists of expanding beyond the RGB color space. HSL model is a natural candidate here. It's components are meaningful for perceiving humans, and the transformation between RGB and HSL is nonlinear, thus beyond abilities of linear model to learn. Including these values thus brings new information in the recognition task.

The resulting descriptor size is determined by the number of the color channels, and the size of histogram bins. The number of the bins is usually kept low for better generalization. Otherwise, filtering the resultant histogram may be necessary to prevent the effects of the noise.

3.2 Edge Histogram Descriptor

Another way of characterizing the image is by describing its texture. The methods, collectively called *texture descriptors*, attempt to describe the image in the terms of patterns and their prevalence in the different parts of the image.

For this work, a method called Edge Histogram Descriptor [15] was chosen. It performs the task by cataloguing the prevailing direction of the edges present in

the image. This is done by processing the image by a bank of convolution filters and looking for the highest response at each pixel. The distributions of the responses are summarized in the histogram. Depending on the implementation, the descriptor is formed by the histogram of the whole image, several histograms of various subregions of the image, or some combination.

To maintain the rotational invariance, only the whole image histogram is usable in this case. Relaxing this limitation to the invariance to rotation along the z axis (panning), it is possible to employ subregions that are horizontal slices of the image.

4 Experiments

To investigate the benefits of extended descriptors, their efficacy was tested on the data recorded with the real robot. A place recognition was performed, first using the unmodified BoW descriptor (i.e. running FABMAP2 algorithm as implemented in the openFABMAP [8] project), color histogram descriptor, edge histogram descriptor, and possible combinations.

Computing this additional information increases the processing time necessary for each frame, as well as the computational complexity of the matching process. This cost is fairly straightforward, time to process each frame is increased by the amount needed for each descriptor calculation. Please note that the algorithms used are generally not optimized, and these values serve only for relative comparison, not as a definitive statement on the method efficacy. For that reason, relative time requirements are shown side-by-side with the actual values in the Table 1. As we can see, the increase in the processing time is modest, between 5 % and 7 %.

As the first step, the properties of extended descriptor as a dissimilarity measure were investigated. The descriptor was constructed according to the Sect. 3. The relation between distance in the real world and the distance in the descriptor space is shown on the Fig. 1. The inclusion of additional information has changed the shape of graph, making the two variables more dependent. Thus, the utility of the descriptor has increased and it is better representative of real world differences.

Table 1. Computation time to construct descriptors

Descriptor	Computation time [s]	Relative computation time
BoW	0.834	1.0
Color histogram	0.014	0.016
Edge density histogram	0.037	0.044
BoW + color histogram	0.879	1.054
BoW + edge histogram	0.880	1.055
Full descriptor	0.894	1.071

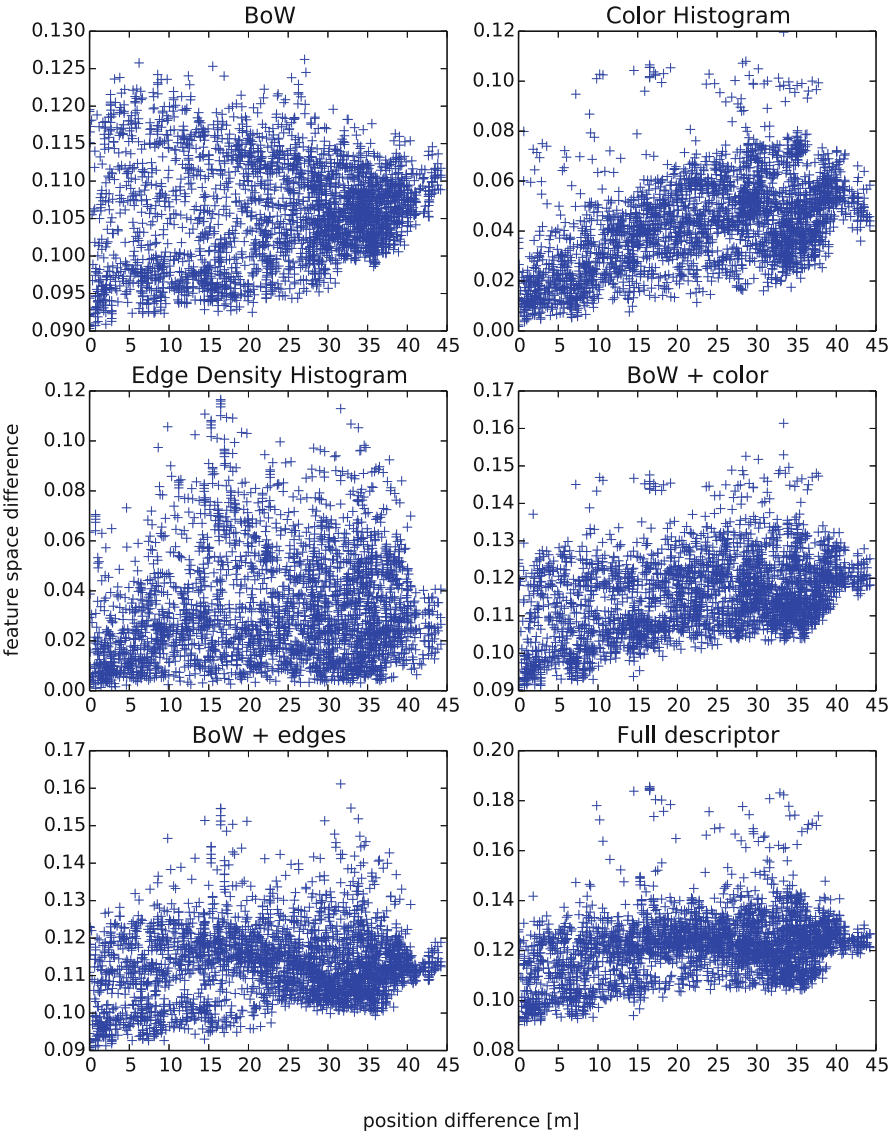

Fig. 1. The real-world position difference plotted against distance in the feature space. The feature space distance is an abstract measure without direct physical interpretation.

The effectiveness of the combined descriptor in a complete place recognition pipeline was tested in the next step. From the descriptors of target locations, a map was built and consequently used to identify the most similar place for each frame in the experiment. As in the previous step, the descriptors and their combinations have been investigated separately.

Table 2. Accuracy of the place recognition. Also reported is the confidence of the algorithm, calculated as the median relative likelihood of the true solution, compared to next most probable candidate. High value imply greater measure of certainty. Confidence 1.0 indicates case where several candidates share the first position.

Descriptor type	Recognition accuracy	Confidence
BoW	0.515	555.7
Color histogram	0.330	1.0
Edge density histogram	0.208	1.0
BoW + color histogram	0.689	719.78
BoW + edge histogram	0.679	719.6
Full descriptor	0.689	719.78

The Table 2 show that the stand-alone performance of the selected global descriptors is not very good under the conditions of the experiment. Nevertheless, incorporating each of them provides considerable benefits. The combination of the BoW descriptor with any of the two global descriptors exhibits greater recognition accuracy. The results obtained with color histogram are slightly better. Combining all three descriptors does not provide further increases in the recognition accuracy or confidence. Hence, it seems advisable to use BoW descriptor extended by a color histogram.

5 Conclusion

In this paper, a method of place recognition is presented. An image descriptor is constructed by combining a bag-of-words approach and several global image descriptors. The probability dependency of the descriptor components is modeled by a Chow-Liu tree. We have found such descriptor to be a good similarity measure for place recognition. Its discriminative powers are greater than that of the component descriptors. The disadvantage is a modest increase in the computing time.

The improvements in discrimination and time requirements were investigated and results presented for comparison. Investigation was performed on data collected by the real robot.

The result show that introducing additional features in the place descriptor increases the discriminative power. There is a trade-of in accuracy and computational complexity, but the gains clearly outweigh the costs, at least in the investigated conditions. The selection of particular descriptors and their efficacy specifically in such difficult conditions will need to be a subject of further investigation. Also, the combined descriptor does not preclude the use of the component descriptors individually, to speed up the recognition in the uncomplicated cases. Efficacy of such setup is yet to be evaluated.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**, 91–110 (2004)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., Gool, L.V., Baya, H., Essa, A., Tuytelaarsb, T., Van Goola, L.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**, 346–359 (2008)
3. Agrawal, M., Konolige, K., Blas, M.R.: CenSurE: center surround extremas for realtime feature detection and matching. In: ECCV 2008, IV, pp. 102–115 (2008)
4. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximaly stable estreal regions. In: Proceeding of the Britsh Machine Vision Conference, pp. 384–393 (2002)
5. Chung, J., Kim, T., Nam Chae, Y., Yang, H.S.: Unsupervised constellation model learning algorithm based on voting weight control for accurate face localization. Pattern Recogn. **42**, 322–333 (2009)
6. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: 2003 Proceedings of Ninth IEEE International Conference on Computer Vision, vol. 2, pp. 1470–1477 (2003)
7. Cummins, M., Newman, P.: Fab-map: probabilistic localization and mapping in the space of appearance. Int. J. Robot. Res. **27**, 647–665 (2008)
8. Cummins, M., Newman, P.: Appearance-only slam at large scale with FAB-MAP 2.0. Int. J. Robot. Res. **30**, 1100–1123 (2011)
9. Ulrich, I., Nourbakhsh, I.: Appearance-based place recognition for topological localization. In: Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065), vol. 2, pp. 1023–1029 (2000)
10. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**, 145–175 (2001)
11. Liu, Y., Zhang, H.: Visual loop closure detection with a compact image descriptor. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1051–1056 (2012)
12. Snderhauf, N., Protzel, P.: Brief-gist - closing the loop by simple means. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1234–1241 (2011)
13. Murillo, A.C., Guerrero, J.J., Sagues, C.: Surf features for efficient robot localization with omnidirectional images. In: Proceedings of 2007 IEEE International Conference on Robotics and Automation, pp. 3901–3907 (2007)
14. Siagian, C., Itti, L.: Biologically inspired mobile robot vision localization. IEEE Trans. Robot. **25**, 861–873 (2009)
15. Park, D.K., Jeon, Y.S., Won, C.S.: Efficient use of local edge histogram descriptor. In: Proceedings of the 2000 ACM Workshops on Multimedia, MULTIMEDIA 2000, pp. 51–54. ACM, New York (2000)