Targeting a Practical Approach for Robot Vision with Ensembles of Visual Features

Emanuela Boroş

Alexandru Ioan Cuza University
Faculty of Computer Science, Iaşi, Romania
emanuela.boros@info.uaic.ro

Abstract

We approach the task of topological localization in mobile robotics without using a temporal continuity of the sequences of images. The provided information about the environment is contained in images taken with a perspective colour camera mounted on a robot platform. The main contributions of this work are quantifiable examinations of a wide variety of different global and local invariant features, and different distance measures. We focus on finding the optimal set of features and a deepened analysis was carried out. The characteristics of different features were analysed using widely known dissimilarity measures and graphical views of the overall performances. The quality of the acquired configurations is also tested in the localization stage by means of location recognition in the Robot Vision task, by participating at the ImageCLEF International Evaluation Campaign. The long term goal of this project is to develop integrated, stand alone capabilities for real-time topological localization in varying illumination conditions and over longer routes.

1998 ACM Subject Classification I.2.10 Vision and Scene Understanding, I.4.3 Enhancement, I.4.6 Segmentation, I.4.10 Image Representation

Keywords and phrases Visual Place Classification, Robot Topological Localization, Visual Feature Detectors, Visual Feature Descriptors

Digital Object Identifier 10.4230/OASIcs.ICCSW.2012.22

1 Introduction and Related Work

Topological localization is a fundamental problem in mobile robotics. Most mobile robots must be able to locate itself in their environment in order to accomplish their tasks. Robot visual localization and place recognition are not easy tasks, and this is mainly due to the perceptive ambiguity of acquired data and the sensibility to noise and illumination variations of real world environments. In order to help reduce this gap, this work addresses the problem of topological localization of a robot that uses a single perspective camera in an office environment. The robot should be able to answer the question *where are you?* when presented with a test sequence representing a room category seen during training [25, 28, 17].

Many approaches during last years have been developed using different methods for robotic topological localization such as topological map building which makes good use of temporal continuity [30], simultaneous localization and mapping [8], using Monte-Carlo localization [32], appearance-based place recognition for topological localization, panoramic vision creation [31].

The problem of topological mobile localization has mainly three dimensions: a type of environment (indoor, outdoor, outdoor natural), a perception (sensing modality) and a localization model (probabilistic, basic). Numerous papers deal with indoor environments [30, 31, 10, 15] and a few deal with outdoor environments, natural or urban [29, 13]. Experimental

2012 Imperial College Computing Student Workshop (ICCSW'12).
Editor: Andrew V. Jones; pp. 22–28
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany
ICCSW

results for wide baseline image matching suggest the need for local invariant descriptors of images. Earlier research into invariant features focused on invariance to rotation and translation. There has also been research into the development of fully invariant features [5, 18, 19]. In his milestone paper [16], D. Lowe has proposed a scale-invariant feature transform (SIFT) for recognition based on local extrema of difference-of-Gaussian filters in scale-space that is invariant to image scaling and rotation, illumination and viewpoint changes. Lately, a new method has been proposed, Affine-SIFT (ASIFT) that simulates all image views obtainable by varying the two camera axis orientation parameters, namely the latitude and the longitude angles, left over by the SIFT method [21]. However, full affine invariance has not been achieved due partly to the impractically large computational cost. SIFT is a 128 dimensional feature vector that captures the spatial structure and the local orientation distribution of a region surrounding a keypoint. The SIFT method has been popularly applied for scene recognition [33, 1] and detection [11, 23] and robot localization [2, 24, 22].

We analyze the problem of topological localization without taking in consideration the use of the temporal continuity of the sequences of images which could be considered an advantage by adding an additional understanding of the space. Our approach represents an extension of our previous work [3, 4] where each RGB image is processed to extract sets of SIFT keypoints from where the descriptors are defined. In this paper the comparison is carried out for different configurations of features and matching distances of a topological localization system. We perform an exhaustive evaluation and introduce new analysis statistics between the quantization solutions.

2 Experimental Setup

2.1 Feature Matching

In this section we introduce different dissimilarity measures to compare features. That is, a measure of dissimilarity between two features and thus between the underlying images is calculated. Many of the features for images are in fact histograms (color histograms, invariant feature histograms, texture histograms, local feature histograms, and other feature histograms). As comparison of distributions is a well known problem, a lot of comparison measures have been proposed and compared before [26]. In the following, dissimilarity measures to compare two histograms H and K are proposed. Each of these histograms has n bins and H_i is the value of the i-th bin of histogram H.

- **Minkowski-form Distance** (L_1 distance is often used for computing dissimilarity between color images, also experimented in color histograms comparison [14]):

$$\mathrm{D}_{Lr}(H,K) = (\sum_{i=1} |H_i - K_i|)^{\frac{1}{r}} \tag{1}$$

- **Jensen Shannon Divergence** (also referred to as **Jeffrey Divergence** [9], is an empirical extension of the Kullback-Leibler Divergence. It is symmetric and numerically more stable):

$$\mathrm{D}_{JSD}(H,K) = \sum_{i=1} H_i \log \frac{2H_i}{H_i + K_i} + K_i \log \frac{2K_i}{K_i + H_i} \tag{2}$$

- χ^2 **Distance** (measures how unlikely it is that one distribution was drawn from the population represented by the other, [20]):

$$\mathrm{D}_{\chi^2}(H,K) = \sum_{i=1} \frac{(H_i - K_i)^2}{H_i} \tag{3}$$

- **Bhattacharyya Distance [7]** (measures the similarity of two discrete or continuous probability distributions). For discrete probability distributions H and K over the same domain, it is defined as:

$$\mathrm{D}_B(H,K) = -\ln \sum_{i=1} \sqrt{(H_i K_i)} \tag{4}$$

2.2 Datasets (Benchmark)

The chosen dataset contains images from nine sections of an office obtained from **CLEF (Conference on Multilingual and Multimodal Information Access Evaluation)**. Detailed information about the dataset are in the overview and ImageCLEF publications [25, 28, 17]. This dataset contains images that are widely used in topological localization image classification papers and it has already been split into three training sets of images, as shown in Table 1 one different from another. The provided images are in the RGB color space. The sequences are acquired within the same building and floor but there can be variations in the lighting conditions (sunny, cloudy, night) or the acquisition procedure (clockwise and counter clockwise).

Areas	Training1	Training2	Training3
Corridor	438	498	444
ElevatorArea	140	152	84
LoungeArea	421	452	376
PrinterRoom	119	80	65
ProfessorOffice	408	336	247
StudentOffice	664	599	388
TechnicalRoom	153	96	118
Toilet	198	240	131
VisioConference	126	79	60

Table 1 Training Sequences of An Office Environment.

2.3 Comparison of Different Distance Functions for Global Features

Global features capture the diagnostic structure of the image, an overall view of the image that is transformed in histograms of frequencies. Existing color-based general-purpose image retrieval systems as [27, 6] roughly fall into three categories depending on the signature extraction approach used: histogram, color layout, and region-based search. In this paper, histogram-based search methods are investigated in two different color spaces, RGB (**R**ed, **G**reen, and **B**lue) and HSV (**H**ue, **S**aturation, and **V**alue). RGB and HSV color histograms are subject to tests with Jeffrey Divergence, χ^2, Bhattacharyya, Minkowski and respectively

the widely used Euclidean distance measure. These were chosen considering the literature that underlies them as achieving the best results in image matching [7, 26].

The retrieved classes for images (*Corridor*, *LoungeArea* etc.) depend on a threshold, those below this value being rejected. This becomes an optimization problem of finding the best value that will cut the unwanted results, considering that it is better to have no results than inconsistent results. To accomplish this, we used the genetic algorithm explained in detail in [12]. For these experiments, we used a population of 200 individuals, the mutation probability of 0.15, and the crossover, of 0.7. The optimization process is stopped after 1000 generations. We used a selection scheme *rank selection* with *elitism*. For RGB histograms, as can be seen

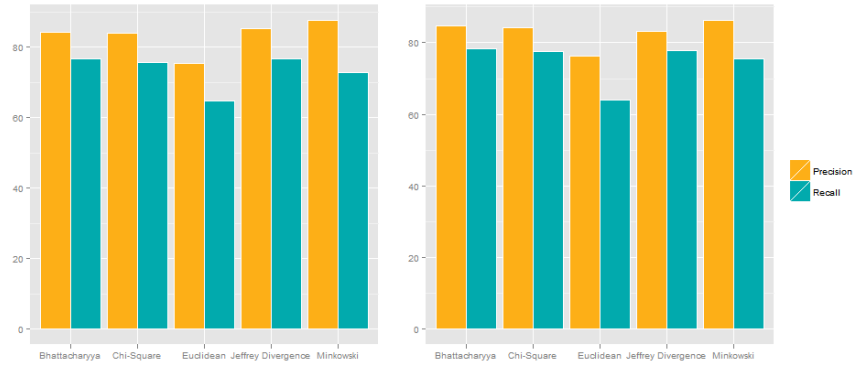

Figure 1 Precision and recall depending on measure distance (RGB & HSV Histograms).

in Figure 1, Bhattacharyya and Jeffrey Divergence obtained the highest recall and also, high precision, the highest F-measure being obtained by Jeffrey Divergence (0.806) extremely close to Bhattacharyya (0.802). The lowest performance is with Euclidean distance, having not only a low recall which means that this solution will bring more irrelevant results than using the other distances, but also a lower precision. In the case of using HSV histograms, the Bhattacharyya distance leaded to good results with a F-measure of 0.81 close to χ^2 distance with 0.807 and Minkowski with a F-measure of 0.805 . Folowing these chosen metrics, we adopted the vizualization with confusion matrices. Entries on the diagonal of the matrix, in blue, count the correct calls. Entries off the diagonal, in fading blue, count the misclassifications. Corresponding to the confusion matrix represented in Figure 2, the results show that HSV histogram with Bhattacharyya distance yielded very similar results with RGB choices of distances but clearly outperforms RGB histogram comparison with Jeffrey Divergence distance, similarity probabilty peaking at 100% in some of the office sections (*PrinterRoom, StudentOffice*).

2.4 Comparison of Different Distance Functions for Local Features

The two types of features used in the experiments are SIFT (Scale Invariant Feature Transform) and ASIFT (Affine Scale Invariant Feature Transform). The advantages of using these features are that they describe localized image regions (*patches*), the descriptors are computed around interest points, there is no need for segmentation and they are robust to occlusion and clutter. The disadvantage is that images are represented by different size sets of feature vectors and they do not lend themselves easily to standard classification techniques.

These results were obtained performing experiments on local feature histograms obtained using the *bag-of-visual-words* model. The descriptors are quantized and normalized. Different

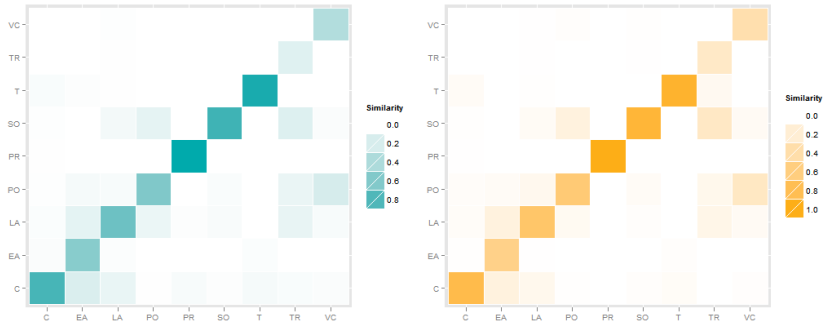

Figure 2 Confusion Matrix (RGB/HSV Histograms) using Jeffrey Divergence/Bhattacharyya Distances.

dissimilarity measures for the different types of features are compared experimentally and the performance for the different types of features is quantitatively measured.

For matthcing features, we chose literature-based distances known as having the best results: Euclidean, Minkowski, χ^2 and Jeffrey Divergence distances. For each of the local features descriptors we created Precision/Recall graphs from which we determine the superior runs. Figure 3 shows the Precision/Recall graphs for SIFT, respectively ASIFT and also shows that there is still vast room for improvement but the most promising results were obtained in the case of the usage of SIFT descriptors with Minkowski and Euclidean distance. The results show that Euclidean and Minkowski distance yielded very similar results, in the case of SIFT features matching.

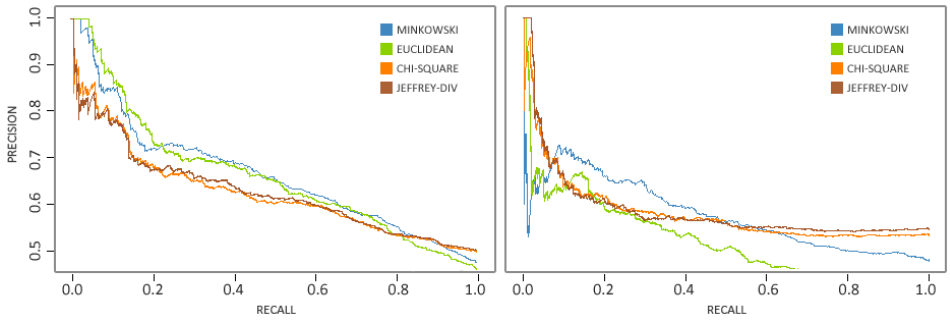

Figure 3 PR curves using different distance measures (SIFT & ASIFT).

3 Conclusions and Future Work

In this work, we approached the task of topological localization without using a temporal continuity of the sequences of images using a broad variety of features for image recognition. The provided information about the environment is contained in images taken with a perspective color camera mounted on a robot platform and it represents a know office environment dataset offered by ImageCLEF.

A large scale of global and local invariant features of images was presented, investigated, and experimentally evaluated. To analyze the features various dissimilarity measures were implemented and tested, as different features require different comparison methods.

The experiments show that the configurations from different feature descriptors and distance measures depends on the proper combinations. One important aspect is to use a selection of features accounting for the different properties of the images as there is no feature capable of covering all aspects of an image. The experiments showed the following features are suitable:

- RGB & HSV color histograms
- SIFT (Scale Invariant Feature Transform) as visual words with an Euclidean 100-means

The experiments showed also that the following image matching settings are suitable:

- RGB color histograms with Jeffrey Divergence distance & HSV color histograms with Bhattacharyya distance
- SIFT (Scale Invariant Feature Transform) matched with Minkowski distance

From the fact that most of the works cited are from the last couple of years, topological localization is a new and active area of research. which is increasingly producing interest and enforces further development. A first starting point for this field is given in this thesis, along with notable experimental results, but there is still room for improvement and further research.

References

1 M. Bennewitz, C. Stachniss, W. Burgard, and S. Behnke. Metric localization with scale-invariant visual features using a single perspective camera. *European Robotics Symposium 2006, ser. STAR Springer tracts in advanced robotics*, 22, 2006.

2 M. Bennewitz, C. Stachniss, W. Burgard, and S. Behnke. Metric localization with scale-invariant visual features using a single perspective camera. *European Robotics Symposium 2006*, 22 of STAR Springer tracts in advanced robotics:143–157, 2006.

3 E. Boroş, G. Roşca, and A. Iftene. Uaic: Participation in imageclef 2009 robotvision task. *Proceedings of the CLEF 2009 Workshop*, Sep 2009.

4 E. Boroş, G. Roşca, and A. Iftene. Using sift method for global topological localization for indoor environments. *Multilingual Information Access Evaluation II. Multimedia Experiments [Lecture Notes in Computer Science Volume 6242 Part II]*, 6242:277–282, 2009.

5 M. Brown and D.G Lowe. Invariant features from interest point groups. *The 13th British Machine Vision Conference, Cardiff University, UK*, pages 253–262, 2002.

6 R. Chakravarti. A study of color histogram based image retrieval. *Information Technology: New Generations, 2009. ITNG '09*, 2009.

7 E. Choi and C. Lee. Feature extraction based on the bhattacharyya distance. *Pattern Recognition*, 36:1703—1709, 2003.

8 H. Choset and K. Nagatani. Topological simultaneous localization and mapping (slam): toward exact localization without explicit localization. *IEEE Trans. Robot. Automat.*, 17(2):125–137, 2001.

9 T. Deselaers, D. Keysers, and H. Ney. Features for image retrieval: An experimental comparison. *Information Retrieval*, 2008.

10 G. Dudek and D. Jugessur. Robust place recognition using local appearance based methods. *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1030–1035, 2000.

11 G. Fritz, C. Seifert, M. Kumar, and L. Paletta. Building detection from mobile imagery using informative sift descriptors. *Lecture Notes in Computer Science*, pages 629–638, 2005.

12 A. L. Gînscă and A. Iftene. Using a genetic algorithm for optimizing the similarity aggregation step in the process of ontology alignment. *Proceedings, of 9th International Conference RoEduNet IEEE*, pages 118–122, Jun 2010.

13 J.-J. Gonzalez-Barbosa and S. Lacroix. Rover localization in natural environments by indexing panoramic images. *Proceedings of the 2002 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1365–1370, 2002.

14 A. B. Kurhe, S. S. Satonka, and P. B. Khanale. Color matching of images by using minkowski- form distance. *Global Journal of Computer Science and Technology, Global Journals Inc. (USA)*, 11, 2011.

15 L. Ledwich and S. Williams. Reduced sift features for image retrieval and indoor localisation. *Australasian Conf. on Robotics and Automation*, 2004.

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

17 W. Lucetti and E. Luchetti. Combination of classifiers for indoor room recognition, cgs participation at imageclef2010 robot vision task. *Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2010)*, 2010.

18 K. Mikolajczyk and C. Schmid. An afine invariant interest point detector. *Proceedings of the 7th European Conference on Computer Vision*, pages 128–142, 2002.

19 K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1), 2004.

20 K. Mikolajczyk, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition. *Visual Recognition Challange*, 2007.

21 J. Morel and G. Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.

22 A. Murarka, J. Modayil, and B. Kuipers. Building local safety maps for a wheelchair robot using vision and lasers. *Proceedings of the The 3rd Canadian Conference on Computer and Robot Vision*, 2006.

23 A. Negre, H. Tran, N. Gourier, D. Hall, A. Lux, and JL Crowley. Comparative study of people detection in surveillance scenes. structural, syntactic and statistical pattern recognition. *Proceedings Lecture Notes in Computer Science*, 4109:100–108, 2006.

24 D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. *CVPR*, 2:2161–2168, 2006.

25 A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt. Multi-modal semantic place classification. *Int. J. Robot. Res.*, 29(2-3):298–320, February 2010.

26 J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *Proc. International Conference on Computer Vision, Vol. 2*, pages 1165–1173, 1999.

27 J. Sangoh. Histogram-based color image retrieval. *Psych221/EE362 Project Report*, 2001.

28 O. Saurer, F. Fraundorfer, and M. Pollefeys. Visual localization using global visual features and vanishing points. *Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2010)*, 2010.

29 Y. Takeuchi and M. Hebert. Finding images of landmarks in video sequences. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 1998.

30 S. Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artificial Intelligence*, 99:21–71, February 1998.

31 I. Ulrich and I. Nourbakhsh. Appearance-based obstacle detection with monocular color vision. *Proceedings of AAAI Conference*, pages 866–871, 2000.

32 J. Wolf, W. Burgard, and H. Burkhardt. Robust vision-based localization for mobile robots using an image retrieval system based on invariant features. *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2002.

33 J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–238, Jun 2007.