UAIC participation at Robot Vision @ 2012 An updated vision

Emanuela Boroş, Alexandru Lucian Gînscă, and Adrian Iftene

Alexandru Ioan Cuza University, Faculty of Computer Science
General Berthelot, 16, 700483, Iaşi, Romania
{emanuela.boros,lucian.ginsca,adiftene}@info.uaic.ro

Abstract. In this paper we describe a system that participated in the fourth benchmarking activity ImageCLEF, in the Robot Vision task, for which we approach the task of topological localization without using a temporal continuity of the sequences of images. We provide details for the state-of-the-art methods that were selected: Color Histograms, SIFT (Scale Invariant Feature Transform), ASIFT (Affine SIFT) and RGB-SIFT, Bag-of-Visual-Words strategy inspired from the text retrieval community. We focused on finding the optimal set of features and a deepened analysis was carried out. We offer an analysis of the different features, similarity measures and a performance evaluation of combinations of the proposed methods for topological localization. Also, we detail a genetic algorithm that was used for eliminating the false positives results. In the end, we draw several conclusions targeting the advantages of using proper configurations of visual-based appearance descriptors, similarity measures and classifiers.

Keywords: Robot Topological Localization, Global Features, Invariant Local Features, Visual Words, SVMs, Genetic Algorithm

1 Introduction and Related Work

In this paper, we present an approach to vision-based mobile robot localization that uses a single perspective camera taken within 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 [30, 33, 25]. We analyze the problem without taking in consideration the use of the temporal continuity of the sequences of images. We perform an exhaustive evaluation and introduce a new analysis statistic between quantization techniques of a large set of features, from which different system configurations are picked and tested.

Traditionally, robot vision systems heavily relied on different methods for robotic topological localization such as topological map building which makes good use of temporal continuity [37], panoramic vision creation [38], simultaneous localization and mapping [7], appearance-based place recognition for topological localization [38], Monte-Carlo localization [41].

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 [37, 38, 10, 21] and a few deal with outdoor environments, natural or urban [36, 13].

Current work on robot localization in indoor environments has focused on introducing probabilistic models to improve local feature matching and the integration of specific kernels. Experimental results for wide baseline image matching suggest the need for local invariant descriptors of images. Invariant features have achieved relative success with object detection and image matching. There has also been research into the development of fully invariant features [4, 26, 27]. In his milestone paper [23], D. Lowe has proposed a scale invariant feature transform (SIFT) 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 the views obtainable by varying the two camera axis orientation parameters, namely the latitude and the longitude angles [29].

The *Bag-of-Visual-Words* [8, 12] model is a great addition to place recognition and was initially inspired by the *bag-of-words* models in text classification where a document is represented by an unsorted set of the contained words. This data modeling technique was first been introduced in the case of video retrieval [35]. Due to its efficiency and effectiveness, it became very popular in the fields of image retrieval and classification [20, 43].

The classification level of images relies more on unsupervised then supervised learning techniques. Categorizing in unsupervised learning scenarios is a much harder problem, due to the absence of class labels that would guide the search for relevant information. In supervised learning scenarios, image categorizing has been studied widely in the literature. Among supervised learning techniques, the most popular in this context are Bayesian classifiers [8, 18, 12, 19] and Support Vector Machines (SVM) [39, 8, 18, 44]. [3] also uses random forests. Actually, state-of-the-art results are due to SVM classifiers: the method described in [44] combines a local matching of the features and specific kernels based on the Earth Movers Distance [32] or χ^2 [28] yielded the best results.

Our approach represents an extension of our previous work [1, 2] where each RGB image is processed to extract sets of SIFT keypoints from where the descriptors are defined. Making use of global and local features, a quantization technique, SVMs and a genetic algorithm that aims at eliminating the false positives, we approached the task of recognition with different configurations and the one that got the best results has been reviewed in the 2012 Robot Vision task in ImageCLEF international campaign.

2 Image Analysis

In this section, we describe the image features that have been used in this work in order to obtain a precise and effective model for the topological localization task. In order to obtain an image representation which captures the essential appearance of the location and is robust to occlusions and changes in image brightness, we compare two different image descriptors and their associated dis-

tance measure. In the first case, we use color histograms integrated and in the second case each image is represented by a set of local scale-invariant features, quantized in bags of *visual words*.

2.1 Global Features

Many recognition systems based on images use global features that describe the entire image, an overall view of the image that is transformed in histograms of frequencies. Adopting the analysis of global features has brought great improvement in robot localization systems as in [33] or in content based image retrieval systems as in medical related images analysis in [34]. Such features are important because they produce very compact representations of images, where each image corresponds to a point in a high dimensional feature space.

In the following, we attempt to model image densities using two different color spaces, RGB and HSV.

RGB (Red, Green, and Blue) Color Model is composed of the primary colors Red, Green, and Blue. They are considered the additive primaries since the colors are added together to produce the desired color. White is produced when all three primary colors at the maximum light intensity (255). The RGB space has the major deficiency of not being perceptually uniform, this being the motivation of adding HSV color histograms.

HSV (Hue, Saturation, and Value) Color Model defines colors in terms of three constituent components: *hue*, *saturation* and *value* or *brightness*. The *hue* and *saturation* components are intimately related to the way human eye perceives color because they capture the whole spectrum of colors. The *value* represents intensity of a color, which is decoupled from the color information in the represented image. This color model is attractive because color image processing performed independently on the color channels does not introduce false colors (hues). However, it has also inconvenient due to the necessary non-linearity in forward and reverse transformations with RGB space.

A color histogram denotes the joint probabilities of the intensities of the three color channels and is computed by discretizing the colors within the image and counting the number of pixels of each color. Since the number of colors is finite, it is usually more convenient to transform the three channel histogram into a single variable histogram, therefore a quantization of the histograms is needed. The histogram dimension (the number of histogram bins) n is determined by the color representation scheme and quantization level. Most color spaces represent a color as a three-dimensional vector with real values (e.g. RGB, HSV). We quantize the color space of three axes into k bins for the first axis, l bins for the second axis and m bins for the third axis. The histogram can be represented as an n-dimensional vector where $n = k \cdot l \cdot m$. Because the retrieval performance is saturated when the number of bins is increased beyond some value, normalized color histogram difference can be a satisfactory measure of frame dissimilarity, even when colors are quantized into only 64 bins (4 Green $\times$ 4 Red $\times$ 4 Blue). As a conclusion, we chose a $18 \cdot 10 \cdot 10$ multidimensional HSV histogram, and a $10 \cdot 10 \cdot 10$ multidimensional RGB histogram, as differences between colors of

the office environment have a high level of similarity and have slight changes in hues.

2.2 Local Features

A different paradigm is to use local features, which are descriptors of local image neighborhoods computed at multiple interest points. There are many local features developed in the last years for image analysis, with the outstanding SIFT as the most popular. In the literature, there are several works studying the different features and their descriptors, for instance [22] evaluates the performance of local descriptors, and [44] shows a study on the performance of different feature for object recognition.

The three types of features used in our experiments are SIFT (Scale Invariant Feature Transform), ASIFT (Affine Scale Invariant Feature Transform) and RGB-SIFT (RGB Scale Invariant Feature Transform).These features were extracted using [14]. Also, the localization experiments using these features show advantages and disadvantages of using one or another.

SIFT (Scale Invariant Feature Transforms)[23, 4, 24] features correspond to highly distinguishable image locations which can be detected efficiently and have been shown to be stable across wide variations of viewpoint and scale. The algorithm basically extracts features that are invariant to rotation, scaling an partially invariant to changes in illumination an affine transformations. This feature has been explained in our previous work being one of our key level of our systems [1, 2].

ASIFT (Affine Scale Invariant Feature Transforms), as described in [29], simulates with enough accuracy all distortions caused by a variation of the camera optical axis direction. Then it applies the SIFT method. In other words, ASIFT simulates three parameters: the scale, the camera longitude angle and the latitude angle and normalizes the other three (translation and rotation), what SIFT lacked.

RGB-SIFT (RGB Scale Invariant Feature Transforms) descriptors are computed for every RGB channel independently. Therefore, each channel is normalized separately which brings another important aspect for SIFT, the invariance to light color change. For a color image, the SIFT descriptions independently from each RGB component and concatenated into a 384-dimensional local feature (RGB-SIFT) [5].

2.3 Feature Matching

In this subsection 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 presented are in fact histograms (color histograms, invariant feature histograms). As comparison of distributions is a well known problem, a lot of comparison measures have been proposed and compared before [31].

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 [17]):

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

- **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, [28]):

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

- **Bhattacharyya Distance [6]** (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:

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

3 Classification

Many of the features presented in Section 2 are in fact histograms (color histograms, invariant feature histograms, texture histograms, local feature histograms). As comparison of distributions is a well known problem, a lot of comparison measures have been proposed in Section 2.3. To analyze the different measure distances we summarize a well known choice for supervised classification.

Support Vector Machines are the state-of-the-art large margin classifiers which recently gained popularity within visual pattern and object recognition [15, 8, 18, 44, 40, 42]. Choosing the most appropriate kernel highly depends on the problem at hand - and fine tuning its parameters can easily become a tedious task. For our experimental setup, we chose the *linear kernel* (which is trivial and won't be presented), the *radial basis* function and the χ^2 kernel, presented below.

The Gaussian Kernel is an example of radial basis function kernel.

$$\mathrm{K_g(x, y)} = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{5}$$

The χ^2 Kernel comes from the χ^2 distribution.

$$\mathrm{K_{\chi^2}(x, y)} = 1 - \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\frac{1}{2}(x_i + y_i)} \tag{6}$$

3.1 Bag-of-Visual-Words (BoVW)

Recent advances in the image recognition field have shown that **bag-of-visual-words** [8, 12] - a strategy that draws inspiration from the text retrieval community - approaches are a good method for many image classification problems. **BoVWs** representations have recently become popular for content based image classification because of their simplicity and extremely good performance.

Basically, to give an estimation of the distribution we create histograms of the local features. The key idea of the *bag-of-visual-words* representation is to quantize each keypoint into one of the *visual words* that are often derived by clustering. Typically k-means clustering is used. The size of the vocabulary k is a user-supplied parameter. The *visual words* are the k cluster centers. The baseline of our tests are based on a bag-of-visual-words with a 100 *visual words*, meaning a 100-means clustering. The resulting k n-dimensional cluster centers c_j represent the *visual words*.

4 Experimental Setup

In this section, we explain the experimental setup, then we present and discuss the results. The different choices of distance measures and classification parameters are analyzed performing also a comparison with previous work results. Conclusions are drawn in benefit of an accurate solution for topological localization, data modeling and classification.

4.1 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 overviews and ImageCLEF publications [30, 33, 25]. The dataset 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	# Images		
	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

4.2 Eliminating False Positives

Finally, a method for the elimination of the unwanted results is performed, therefore the retrieved classes for images (*Corridor*, *LoungeArea* etc.) depend on a threshold, those below this value being rejected, with the meaning that the system doesn't recognize the image. 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.

We adapted the implementation of the genetic algorithm described in [11]. In order to capture the particularities of the distance measures that are correlated with the rooms on which they are used, we considered a different threshold for each room. As a justification for choosing multiple thresholds rather than a single one, let us consider the case in which we are trying to classify images taken from a room that is more distinguishable from the others. The values returned by the similarity measures when comparing these images to others taken from the same room are further apart from the values returned in the case of comparing these images with others taken from different rooms. In contrast, if we consider a room that is visually similar to others, these values will be closer on the real axis. This is why it is harder to correctly separate erroneous classifications for the good ones with a single threshold.

For the genetic algorithm, the chromosomes are vectors of length 9, representing the thresholds for the 9 rooms. For the genetic operators we used the binary representation of these vectors. The fitness function evaluates the quality of the thresholds and it is the measure used to score runs in the Robot Vision task. As a selection strategy, we used the rank selection, which sorts the chromosomes accordingly to their value given by the fitness function. In the crossover process, we don't allow the parent chromosomes which are the input for the crossover to be the same individual as it could lead to early convergence. To prevent this from happening, we first select one chromosome from the population and then run the selection process in a loop until a different chromosome is returned. We also used elitism in order to assure the survival of the best chromosomes of each

generation. In order to balance the diversity of the population, this method is accompanied by a slightly increased mutation probability.

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.

4.3 Results Interpretation

We are interested in observing the performances of the final configurations to see which features/dissimilarity measures lead to good results and which do not. As it is well known that combinations of different methods lead to good results [16], an objective is to combine the briefly presented features. However, it is not obvious how to combine the features. To analyze the characteristics of

Method	R[%]	P[%]	F
RGB-Only	73.73	82.02	0.77
HSV-Only	76.46	82.34	0.79
RGB-HSV	76.42	79.66	0.780
Basic-BoVW-SIFT	45.10	46.51	0.45
Basic-BoVW-SIFT+HSV+RGB	76.85	79.26	0.780
Basic-BoVW-ASIFT+HSV+RGB	77.60	79.97	0.787
SVM-RBF-BoVW-SIFT+HSV+RGB	78.87	78.87	0.788
SVM-LINEAR-BoVW-SIFT+HSV+RGB	78.63	78.94	0.787
SVM-χ^2-BoVW-SIFT+HSV+RGB	78.43	78.52	0.784

Table 2. Performance Comparison for Topological Localization

features and which features have similar properties, we perform an evaluation on selected configurations as shown in Table 2. The evaluation was performed choosing *Training 1* and *3* (Table 1) for training and *Training 2* for testing.

The first column gives a description of the used training method. The descriptions of the configurations are straight forward, for example, *Basic-BoVW-SIFT+HSV+RGB* means a configuration of a combination of RGB and HSV color histograms and *Basic-BoVW-SIFT* a bag of visual words formed with SIFT feature vectors. The chosen measure distances were decided like this: Jeffrey Divergence for RGB histograms, Bhattacharyya for HSV histograms and Minkowski for SIFT feature vectors. The second column gives the recall values for the training data, the third - the precisions. The F-measure is computed and represented in the fourth column of the table. The table also shows that feature selection only is not sufficient to increase the recognition rate but more flexibility is needed here and this fact led to different combinations.

The results are improved by the addition of the SVM classification step. We also add the observation that a SVM classification of SIFT mapped on visual words can get to a maximum of 52% accuracy, but these results are very assuring

in the context of a configuration in which are implied the usage of other feature descriptors. Thereby, the configuration that combines SIFT words, HSV and RGB histograms and a classification with a SVM with a RBF kernel yielded the most satisfying result.

4.4 ImageCLEF 2012 Robot Vision Task

The fourth edition of the Robot Vision challenge focused on the problem of multi-modal place classification. We had to classify functional areas on the basis of image sequences, captured by a perspective camera and a kinect mounted on a mobile robot within an office environment with nine rooms. We ranked **third** out of seven registered groups.

#	Group	Score
1	CIII UTN FRC, Universidad Tecnológica Nacional, Ciudad Universitaria, Cordoba, Argentina	2071.0
2	NUDT, Department of Automatic Control, College of Mechatronics and Automation, National University of Defense Technology, China	1817.0
3	**Faculty of Computer Science, Alexandru Ioan Cuza University (UAIC), Iaşi, România**	1348.0
4	USUroom409, Yekaterinburg, Russian Federation	1225.0
5	SKB Kontur Labs, Yekaterinburg, Russian Federation	1028.0
6	CBIRITU, Istanbul Technical University, Turkey	551.0
7	Intelligent Systems and Data Mining Group (SIMD), University of Castilla-La Mancha, Albacete, Spain	462.0
8	BuffaloVision, University at Buffalo, NY, United States	-70.0

Table 3. ImageCLEF 2012 Robot Vision final results

5 Discussion

Our approach on topological localization is currently applied on an office environment of nine sections: *Corridor, ProfessorOffice, StudentOffice, LoungeArea, PrinterRoom, Toilet, VisioConference, ElevatorArea* and *TechnicalRoom*. To address the problem of recognizing these sections separately, we approached the classification with specific thresholds in taking the final decision over the selected room. These thresholds create constraints that have to be loosened in order to obtain an accurate result in treating situations of great similarity between two different rooms. As an example, note that one of the main inconvenient that can appear in this case is that the rooms are very connected and difficult situations can rise as the robot moves around the office. For example, if the robot is in the *Corridor*, it looks to its right and sees the *LoungeArea* but its position is still in the *Corridor*. This type of situation creates noise that cannot be neglected,

therefore a proper threshold needs to treat these results that correspond to a humanized interaction with the medium. The threshold on the final decision quality was chosen to avoid erroneous localizations, thus favoring a result that doesn't specify any room and giving less correct localizations but also, less false assumptions.

6 Conclusions

In this work, we approached the task of topological localization without using a temporal continuity of the images and involving 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 an office environment dataset offered by ImageCLEF.

The main contribution of this work stays in quantifiable examinations of a wide variety of different configurations for a computer vision-based system and significant results. The experiments show that the configurations from different feature descriptors and distance measures depend on the proper combinations.

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. An important contribution to this field is given in this paper, along with notable experimental results, but there is still room for improvement and further research.

7 Acknowledgement

The research presented in this paper was funded by the Sector Operational Program for Human Resources Development through the project "Development of the innovation capacity and increasing of the research impact through post-doctoral programs"POSDRU/89/1.5/S/49944.

References

1. E. Boroş, G. Roşca, and A. Iftene. Uaic: Participation in imageclef 2009 robotvision task. *Proceedings of the CLEF 2009 Workshop.*
2. 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.
3. A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. *ICCV*, 2007.
4. 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.
5. G. J. Burghouts and J. M. Geusebroek. Performance evaluation of local color invariants. *CVIU, 13(113):4862*, 2009.

6. E. Choi and C. Lee. Feature extraction based on the bhattacharyya distance. *Pattern Recognition*, 36:1703–1709, 2003.
7. 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.
8. C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. *ECCV International Workshop on Statistical Learning in Computer Vision, Prague*, 2004.
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. 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.*
12. D. Gokalp and S. Aksoy. Scene classification using bag-of-regions representations. *Proceedings of CVPR*, pages 1–8, 2007.
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. J. Hare, S. Samangooei, and D. Dupplaw. Openimaj and imageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images. *ACM Multimedia 2011.*
15. Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 511–517, 2004.
16. J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classiffers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20 (3):226–239, 1998.
17. 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.
18. D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization. *BMVC*, 2006.
19. D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization and segmentation. *Journal of Image & Vision Computing*, 27(5):523–534, April 2009.
20. N. Lazic and P. Aarabi. Importance of feature locations in bag-of-words image classification. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1:I641–I644, 2007.
21. L. Ledwich and S. Williams. Reduced sift features for image retrieval and indoor localisation. *Australasian Conf. on Robotics and Automation*, 2004.
22. H. Lejsek, F.H. Ásmundsson, B. Thór-Jónsson, and L. Amsaleg. Scalability of local image descriptors: A comparative study. *ACM Int. Conf. on Multimedia*, 2006.
23. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2(60):91–110, 2004.
24. D. G. Lowe. Object recognition from local scale-invariant features. *Proceedings of the 7th International Conference on Computer Vision*, pages 1150–1157, 1999.
25. 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.

26. K. Mikolajczyk and C. Schmid. An afine invariant interest point detector. *Proceedings of the 7th European Conference on Computer Vision*, pages 128–142, 2002.
27. K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1), 2004.
28. K. Mikolajczyk, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition. *Visual Recognition Challange*, 2007.
29. 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.
30. A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt. Multi-modal semantic place classification. *Int. J. Robot. Res.*, 29(2-3):298320, February 2010.
31. 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*, page 11651173, 1999.
32. Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.
33. 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.
34. C. R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. Aisen, and L. Broderick. Local versus global features for content-based image retrieval. *Proc. IEEE Workshop of Content-Based Access of Image and Video Databases*, pages 30–34, June 1998.
35. J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. *Proceedings of the 9th International Conference on Computer Vision*, pages 1470–1478, 2003.
36. 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.
37. S. Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artificial Intelligence*, 99:21–71, February 1998.
38. I. Ulrich and I. Nourbakhsh. Appearance-based place recognition for topological localization. *IEEE Intl. Conf. on Robotics and Automation*, 2000.
39. V. Vapnik. Statistical learning theory. 1998.
40. C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. *Proc. ICCV*, pages 257–264, 2003.
41. 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.
42. L. Wolf and A. Shashua. Kernel principal angles for classification machines with applications to image sequence interpretation. *Proc. CVPR*, I:635–640, 2003.
43. J. Yang, Y. G. Jiang, A. G. Hauptmann, and C. W. Ngo. Evaluating bag-of-visual-words representations in scene classification. *ACM MIR*, 2007.
44. 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.