Using Scale Space Image Histograms for Global Localization of Mobile Robots

Hashem Tamimi and Andreas Zell
Computer Science Dept., University of Tübingen,
Sand 1, 72076 Tübingen, Germany,
{tamimi, zell}@informatik.uni-tuebingen.de

ABSTRACT

The scale invariant feature transform and the integral invariants are two well known approaches for visual feature extraction. Each of these approaches has been successfully applied to global localization of mobile robots. In this paper, we propose applying a combination of the two concepts. We demonstrate that extracting the integral invariants from the scale space does indeed improve the localization accuracy. We also show that the computation time of the proposed approach is much less than the scale invariant feature transform.

1 INTRODUCTION

The problem of robot localization can be classified as either global or local localization. In global localization, the robot tries to discover its position without previous knowledge about its location. In local localization, the robot updates its position using its current data from its sensors as well as the previous information that it has already accumulated. The lack of any historical information about its surroundings makes global localization more challenging [9].

Vision based robot localization demands image features with many properties. On the one hand the features should exhibit invariance to scale and rotation as well as robustness against noise and illumination. On the other hand they should be extracted very quickly so as not to hinder the other tasks that the robot plans to perform. In Wolf et al. [18] the problem of robot localization is dealt with by means of visual features that are also applied to image retrieval systems. The difference between the concept of image retrieval and robot global localization is only within their applications rather than in the methodologies used. Nevertheless, robot localization is a real time issue. It is also more elaborated when the visual surroundings have various similarities.

Visual features for robot localization can be generally classified into local or global features. This should not be mixed with the definition of global and local localization. Local features can be part of the images under investigation, and can also be referred to as local landmarks or local descriptors. These features should remain unchanged while the robot change its position, they should also maintain their values as well as position in the images under different illumination changes. Local features are commonly employed in robot localization because they are resistant to partial occlusion and are relatively insensitive to changes in viewpoint.

Some examples of robot localization using local features are: The work of Sim et al.[14], where Principal Component Analysis, PCA, is applied to local patches of the images. The paper of Lowe et al. [8], which is discussed thoroughly later on, applies local descriptors at different scales of the images and maps them into a set of histograms. Another example is our work in [16], where wavelet features are extracted around wavelet based salient points. In our paper [17], kernel PCA is applied to local patches in order to extract nonlinear features from the images. Although local features are robust to occlusions and are suitable for dynamic environments, they have some major problems when the features are not able to maintain their positions as the robot moves or when the robot is in an environment of highly changing illumination or noise.

In contrast with local features, global features are extracted from the whole images. In Wolf et al. [18], discussed in detail later on, a set of integral invariant features are extracted from the images, Monte Carlo integration is used to reduce the computation power, their features are finally represented in global histograms with both color and texture information. Another example of global features is the work of Jogan et al. [6], which is to apply robust PCA for robot localization with illumination invariance advantage. Although the extraction of global features can be more time consuming than local ones, the matching time of two global features is generally less than matching a set of local ones because matching local features is most of the time a correspondence problem.

This paper combines the work of two state of the art approaches for robot localization. The first one is the scale invariant feature transform approach [8], which is well known for robust localization but suffers from high computation

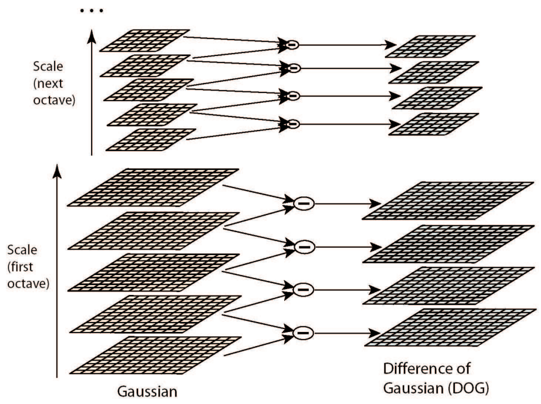


Fig. 1. A scale space of an image that shows 2 octaves and 5 scales in each octave (from [8]).

cost. The second one is robot localization based on image retrieval features [18], which we refer to here as integral invariants. These features are invariant to translation and rotation but when used in robot localization, they lack accuracy. The combination of the two approaches is done as follows: From the SIFT approach, we obtain a set of stable points to scale and illumination around which, features can be extracted. These features also hold information from course and fine structures found in the image. In out approach we extract the integral invariants around the stable points found by SIFT. We demonstrate that this can be accomplished with less computation time than SIFT.

The remaining sections of this paper are organized as follows: Section (2) reviews the SIFT approach. Section 3 reviews the integral invariants. Section (4) discusses the proposed approach. Section (5) defines the problem of global localization and discusses the similarity measure used. Section (6) presents the experimental work and finally section(7) concludes this paper.

2 SCALE INVARIANT FEATURE TRANSFORM

The Scale Invariant Feature Transform (SIFT), developed by Lowe [8], is invariant to image translation, scaling and rotation. SIFT features are also partially invariant to illumination changes and affine for 3D projection. These features have been widely used in the robot localization field. Se et al. [10] employ the SIFT scale and orientation constraints for matching stereo images. Andreasson et al. [1] propose a modified version of the SIFT approach to solve the global robot localization using panoramic images. Kosecka et al. [7] propose a method to further minimize the classification errors during localization by extracting SIFT features from each image and then using spatial relationships among the locations by means of a hidden Markov model. Silpa-Anan et al. [5] use an image map based on SIFT and Harris corners and use it later for robot localization.

The SIFT algorithm has 4 major stages:

1. **Scale-space extrema detection:** The first stage searches over scale space using a Difference of Gaussian (DoG) function to identify potential interest points.

2. **Keypoint localization:** The location and scale of each candidate point are determined and keypoints are selected based on measures of stability.

3. **Orientation assignment:** One or more orientations are assigned to each keypoint based on local image gradients.

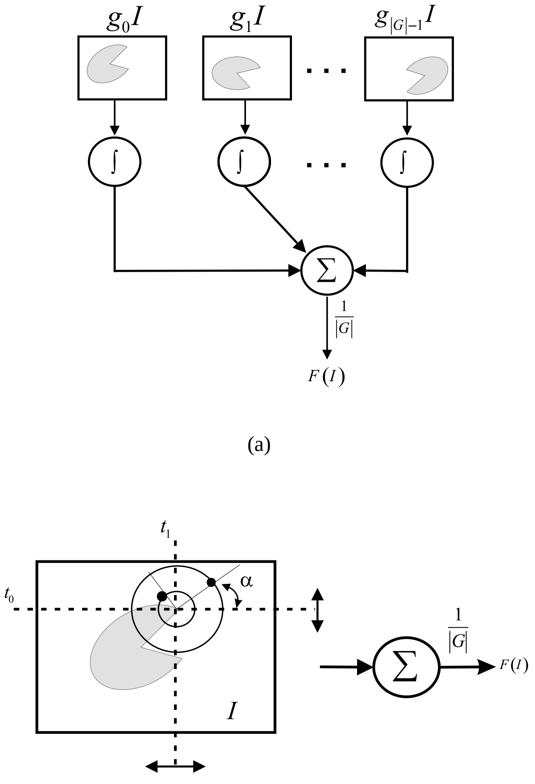
4. **Keypoint descriptor:** A descriptor is generated for each keypoint from local image gradients information at the scale found in stage 2.

The SIFT keypoints are found as scale-space extrema located in $D(x, y, \sigma)$, the Difference of Gaussians (DoG) function, which can be computed from the difference of two nearby scaled images separated by a multiplicative factor k:

$$\begin{aligned} D(x,y,\sigma) &= (G(x,y,k\sigma) - G(x,y,\sigma)) * \mathbf{I}(x,y) \\ &= L(x,y,k\sigma) - L(x,y,\sigma) \end{aligned} \tag{1}$$

where $L(x, y, \sigma)$ is the scale space of an image, built by convolving the image $\mathbf{I}(x, y)$ with the Gaussian kernel $G(x, y, \sigma)$, as seen in figure 1. Points in the DoG function which are local extrema in their own scale and one scale above and below are extracted as keypoints. Generation of extrema in this stage is dependent on the frequency of sampling in the scale space k and the initial smoothing σ_0. The keypoints are then filtered for more stable matches, and more accurately localized to scale and subpixel image location using methods described in [3]. For a more detailed discussion of the keypoint generation and factors involved see [8].

SIFT features are distinctive and invariant features used to robustly describe and match digital image content between different views of a scene. While invariant to scale and rotation, and robust to other image transforms, the SIFT feature description of an image is typically large and slow to compute. For example, the work in [2] presents a study of SIFT features for outdoor robot localization. Although their approach is able to pick up features that are stable despite the varying illumination, the authors reported some disadvantages of using SIFT, specifically that it takes a long time to extract the features from an image. Furthermore, the number of features is immense, which poses problems when searching for the matching pairs, along with having to store a large amount of data.



(a)

(b)

Fig. 2. illustration of Equation 2 using G elements. (a) The calculation of the features under different transformations. (b) The feature extraction around a point in the image (from Siggelkow)

3 INTEGRAL INVARIANTS

In this section we review another type of invariant features, which has been successfully used in image retrieval as well as robot localization areas. Unlike SIFT, these features are extracted globally from the image, which ease matching two images to a high extent [11]. The features in this approach are defined as follows:

Given an intensity image **I** of size $M \times N$, we can extract the integral invariants $IF(\mathbf{I})$ as given in equation 2

$$IF(\mathbf{I}) = \frac{1}{2\pi MN} \int_{t_0=0}^{M} \int_{t_1=0}^{N} \int_{\theta=0}^{2\pi} f\left(g\left(t_0, t_1, \theta\right)\mathbf{I}\right) d\theta dt_1 dt_0 \tag{2}$$

where $f(\mathbf{I})$ is a non-linear kernel function and g is an element in the transformation set $\mathcal{G}$, which consists here of rotation and translation. The application of one element g to the image **I** is denoted by $g\mathbf{I}$ as seen above. For the group of Euclidean motion there exists an angle $\varphi \in [0, 2\pi]$ and a translation vector $(t_0, t_1)^T \in I\!R^2$ such that:

$$(g\mathbf{I})(i,j) = \mathbf{I}(k,l) \tag{3}$$

where

$$\begin{pmatrix} k \\ l \end{pmatrix} = \begin{pmatrix} cos\varphi & sin\varphi \\ -sin\varphi & cos\varphi \end{pmatrix} \begin{pmatrix} i \\ j \end{pmatrix} - \begin{pmatrix} t_0 \\ t_1 \end{pmatrix} \tag{4}$$

Figure 2 illustrates how the features are calculated.

The choice of the non-linear kernel function f can vary. For example, invariant color features can be computed by applying the so-called monomial kernel, equation(5). On the other hand, invariant texture features can be contracted using the so-called relational kernel function as seen in equation (6)

$$f(\mathrm{I}) = \left(\prod_{p=0}^{P-1} \mathrm{I}\left(\mathrm{x_p y_p}\right)\right)^{\frac{1}{P}} \tag{5}$$

$$f(\mathrm{I}) = rel\left(\mathrm{I}\left(\mathrm{x_1}, \mathrm{y_1}\right) - \mathrm{I}\left(\mathrm{x_2}, \mathrm{y_2}\right)\right) \tag{6}$$

where

$$rel\left(\gamma\right) = \begin{cases} 1 & \text{if } \gamma < -\varepsilon \\ \frac{\varepsilon-\gamma}{2\varepsilon} & \text{if } -\varepsilon \leq \gamma \leq \varepsilon \\ 0 & \text{if } \varepsilon < \gamma \end{cases} \tag{7}$$

Applying the integral invariants to each pixel on the image is time consuming, as an alternative the work of Siggelkow et al. [13][11] is to estimate the invariant features using Monte Carlo integration method. Also in their work, it is discussed that integrating the outcomes can be destructive and they suggest using histograms instead. Another modification to these features was done by Halawani et al. [4], where the integral invariants are applied around a set of local patches in the image rather than the randomly generated ones. This led to more discriminate representation of the images. Wolf et al. [18][19][20] apply the integral invariants to robot localization using indoor images. Although the rotation invariance is not needed for an indoor mobile robot, the features have proved to be suitable because they are able to maintain the local structure held in the images [12].

4 COMBINING THE TWO APPROACHES

In this paper we aim to apply the integral invariants instead of the orientation histograms proposed in SIFT. This means we will calculate each feature of the integral invariants around the positions of the interesting points in each scale space, as seen in figure 3. By doing so we obtain even more distinctive features which eventually leads to more accuracy in localization. After calculating the features around each keypoint, a histogram is constructed for each scale of each octave. Our approach can be described in the following steps:

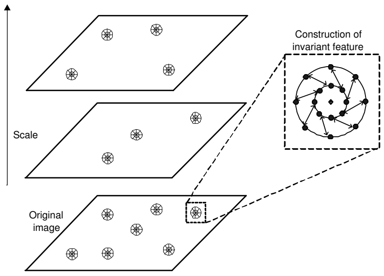


Fig. 3. The proposed approach: The construction of integral invariants from the scale space and the original image

1. **Scale-space extrema detection:** The first stage searches over scale space using a Difference of Gaussian function to identify potential interest points. Assume we end up with the scales $S_1..S_n$.

2. **Keypoint localization:** The location and scale of each candidate point are determined and keypoints are selected based on measures of stability. Let the number of keypoints in the scale S_i be K_i.

3. **Invariant features initial construction:** For each interest point c_j in the scale S_i where $j = 1..K_i$, we calculate a set of all points, which are at a distance r_1 from c_j, we denote this set as M^a. We use bilinear interpolation for sub-pixel calculation. Another set of points of distance r_2 from c_j are calculated in the same manner, we refer to this as M^b.

4. **Nonlinear kernel application:** A nonlinear kernel, explained in section (3), is employed on the values of the points in M_j^a and M_j^b. In the case of a monomial kernel, each point $p_k^a \in M^a$ is multiplied with another point $p_l^b \in M_b$. In the case of relational kernel, we employ equation (6) i.e: $rel\left(p_j^a - p_k^b\right)$. We end up with a new set of points for each interest point, let us refer to these set of points as M_j^c.

5. **Histogram establishment for each scale:** A histogram of b bins is constructed for all the sets M_j^c $j = 1..K_i$ in the scale S_i we refer to this histogram as h_i. We repeat this for all the scales $S_1..S_n$.

6. **Histogram establishment for the original image:** This is an optional step, where the positions of interest points in each scale are tracked down to the original image. Notice that the interest points in the octaves after the first one require scaling their positions. The number of positions at the end should be $\sum_{i=1}^{n} K_i$. We then apply the nonlinear kernel and construct a histogram h_0 from the original image.

5 GLOBAL ROBOT LOCALIZATION

5.1 Problem definition

The Localization is mainly performed in two phases: First, the robot performs an exploration phase, during which it discovers the environment for its first time, collects images from different positions and extracts features from these images. The features are stored in the robot memory along with the corresponding robot positions, usually in (x, y, θ) terms. Then the robot performs the localization phase, where robot retrieves its position by comparing the features from its current image with the features in the database.

5.2 Similarity measure

When comparing images through their corresponding features using SIFT, we apply the following similarity measure between each two images: For each keypoint in a given image k_a we find the two closest matching keypoints k_b and k_c from the other image. The matches are calculated through the squared distance measure in equation (8).

$$d_1\left(k_x, k_y\right) = \sum_{i=1}^{b} \left(k_{x_i} - k_{y_i}\right)^2 \qquad (8)$$

A positive match of the keypoint k_a with k_b is recognized if $4 * d_1\left(k_a, k_b\right) < d_1\left(k_a, k_c\right)$. The final decision, which image is similar to which, is then given by the one with the maximum number of positive matches. This usually leads to robust matching, nevertheless, the search for correspondences in this manner is time consuming. On the other hand, our approach involves extracting a single histogram for each scaled image, as well as an additional histogram for the base image itself. The difference d_2 between two images x and y is then defined in equation 9, where h_i^a is the histogram of the image a at scale i. The most similar image to the one at hand is then the one which has the minimum difference d_2 based on its features.

$$d_2\left(h_x, h_y\right) = \sum_{i=0}^{n} \left(d_1(h_i^x, h_i^y)\right) \qquad (9)$$

6 EXPERIMENTAL RESULTS

We use the same image set as in [15]. For localization we use a set of 121 gray scale indoor images each of which has 320×240 pixels. The images are taken in a 11×11 grid in a robot lab, 20 cm apart from each other. For exploration, another 30 images distributed in the robot lab are used. The camera in all the experiments is always heading in the same

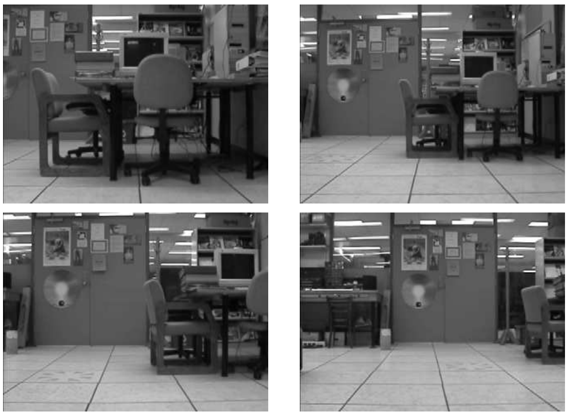


Fig. 4. Four sample images from the test set

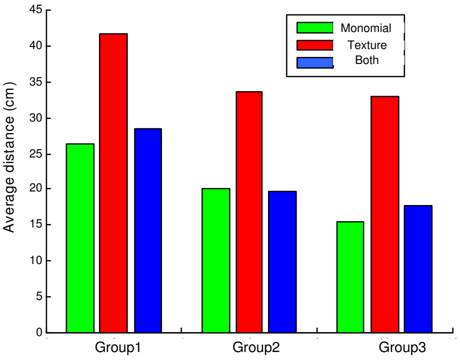


Fig. 5. Localization average distance using the three approaches.

direction, which means that we deal only with (x, y) coordinates and can neglect the orientation θ. We use these images specifically because of two reasons. First, their availability on the web enables many comparatives studied to take place. Second, they facilitate the accuracy measure, by means of their corresponding locations, because it is considered that the similar images are located closer to each other than different ones. This is not the case in many indoor images where similar images could be found in different locations and vice versa. Figure 4 includes some sample images.

Experimental work have shown that the following parameters lead to the best localization accuracy: For each image a pyramid with 3 octaves and 3 scales in each octave is built. Then, the interesting points are detected. The interpolation of two circles is calculated with radii $r_1 = 6$ and $r_2 = 9$ around each interesting point and the nonlinear kernels are applied. A histogram from the nonlinear kernel outcome is calculated $(h_1...h_9)$. The number of bins in each histogram is $b = 64$. An additional histogram is also calculated from each original image (h_0) with the same number of pins. The similarity is measured as explained in section (5.2). The experimental work is divided into two parts.

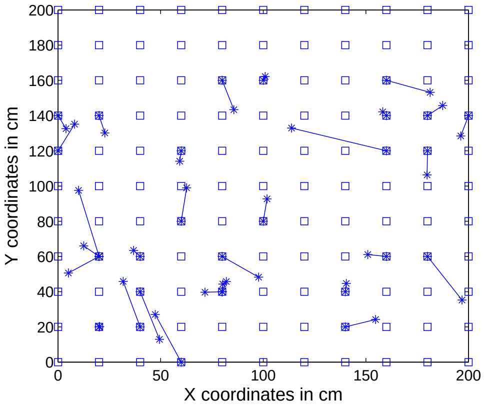


Fig. 6. Robot pose estimates and corresponding ground truth using the proposed approach. Average distance =15.27 cm.

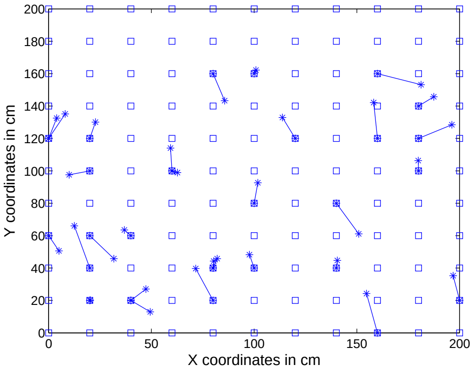


Fig. 7. Robot pose estimates and corresponding ground truth using SIFT. Average distance =12.81 cm.

The first part concentrates on comparing different parameters of the integral invariants with and without the construction of the scale space, whereas the second part compares the proposed work with the SIFT approach.

Experiment 1: In the first experiment we compare the result of using either a monomial function, a texture function or a combination of them. The results are illustrated in figure (5). In the figure, group1 is the average localization result when applying the integral invariants on the original image only, i.e. only deals with studying the histogram h_0 which is similar to [18][4]. Group2 shows the result of our proposed work using the histograms $h_1...h_9$. Group3 is done with the histograms $h_0..h_9$. The number of interesting points used in group1 and group2 are the same. Still, group2 has better results because those points are studied in the higher scales. In each of the three groups it can be seen that the monomial kernel alone leads to better results than either the texture kernel or even the results of using both kernels together. Initial experimental results showed that the best weights for combining the two kernels are 75% mono-

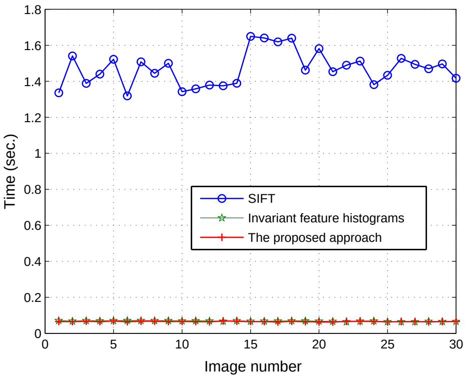


Fig. 8. The time required to extract features from a single image and compare them with those of 121 images using SIFT, the integral invariants and the proposed approach. The average time is 1.470, 0.068 and 0.065 seconds respectively

mial and 25% texture. From the results we can also conclude that group3 has the best results, specially when using the monomial kernel.

Experiment 2: We compare the results of group3 with the SIFT approach. Figure (6) is the robot map using the proposed approach with squares representing positions that belong to the exploration set of images and stars representing the position of the localization set. The lines between starts and squares represent the matching results between the two sets according to our proposed approach. The average distance between the matched images is 15.27 cm. Figure (7) shows another map with the matching that is done by the SIFT approach. The average distance is here 12.81 cm. which is a little more accurate than our proposed approach. Figure 8 shows the time required for the localization for each image using the proposed approach compared with SIFT. The time is variable because it depends on the number of interesting points which is different in each image. From the figure we can see that the proposed approach requires 0.068 seconds while SIFT needs 1.470 seconds. This is a major advantage to our approach. Figure (9) shows a third map done by the integral invariants, i.e. the first bar in group1 in figure (5).

7 CONCLUSION

Our proposed approach is based on extracting integral invariants from the scale space of the images. We have discussed this approach and compared it with two well known approaches. Our proposed work inherits the translation and rotation invariant properties form the integral invariants. Unlike it, the proposed approach has better localization accuracy, while maintaining nearly similar computation power. This is because our approach extracts more distinctive features. Although more distinctive features could be obtained from the original image, this information deals only with the fine and detailed structure of the image contents, whereas our approach deals with both course and fine details found around a set of stable points that are taken into consideration. When comparing our approach with SIFT, SIFT shows slight better accuracy but it requires much more computation power (up to a factor of 1:20). Finally, It is also important to say that, unlike SIFT, the proposed approach can be also extended to using colored images.

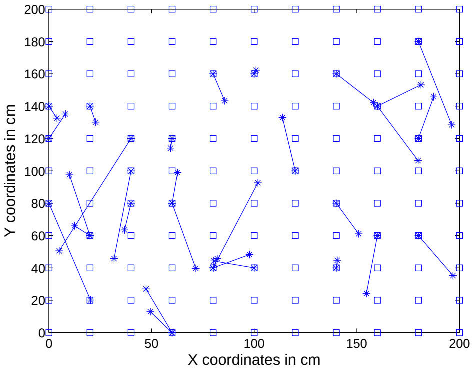


Fig. 9. Robot pose estimates and corresponding ground truth using integral invariants. Average distance =26.35 cm.

ACKNOWLEDGMENT

The authors would like to thank Prof. David Lowe for providing them with the source code of SIFT features. The first author would also like to acknowledge the financial support by the German Academic Exchange Service (DAAD) of his Ph.D scholarship at the University of Tübingen.

REFERENCES

[1] H. Andreasson and T. Duckett. Topological localization for mobile robots using omni-directional vision and local features. In *Proceedings IAV 2004, the 5th IFAC Symposium on Intelligent Autonomous Vehicles*, Lisbon, Portugal, 2004.

[2] Matej Artač and Aleš Leonardis. Outdoor mobile robot localisation using global and local features. In Danijel Skočaj, editor, *Computer vision - CVWW '04 : proceedings of the 9 th Computer Vision Winter Workshop*, pages 175–184, Piran, February 2004. Slovenian Pattern Recognition Society.

[3] M. Brown and D. Lowe. Invariant features from interest point groups. In *British Machine Vision Conference, BMVC 2002*, pages 656–665, Cardiff, Wales, September 2002.

[4] A. Halawani and H. Burkhardt. Image retrieval by local evaluation of nonlinear kernel functions around salient points. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, volume 2, pages 955–960, Cambridge, United Kingdom, August 2004.

[5] C. Silpa-Anan R. Hartley. Localisation using an image-map. In *Proceedings of the 2004 Australasian Conference on Robotics and Automation*, Canberra, Australia, 2004.

[6] M. Jogan, A. Leonardis, H. Wildenauer, and H.Bischof. Mobile robot localization under varying illumination. In *16th International Conference on Pattern Recognition*, volume II, pages 741–744, 5–11 August 2002.

[7] J. Kosecka and F. Li. Vision based topological markov localization. In *IEEE International Conference on Robotics and Automation (ICRA 2004)*, pages 1481–1486, New Orleans, USA, 2004.

[8] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[9] S. Se, D. Lowe, and J. Little. Local and global localization for mobile robots using visual landmarks. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 414–420, Maui, Hawaii, October 2001.

[10] S. Se, D. Lowe, and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) 2001*, pages 2051–2058, Seoul, Korea, May 2001.

[11] S. Siggelkow. *Feature Histograms for Content-Based Image Retrieval*. PhD thesis, Albert-Ludwigs-Universität Freiburg, Fakultät für Angewandte Wissenschaften, Germany, December 2002.

[12] S. Siggelkow and H. Burkhardt. Image retrieval based on local invariant features. In *IASTED International Conference on Signal and Image Processing (SIP) 1998*, pages 369–373, Las Vegas, USA, October 1998.

[13] S. Siggelkow and H. Burkhardt. Improvement of histogram-based image retrieval and classification. In *IAPR International Conference on Pattern Recognition (ICPR)*, volume 3, pages 367–370, Quebec City, Canada, September 2002.

[14] R. Sim and G. Dudek. Learning landmarks for robot localization. In *Proceedings of the National Conference on Artificial Intelligence SIGART/AAAI Doctoral Consortium*, pages 1110–1111, Austin, TX, July 2000. SIGART/AAAI, AAAI Press.

[15] R. Sim and G. Dudek. Learning generative models of invariant features. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, pages 3481–3488, Sendai, Japan, 2004.

[16] H. Tamimi and A. Zell. Vision-based global localization of a mobile robot using wavelet features. In Informatik aktuell, editor, *Autonome Mobile Systeme (AMS), 18. Fachgespräch, Karlsruhe, 4. - 5. December*, pages 32–41, Karlsruhe, Germany, 2003.

[17] H. Tamimi and A. Zell. Vision based localization of mobile robots using kernel approaches. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*, pages 1896–1901, 2004.

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

[19] J. Wolf, W. Burgard, and H. Burkhardt. Using an image retrieval system for vision-based mobile robot localization. In *Proc. of the International Conference on Image and Video Retrieval (CIVR), 2002*, 2002.

[20] J. Wolf, W. Burgard, and H. Burkhardt. Robust vision-based localization by combining an image retrieval system with monte carlo localization. *IEEE Transactions on Robotics*, 21(2):208–216, 2005.