2012 Second Brazilian Conference on Critical Embedded Systems

Autonomous Self-Localization of Mobile Robots through Reference Point Based Computer Vision

Leandro Nogueira Couto, Fernando Santos Osório

***Abstract*— This paper demonstrates a method for global localization of autonomous mobile robots based on the creation of visual memory maps, through detection and description of reference points from captured images, associated to odometer data in a specific environment. The proposed procedure, coupled with specific knowledge of the environment, allows for localization to be achieved through the pairing of these memorized features with the scene being observed in real time. Experiments are conducted to show the effectiveness of the proposed method for the localization of mobile robots in indoor environments. The results are analyzed and navigation alternatives and possible future refinements are discussed.**

***Index Terms*—Mobile Robots, Computer Vision, Feature Extraction, Robot Vision Systems**

I. Introduction

With the development of novel methods and the improvement of computer power, research on the field of mobile robots has grown over the years, aiming to create practical solutions by the use of autonomous mobile robots. The intersection between Mobile Robotics and Computer Vision is also a fertile area of research, for which there are several approaches. This paper is included in these particular areas of interest, and proposes a method aimed at enabling a robot to locate itself using a single monocular camera as the main sensor, through extraction of general visual features. The task for which the system was tested is an indoor robot patrolling system through a pre-defined route. We show that good localization accuracy can be achieved through the feature extraction method chosen, coupled with specific knowledge about the nature of the environment.

Global localization is one of the most important aspects to be considered in Mobile Robotics [1]. The challenges of global localization are frequently denoted by the kidnapped robot problem. [2], in which a robot is initialized or repositioned at an arbitrary point of the environment, in an arbitrary pose. This is not an uncommon situation in practical robotics tasks, and solutions to this problem relate to a robot's ability to recover from possible failures in maintaining its localization known. Thus, this kind of analysis is of broad importance.

The video camera, the main sensor used in this work, was chosen for its low cost (relative to other sensors), its ubiquitous nature and the richness of the visual information provided. The feature detection and description method used also allows for robust image signature generation and recall that is resistant or invariant to a number of image deformations. Vision is also the main sense used by humans to locate themselves, and this also allows for an immediate analogy. In fact, some works attempting to understand and model human visual recognition, identification and memory show that humans detect robust global image features and properties much like what many computer based feature extraction algorithms do [3].

Many varieties exist of feature extraction methods in Computer Vision, like those based on detection of corners [4], contours [5] or high level features [6]. The chosen method for this particular problem is SURF [7], which is itself inspired by and generally improves upon SIFT [8]. These methods are called scale-space based feature extractors. Features acquired this way will be matched between images from the memory map and the images the robot is currently observing. By knowing the odometry associated to the memorized images, the robot can estimate his current position.

The memory map described is based upon the VSRR (View Sequenced Route Representation), a visual representation proposed by Matsumoto [9], shown on Figure 1. In the VSRR, a recording run is executed by the robot while it is being manually operated by a human being. During this run, the guided robot captures images from the environment at fixed intervals, and the images are associated with the robots current odometric data (the robot's position and orientation). Later, when the robot is activated without being aware of its position or bearing, it is capable of pairing the most similar perspective (the criterion for deciding this may be choosing the image which yields the most matches, or the one whose features are spatially closer to the original in pixels, as will be discussed later) from the images in the memory map with the image seen, and so calculate its approximated localization relative to the memorized scene.

Manuscript received January 7, 2012. This work was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and INCT-SEC (Instituto Nacional de Ciências e Tecnologia em Sistemas Embarcados Críticos)

L. N. Couto. Author, is with the Instituto de Ciências Matemáticas e de Computação – ICMC Universidade de São Paulo - USP Cx. Postal 668 - CEP: 13560-970 - São Carlos – SP – Brazil. (e-mail: landovers@gmail.com).

F. S. Osório. Author, is with the Instituto de Ciências Matemáticas e de Computação – USP (e-mail: fosorio@icmc.usp.br).

978-0-7695-4728-2/12 $26.00 © 2012 IEEE
DOI 10.1109/CBSEC.2012.12

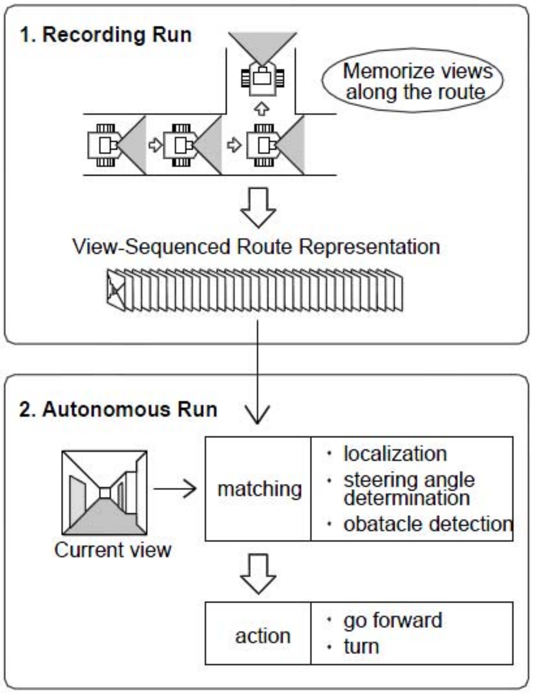

Fig. 1. The VSRR, taken as presented in [9]. It consists of the route representation generated by a recording run of the environment, which enables an autonomous run. This work focuses on localization in the autonomous run.

There is an intrinsic odometry error in this estimated measurement, but the localization is relative to the images, not the origin of the run. Because of this, what's important in this case is that the visual information is accurate, so that it can be used during navigation to mitigate the odometry error, allowing the robot to retrace the memorized course. The aspect of navigation will not be covered in this paper.

Localization itself occurs through matching of SURF signatures (a signature is the collective of features obtained from a given frame). SURF's recognition algorithm specializes in matching features for the same scene elements even when subject to a great variety of image deformities and changes, such as changes in scale, rotation, lighting, perspective, affine deformation, partial occlusion and noise. Even though it is a very robust method, preliminary tests have shown that images from the same scene, although they are recognized from a variety of perspectives, when observed from similar points of view, yield more positive SURF feature matches than when the same scene is observed through slightly more different points of view. This is an important fact to this work, because when navigating indoor areas like, for example, a straight corridor, a robot is bound to observe the same feature for a while, from many different perspectives. The scene will be recognized from most perspectives, but the perspective most faithful to the original will offer more matches. So the best candidate for position is normally the feature with the most matches. In this work we also propose an improvement to this criterion, based on comparisons between signature positions within the image. This shows that context information can be used in tandem with general feature matching methods to allow for better performance through specialization. In the case of ties, the immediate neighbors of the candidate images in the memory map are analyzed. This is done because the best candidate image is more likely to have the best neighbors as well.

Section II of this paper will describe the proposed approach in greater detail, and section III will describe the experiments and tests run and show the acquired results. Section IV discusses the proposed project and draws conclusion based on the work done.

II. Description of the Proposed Method

We now present the hardware and software resources used in this work, as well as the methodology applied. For the development of the proposed work, a Pioneer 3-AT robot was used. The developed systems makes heavy use of the OpenCV Computer Vision library (2.3 version), the Player robot software platform (3.0.2 version) and the OpenSURF (27/05/2010 build) implementation of the SURF feature detector and descriptor method.

A. The Test Platform

The chosen robot for the system testing is the Pioneer 3-AT, developed and produced by MobileRobots Inc. It is versatile and robust, the Pioneer series being one among the most popular mobile robots used in research. A computer can be embedded in the robot, communication between the two occurring through a serial bus. This robot also has extensive software support, with Player drivers implemented for many of the robot's sensors and actuators.

B. The Feature Detector and Descriptor

So that robust and distinct reference points can be extracted from an image, a scale-space based detector-descriptor method was used. The choice of method was not arbitrary, but the choice was based on the Best alternative for the task. Image correlation methods are commonly used for this kind of task, but are not as resistant to changes. Corner detectors such as FAST [10] are not reliable under changes in scale, and scale variance is an essential aspect of indoor environments navigation tasks. The SURF (Speeded Up Robust Features) detector-descriptor method was considered the most adequate option. A good detector should be repeatable (yield the same results from different images despite possible changes), while the descriptor must offer distinctive interest point descriptions that can be later univocally associated with other features.

When compared to other similar feature detection and description methods that use scale-space, SURF offers good performance regarding matching results. Also importantly, an aspect at which SURF frequently outperforms other methods of its kind is in the speed of processing in terms of both feature extraction and features matched per second [11]. As the present work proposes to evaluate relatively large amounts of image data, and aims to be usable in practical monitoring and patrolling tasks, this is a determinant factor, since most of the feature processing will be done during execution time.

Other comparative works [12] also demonstrate that SURF has similar positive matching rates compared to methods such as SIFT, with a bigger tendency to be affected by variations in rotation (roll). This is not a very relevant issue for the proposed applications and chosen environment, because navigation in indoor areas occurs mostly through plain terrain. This in turn, makes for rotational perspective variance on the normal axis relatively to the robot's camera to be minimal, so it can be outright neglected. In fact, a modification has been introduced to the SURF method to reflect this. The step in feature detection that guarantees rotation invariance is bypassed, and this improves the method's performance. This variant is known as Upright-SURF, or U-SURF [7]. It provides an increase in speed, and for this particular case, false positives can be avoided.

Methods like SURF and SIFT are relatively new, and an active field of research. Therefore, more recent alternatives in detector-descriptor methods exist, like CenSurE [13], but further comparisons are needed before they can be proven superior to SURF, while SURF remains a more established option.

C. The Recording Run

Having established the method of feature detection, the robot can perform the recording run through the determined path. The robot was guided with a joystick during the recording. While traversing the path, the robot took pictures of the environment at a rate of one image per second. The density of the memory map generated by this frequency of capture was deemed enough for the results required, and the average distance between two adjacent images in the memorized image database is 5 cm. Simultaneously to the image capture, odometric information was recorded onto a log file, specifically the robot's position (*x*,*y*) and rotation (*z*) parameters, taken in reference to the coordinate where the robot started executing (where *x*, *y* and *z* are zero). The association between the ordered images and the odometry where each of them was taken yields a topological map of the environment. Figure 2 shows examples of images taken during the recording run and used in the tests.

Following this recording, a program able to extract features from the image database through SURF was created. The SURF signatures of every image from the database was obtained and saved, so that they would not need to be recalculated during run-time. The SURF signature used with the descriptor was the 64 bits variety; the 128 bits variety demands further processing and does not provide significant improvement in description results. The resolution for all images used is 640x480. Resolution plays an important role in scale-space methods since they operate by downscaling images to search for features present in all scales (which are resistant to changes in scale).

D. Global Localization

Experiments for this work focused on the global localization algorithm. To test the proposed technique, the robot was initialized in a random point within the recorded path, without further information regarding its localization. The frame collection constituting the memory map and base of reference involved a 20 meters path through a corridor and a small office room. Two different criteria were used to determine which of the memorized images corresponded more accurately to the currently observed image. The first criterion tested to select the best image was to choose the image which yielded the most SURF matches. The reasoning behind this is that closer perspectives generate better matching and so more pairs of features. The second criterion was to measure the average distance in pixels between the paired features from each image and choose the signatures that are spatially closer. The reasoning for this method was that signatures whose feature positions are on average similar when taken relative to their respective images are more likely to have originated from close perspectives. To evaluate the spatial closeness between images, only the Y axis was taken into account. This is because most of the movement during the recording run occurred in the direction normal to the camera, which affects positioning of features in the Y axis the most. In a practical navigation algorithm, the same technique can be applied to the X axis to avoid angular deviation. The experiments were executed considering both criteria. In both cases, the rule adopted to settle ties was to consider the results obtained by the candidates' neighbors, with the preferred image having the best neighbors.

III. RESULTS

The accuracy of global localization yielded positive results, and when the robot was put in a position inside the memory map it's position could be estimated with good precision. We expect the precision in these results is enough in most practical cases for self-localization in robots performing most indoor tasks such as security patrolling. Even in indoor environments with few evident characteristics which make it difficult for the SURF method to acquire distinct features, results were accurate with few false positives. We expect that coupling these results with simple methods of determining movement through observation of optical flow [14] or visual odometry [15] techniques would result in great improvements, but even in its current uninformed state the localization can be used in practice. Both criteria for determining the best image pair (the regular SURF counting of the number of matches and the method proposed here of spatial proximity between features in different images) can be used for localization, and the coupling of these two criteria is also expected to improve results. In the criterion where the raw number of matches is counted, sometimes scenes with very few features would yield matches with many features across many images, a possible problem with the description of that particular feature. In these cases, we determined that a single feature with many matches in different places should not be counted. Other techniques such as navigation by odometry or an extra lateral camera would be beneficial in practice in this kind of situation where the camera captures particularly plain images.

The computer used to process the data was an Intel Core 2 Duo de 2.20GHz with 2GB of RAM. In this configuration, the association between images, including loading the images onto the memory required in average 415 ms to resolve.

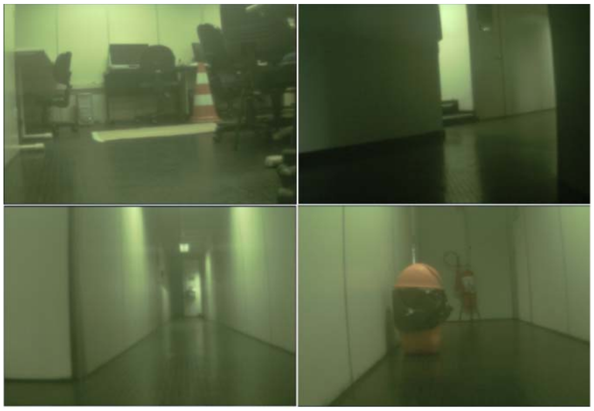

Fig. 2. Captured frames in different points of the paths used on the tests, showing the different types and profiles of scenery the robot found in the runs.

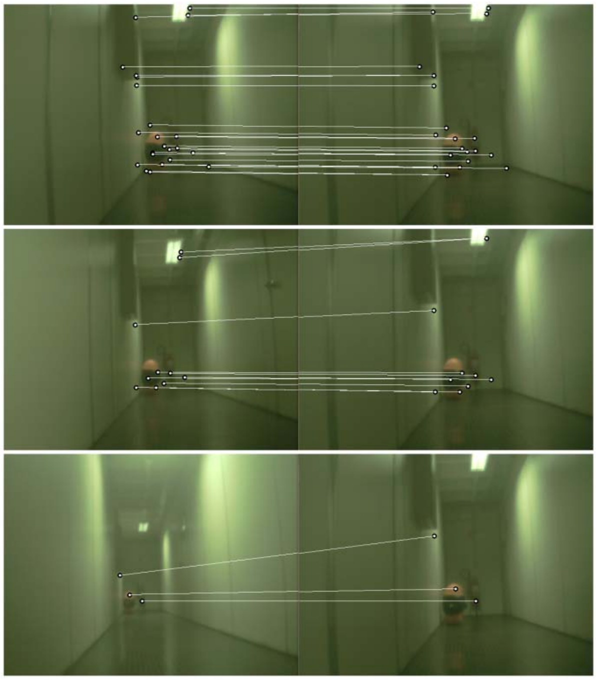

Fig. 3. This montage shows 3 results of the SURF matching algorithm, with paired images being presented side by side, and the horizontal lines across the images showing the matching pairs of features.

In the experiments, the matching was tested on two sets of image data. Each data set was composed of around 250 images each, covering a distance of 20 meters. The main attribute that was measured was the estimated frontal distance (the movement of the robot relative to the axis normal to the plane of the camera). The measured error pertains to this distance, and angular deviations were ignored for the purpose of the experiment. The localization error on the run using the first criterion to determine the best image was of 27.82 cm, with a standard deviation of 47.43 cm. The localization error using the second criterion was of 16.26 cm, and the respective standard deviation was 25.13 cm.

As expected, images taken from the same scenes but in different perspectives, for instance in a corridor, present a gradual change in the number of matches. This effect is noticeable on Figure 3. Closer perspectives generally yield more matches. Also, closer perspectives yield matches that have a smaller offset in the Y axis (this can be perceived in the images since the lines are more horizontal on the closer perspectives).

Figure 4 shows an example of matches for 3 different observed images. Each line corresponds to the matching results for each of the 3 test images being compared to all other images taken in the recording run and stored in the database. For each case, it can be seen that the number of matches is greater in the most accurate image, and also in its neighborhood.

SURF is a method developed to detect the same features even from different perspectives. Taking the limitations of the method and the context of the intended application into account, we could devise methods for electing the image closest to an observed image, even from a database of highly similar images with relatively few features.

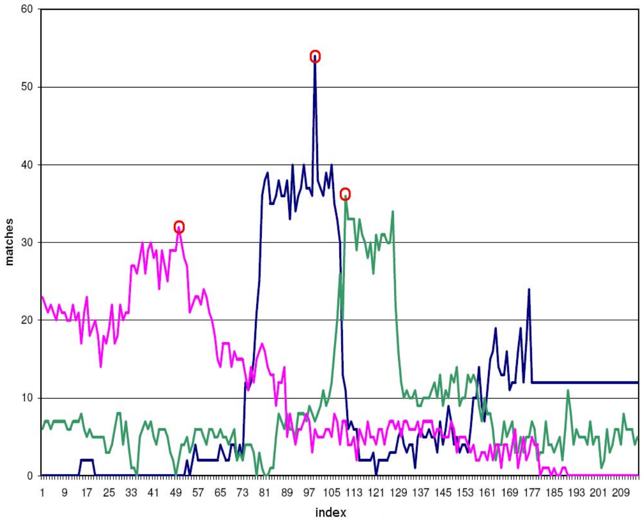

Fig. 4. This graph shows the number of SURF matches (vertical axis) when applying the method comparing 3 different images to every image representing a route through a corridor, shown here sequentially by consecutive image index (horizontal axis). Each line represents the matches of all images in the route relatively to a different observed image. The best match in each case is marked with the circular dot.

IV. Conclusion

This work shows how global localization can be acquired by a simple method using a popular method for detection and description of features. Usually localization is done through the use of range sensors. Omni directional cameras and stereo camera setups can synergize well with scale-space detectors [16], but their cost makes them prohibitive or difficult to use in everyday appliances or commercially viable robotic systems. They also provide an overhead in processing costs. A monocular camera does not directly provide depth information, but this paper shows that explicit knowledge of depth is not always necessary for localization, although depth information is implicitly estimated through the best image choice criteria.

The approach using the straightforward criteria of considering the most matches generated good results. Considering context information about the environment (like the fact that the average position of features on the Y axis changes with perspective changes) allowed us to come up with a new criterion to deciding which the best image is. The new criterion resulted in better results in the sets of data used in the experiments, both in a lower average error and lower standard deviation of the error. Both sets presented outliers, matches that were nowhere near the target (less so with the second criteria). Clear outliers and bigger errors can be fixed during navigation by predicting probabilistic methods and multiple iterations of the localization algorithm, to increase certainty about the estimated information. The tests were performed in indoor environments with good results, and results show the possibility of the system's practical use in cases where the precision obtained is enough for the desired task. In future works, both criteria used can be merged, which is expected to improve results. From the results, we estimate the match count criterion can be used for coarse, more general localization, since it is more effective in different scenes, while the Y axis offset criterion can be used to refine localization results, since it performed better than the match count in fine adjustments, when images were very close, but not as well when the scenes are more distinct.

Another positive aspect of the method demonstrated in this work is that there is no requirement that the robot be in motion for the localization to work. While optical flow, visual odometry and Structure from Motion [17] methods require the robot to move, the proposed method can achieve localization from a still pose, using one single frame.

In practical circumstances requiring further precision and maintenance of localization (local localization), probabilistic methods like Bayesian filters [1] can be applied during navigation to increase certainty regarding the robot's knowledge of its own position.

Among medium and long term objectives for future works we intend to develop Monocular SLAM [18] based upon an extension of the methods described here to recognize and determine relative camera position. This would allow the robot not to be restricted to a pre-defined route specified in the recording run, but being able to map the environment and navigate simultaneously.

REFERENCES

[1] S. Thrun, Wolfram Burgard, Dieter Fox. *Probabilistic robotics*. Cambridge, Mass. : MIT Press, 2006.

[2] C. Howie, *Principles of Robot Motion: theory, algorithms and implementation*. The MIT Press.2005, p. 302.

[3] Oliva, A. & Torralba, A. Building the Gist of a Scene: The Role of Global Image Features in Recognition. *Progress in Brain Research: Visual perception*, 155, 2006. 23-36

[4] Harris C. and Stephens M.J.: A combined corner and edge detector. *Plessey Research Roke Manor, United Kingdom. The Plessey Company plc*. 1988.

[5] Canny J. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, 1986. 679-698.

[6] Duda R.O., Hart P.E.: Use of Hough transform to detect lines and curves in picture. *Communications of the ACM*, 15(1): pp. 11-15. 1972.

[7] H. Bay et al. SURF: Speeded Up Robust Features. *In ECCV*, 2006. Pages 404 – 417.

[8] Lowe D.G.: Distinctive Image Features from Scale-Invariant Keypoints. *Computer Science Department of University of British Columbia.* Vancouver, B.C., Canada. 2004.

[9] Y. Matsumoto, M. Inaba, H. Inoue. Visual navigation using view-sequenced route representation. In *Proceedings of the IEEE International Conference on Robotics and Automation* (ICRA), 1996. Pp. 83 – 88.

[10] E. Rosten, T. Drummond, "Machine learning for high-speed corner detection (Published Conference Proceedings style)", in *Proc European Conference on Computer Vision*, 200, pp. 430 – 443.

[11] J. Bauer, N. Sunderhauf, P. Protzel, "Comparing several implementations of two recently published feature detectors (Published Conference Proceedings style)" iIn *Proc of the International Conference on Intelligent and Autonomous Systems*, IAV, 2007.

[12] L. Juan, O. A. Gwun, Comparison of SIFT, PCA-SIFT and SURF. *International Journal of Image Processing (IJIP)*, vol. 3, Issue 4, 2010. pp. 143 – 152

[13] M. Agrawal, K. Konolige, M. R. Blas, CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching. In *ECCV*, 2008, Part IV. pp. 102 – 115.

[14] B. Horn, B. Schunck. Determining optical flow. *Artificial Intelligence*, vol. 16, 1981, pp.185–203.

[15] D. Nister, O. Naroditsky, J. Bergen, "Visual Odometry (Published Conference Proceedings style)" in *Proc of the 2004 IEEE Computer Society Conference on Computer Vision and pattern Recognition*, 2004. pp. 652 - 659.

[16] Lowe D., Su S., Little J.: Vision-based Mobile Robot Localization And Mapping using Scale-Invariant Features. *Department of Computer Science University of British Columbia Vancouver*, B.C. V6T 1Z4, Canada. 2001.

[17] Kaess M. and Dellaert F.: Probabilistic Structure Matching for Visual SLAM with a Multi-Camera Rig. *Computer Vision and Image Understanding*. 2009.

[18] A. Davison, I. Reid, N. Molton, O. Stasse. " MonoSLAM: Real-Time single camera SLAM. (Published Conference Proceedings style)" in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007. pp 1052 – 1067.