Two-Level Topological Mapping and Localization Based on SIFT and The Wavelet Transform

Sara Elgayar
Computer Science
sara.elgayar@gmail.com

Mohammed A.-Megeed Salem
Scientific Computing
salem@cis.asu.edu.eg

Mohamed I. Roushdy
Computer Science
miroushdy@hotmail.com

Faculty of Computer and Infromation Sciences, Ain Shams University
Abbassia 11566, Cairo, Egypt

Abstract: This paper presents a novel framework for Vision-Based Simultaneous Localization and Mapping which focuses on the class of indoor mobile robots using only a monocular camera to achieve a two-level topological map. Local and global features are combined in the same topological framework. The Scale Invariant Feature Transform is used to extract and build up a global map which provide rough estimation of the robot position. The map is then decomposed into a sub-maps, in which the horizontal, vertical and diagonal details of the wavelet coefficients are used to provide finer estimation of the robot position and pose. The output topological map is validated with the ground truth of the environment. The results show high localization accuracy and low rate of matching time.

Key–Words: V-SLAM, Monocular SLAM, Local Features, Global Features, Computer Vision

1 Introduction

Simultaneous localization and mapping (SLAM) or Concurrent Mapping and Localization (CML) as referred in [1] is one of the most extensively researched field of robotics. To build a truly autonomous robots, it must have the ability to autonomously construct maps. SLAM is the problem of building a map while at the same time localizing the robot within the map. Undoubtedly SLAM is much more complicated than Localization or mapping processes, as, mapping addresses the problem of generating a map using the acquired information by robot's sensors and the given robot's poses. On the other hand, localization addresses the problem of determining the robot's locations within a given map [1].

Sensory input is one of the main issues that must be addressed when working with SLAM. The most common sensors researchers used to exploit are laser & sonar. Nevertheless, recently vision sensors gained more attention for performing SLAM [2]. Imaging sensors offer a variety of desirable properties. Cameras are low cost, provide a huge amount of information, available and passive. Vision also allows the development of a wide set of essential functionalities in robotics such as, obstacle detection, people tracking, visual servoing, and others.

Localization systems based on landmarks rely on either artificial or natural landmarks to represent the environment in the form of points and lines, or more complex patterns and try to determine correspondences between the observed landmarks and a pre-loaded map and to estimate the location of the robot from this correspondence [3]. Discriminative classification is proposed in [4] using a SVM and histogram-based features with the kernel averaging method. The output of the classifier for each frame is a label and its associated margin, which it took as a measure of the confidence of the decision. A multi-level machine learning proposed in [5] is made of a first "weak" classifiers level based on visual features and of a second level performing fusion of first level outputs. The authors in [6] used global visual features for image similarity and a geometric verification step using vanishing points.

Mapping addresses the problem of projecting the information gathered by the sensors into a consistent model of the environment. Different methods used for representing the map of the environment such as, metric, feature-based map or as topological graph [7]. The most popular metric representation is grid-based maps. They represent the environment being mapped in form of cells. Each cell is marked either as free or occupied. It gives detailed and precise model of

ISBN: 978-960-474-308-7

the environment. However, it is not to scale well with the actual environment dimension [8]. The idea of feature-based maps is to extract features from the environment. Detected features are then registered in the map. They scale well with the environment's dimension [9]. Topological maps represent environments as a list of significant places (nodes) that are connected via arcs. Thus they are mainly graphs and are especially useful for path planning. They scale well to large environments, since the amount of information that is stored is limited to the description of the places. The problem is, a same place may be represented more than once. To overcome such problem, many authors proposed to use the topological map in a combination with metric or feature-based maps [10][8].

Topological V-SLAM is a SLAM process which is based on vision for environment sensing and used topological map for representing the environment. In this paper a V-SLAM algorithm is proposed with the use of a single freely moving camera as the only data source. Several research challenges have been considered: (1) How to reduce the number of images needed to describe the environment without losing important details of the test environment? (2) How to reduce number of features and how to track them through images? (3) How to calculate in an efficient manner the similarity of the input image against all the reference images in the map? (4) How to represent the environment by two-level topological map?

The paper is organized as follows: The tools used for feature extraction are introduced in section 2. The architecture of the proposed system is given in section 3. Section 4 describes the data set used and the experimental results followed by the conclusion in section 5.

2 Feature Extraction and Classification

Features extraction is the process of finding some sort of description that can be used later to identify the region or the object of interest and to differ this object from other examples. It is agreed that local features are more robust to scene dynamics and illumination adjustments than global features. This ability makes local features more applicable for wide-range characterization of the environment. To illustrate, it can better recognize a room from one another than to recognize a different location in the same room. By contrast, global features usually have weak resistance to illumination and dynamic changes, which makes the global features suitable for narrow-range characterization of the environment. For example, images captured at adjacent locations own similar signatures even if there are illumination or dynamic changes [11].

2.1 Scale Invariant Feature Transform

Among the available techniques for feature detection and extraction, SIFT has proven to have the ability to find and match features with higher degree of uniqueness and robustness. It has been successfully applied to robot localization and robot SLAM [10][12][13]. SIFT was developed and published by David Lowe in 1999-2004. It aims at representing an image by a set of local interest points which are invariant to image transformation and partially invariant to illumination changes. SIFT algorithm consists of two main stages which are (1) detection of key points and (2) description of the detected key points. The first step is to construct the known difference-of-Gaussian (DOG) images. The second step is to localize the key points by comparing each sample point in DOG images to its neighbours in the current image and in the scales above and below. Key points are described by the orientation histogram for each subregion around the key points.

2.2 The Discrete Wavelet Transform

Wavelet transform has been successfully used for vision based robot localization, vision-based SLAM and image retrieval algorithms [11] for their capability in representing images in a compact way without losing information about location of the image discontinuity, shapes and textures. Mallat [14] has proposed an iterative algorithm to compute the discrete wavelet transform. It is based on the multiresolution analysis. The algorithm is based on computing iteratively an approximation at a lower resolution level j of the original signal $f(t)$. For this an orthogonal set of basis functions $\phi_{k,j}(t), k, j \in \mathbb{Z}$ is used, called the scaling functions. The differences of the information between two approximations at successive resolution levels (the details) are extracted by the orthogonal set of the wavelet functions $\psi_{k,j}(t), k, j \in \mathbb{Z}$. The Haar wavelet ψ_{haar}, is the basis of the simplest wavelet transform. It is orthogonal and have finite vanishing moments, i.e. compact support [15], which ensures local anal-

ISBN: 978-960-474-308-7

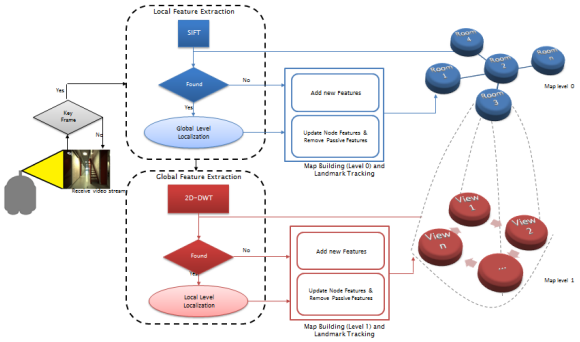


Figure 1: System Architecture

ysis. The scale function ϕ_{haar}, is a simple average function. The 2D wavelet transform is widely used for analysis and processing of images and videos. The results of the analysis at each decomposition level are a low-pass image or a coarser approximation A and three detail images, horizontal details H, vertical details V, and diagonal details D, which contain the details lost while going from the original image to its approximation A. The approximation A represents the image at a coarser resolution. Horizontal edges tend to show up in H and vertical edges in V, while D contains all other details [16].

3 The Proposed System

This paper proposes a novel system for V-SLAM. The input to the system is a sequence of key frames, and the output is an evolved two-level topological map and relationship between nodes in the map, as well as two-levels topological localization for the robot's current position in the environment. An illustration of the system is shown in Figure 1.

3.1 Sensing

When vision-based SLAM uses only a single camera, it is called *Monocular SLAM*, *Mono-SLAM* or *bearing-only SLAM*. One of the main advantages of the single camera setup is its low cost. On the other hand, single cameras don't provide any information about the feature depth. The image sequences of the dataset used in this research were acquired using the MobileRobots PowerBot robot platform equipped with a stereo camera system consisting of two Prosilica GC1380C cameras [17]. However, a monocular vision system is used in this paper.

3.2 Preprocessing

The experiments show that the preprocessing affects strongly the performance and accuracy of the system. Therefore we have used different steps of preprocessing. The captured image is converted to grey scale, and then the *"Next Increase"* [18] procedure is applied to decide whether the captured image is a key frame or not.

3.3 Feature Extraction

Features can be classified as global or local features. Examples of global features are the mean color of the object, image histogram, or the wavelet signature. Strong edges and corners are common examples of local features [19]. We think that local features provide rough level of estimation for the robot's location, while global features provide detailed estimation for the robot's position such as its pose.

In this paper, two-level map is produced and the robot is two-levels topologically localized for example (a) in which room the robot is, (b) in which corner the robot is. Global image signatures and local features are combined in the same framework as shown in Figure 1. SIFT local features are used for the high level estimation of the robot position and map building. Whereas, for the low level estimation of the robot position and map building discrete wavelet signature of images grabbed are chosen due to its simplicity, robustness, scalability and small memory requirements.

3.4 Feature Tracking

First, for implementing the global level: The main idea is to merge all extracted SIFT interest points from multiple frames that belong to the same location in a buffer, so that each reference location is described by a group of SIFT interest points.

When SIFT features from the first frame arrive, a new map node is created. For each extracted SIFT interest point p_i, the features saved are: The 2D position relative to the initial coordinates frame, the scale and orientation of the landmark, the set of descriptors (n x 128), and a count C to indicate how many consecutive frames this landmark has been missed. Initially this count is set to 0.

Over subsequent frames, new entries are added to each node, features are tracked and entries are removed from nodes when appropriate so that a minimum number of features robustly describe each reference location. There are the following types of features to consider: (1) New

ISBN: 978-960-474-308-7

features arrive from a key frame for a previously visited location, so they are added to the node and the missed count for each feature is initialized to 0. (2) This feature was matched before in a previously visited location, so, the missed count remains 0, and it is said to be an active feature. (3) If the missed count C of any feature in the map reaches a predefined limit N (7 was used in experiments), this feature tracking is terminated, it's said to be a passive feature and is removed from the map.

Likewise, for implementing the local level: if a key frame is accepted in the global node, then, the 4^{th} level of the 2D Haar discrete wavelet transform is calculated, and a signature consists of the horizontal, vertical and diagonal details is saved in this node.

3.5 Topological Mapping

A topological representation of the environment is used in our algorithm. Each node in the global topological map is a rich node that contains information about a reference location and the number of SIFT interest points. Similarly, each node in the local topological map contains information about the view and the wavelet signatures of the matched key frames.

The complete process for two-level topological mapping can be summarized in the following steps: First, SIFT points are extracted from the captured key frame as explained before. Second, similarity is computed between the current key frame and all nodes of the global map by means of the number of matched SIFT points. For instance, a captured key frame can represent a wide scene that was captured over multiple sequence of images, accordingly SIFT features extracted from one key frame can match SIFT features from multiple key frames representing same scene, as show in Figure 2. Finally, a ranking with the best n similarity values and its associated locations is obtained. If the similarity value of the highest ranking global node exceeded a predefined threshold (25 SIFT points) then, the test frame is assigned to this global node, otherwise a new node is added to the global map, and a connection between the new node and the last visited node is created.

Likewise, for the local map, the wavelet signature of the test frame is computed, and compared to all wavelet signatures of the matched global node. A ranking with the best n similarity values and its associated views is obtained. If the similarity value of the nearest local node exceeded a predefined threshold (96%), then, the test frame is allocated at this view; else, a new node is added to the sub-map of the global node and an edge between the new node and the last visited node is generated.

Figure 2: Sample example of matching a key frame with node consisting of SIFT features extracted from set of key frames representing same scene.

3.6 Topological Localization

The key element of our two-level topological localization method is the place recognition module. Usually place recognition modules need to determine the reference image that is most similar in appearance to the current input image, by comparing it with images of an entire database which can exceed thousands of images. Our proposed module, treats the previously learned set of images for the same reference location as group of features, assuming that this group of features is representative for each reference location, owing to this, the current input image is compared to features of each reference location which results in a fast matching process.

4 Experimental Results

The system is tested in the indoor office environment of The Computer Vision and Active Perception Laboratory (CVAP) at The Royal Institute of Technology (KTH) in Stockholm, Sweden. The robot was manually driven through the environment while continuously acquiring images at a rate of 5 fps.

ISBN: 978-960-474-308-7

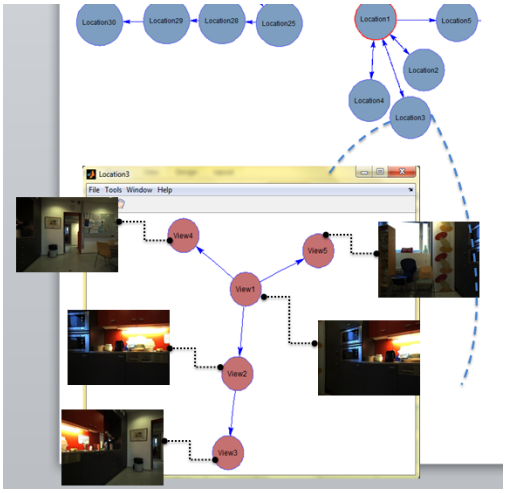


Figure 3: The decomposition of the global node 3 into nodes for different views.

4.1 Dataset Description

The dataset COLD-stochkolm [17] is used which is consisted of 9,564 images, each image in the training sequence is labeled and assigned to an ID and a semantic category of the area (usually a room) in which it was acquired. The environment consists of eight main rooms or areas, Corridor, Kitchen, LargeOffice, MeetingRoom, PrinterArea, RecycleArea, SmallOffice & Toilet. In this paper the right images (4,782 images) are used to simulate the realistic settings.

4.2 Results

Figure 4 shows the output of the experiment, in which the high level topological map is estimated. Figure 4(a) represents the ground truth of the environment and Figure 4(b) shows the output topological map. Figure 4(c) compares the estimated topological map by the robot to the ground truth of the environment which validates that the topological nodes have been correctly recognized by the robot to a high extend of accuracy. For example, the room category 'kitchen' is recognized and represented by three nodes in the global map: reference locations 2, 3 & 4. Figure 3 shows the output low level topological map. It shows the decomposition of node 3 to produce sub-map, which consists of fife views and connection between them.

Table 1 describes the distribution of the dataset, where only 965 images are selected as key frames from a total of 4782 images. The table shows that 906 key frames were correctly matched from the entire key frames.

5 Conclusion

In this paper, a global image signature together with a local feature extractor module is combined in a framework for mobile robot two-level topological localization and mapping. This approach allows a detailed two levels map building and robot localization in an indoor environment. The system reduced the number of images needed to describe the environment without losing important details by applying the key frame selection technique. The detected SIFT features are tracked and maintained and terminated based on the missed count C as explained above. The similarity matching of the global map level is achieved in an efficient way by comparing the test image with a set of features that represent a reference location instead of comparing it by all relevant images in the database. A successful experiment of accuracy 93.8% is presented, the output map is validated with the ground truth, which proved the validity of the proposed system, and the reference locations are correctly detected as well as the robot locations are correctly obtained during operation.

References:

[1] Cyrill Stachniss. *Robotic mapping and exploration*, volume 55. Springer, 2009.

[2] Muhammad Naveed, David Fofi, and Samia Ainouz. *Vision Based Simultaneous Localisation and Mapping for Mobile Robots*. PhD thesis, MasterŠs Thesis, Universit de Bourgogne, 2008.

[3] Hanafiah Yussof, editor. *Robot Localization and Map Building*. InfoTech, Berlin, 2010.

[4] Andrzej Pronobis, Oscar M. Mozos, Barbara Caputo, and Patric Jensfelt. Multi-modal semantic place classification. *The Int. J. of Robotics Research, Special Issue on Robotic Vision*, 29(2-3):298–320, 2010.

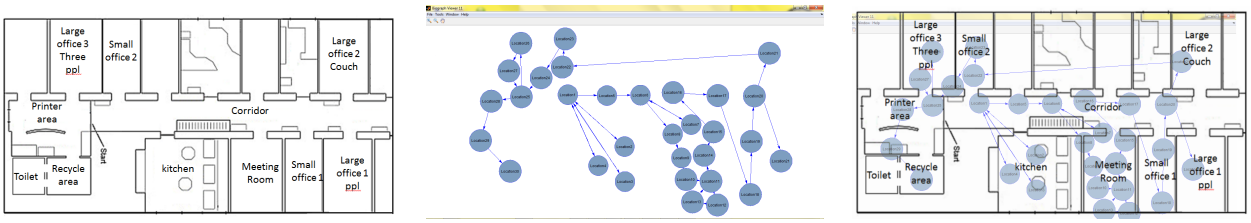
[5] Walter Lucetti and Emanuel Luchetti. Combination of classifiers for indoor room recognition. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[6] Olivier Saurer, Friedrich Fraundorfer, and Marc Pollefeys. Visual localization using global visual features and vanishing points. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[7] Bruno Siciliano and Oussama Khatib, editors. *Springer Handbook of Robotics*. Springer, 2008.

[8] Vivek Pradeep, Gérard G. Medioni, and James Weiland. Visual loop closing using multi-resolution sift grids in metric-topological slam. In *CVPR*, pages 1438–1445, 2009.

ISBN: 978-960-474-308-7



(a) (b) (c)

Figure 4: Output global topological map (a) Real Map of the Environment Floor no. 6, (b) The estimated global map of the system and (c) Comparison between the output global map and the ground truth.

Table 1: Dataset distribution and percentage of correctly classified key frames

Category	Number of Image	Number of Key Frames	Miss Matched	Corrected Matched
Corridor	1,146	240	31	209
Kitchen	600	57	21	36
Meeting Room	578	71	5	66
Small office	696	101	0	101
Large office	1,183	428	0	428
Printer area	210	53	2	51
Recycle area	177	15	0	15
Toilet	192	0	0	0
Total	4782	965	59	906
Accuracy Percentage				93.8%

[9] S. Thrun, D. Hähnel, D. Ferguson, M. Montemerlo, R. Triebel, W. Burgard, C. Baker, Z. Omohundro, S. Thayer, and W. Whittaker. A system for volumetric robotic mapping of abandoned mines. In *Proceedings of the IEEE Int. Con. on Robotics and Automation*, 2003.

[10] Mohammed A.-M. Salem. Multi-stage localization given topological map for autonomous robots. In *The 8th IEEE Int. Conf. on Computer Engineering and Systems (ICCES12)*, Cairo, Egypt, Nov. 29-30, Dec. 1 2012.

[11] Alberto Pretto, Emanuele Menegatti, Yoshiaki Jitsukawa, Ryuichi Ueda, and Tamio Arai. Image similarity based on discrete wavelet transform for robots with low-computational resources. *Robotics and Autonomous Systems*, 58(7):879–888, 2010.

[12] Sidharth Sood. Performance comparison of feature detectors for monocular visual slam. 2008.

[13] David G Lowe. Distinctive image features from scale-invariant keypoints. *Int. Jo. of Computer Vision*, 60(2):91–110, 2004.

[14] Stéphane G. Mallat. A theory for multiresolution signal decomposition, the wavelet representation. *IEEE Tr. on Pattern Analysis and Machine Intelligence*, 2(7):674–693, 1989.

[15] Mohammed A.-M. Salem. On the selection of the proper wavelet for moving object detection. In *The 7th IEEE Int. Conf. on Computer Engineering and Systems (ICCES11)*, Cairo, Egypt, November 29-30, December 1 2011.

[16] Mohammed A.-M. Salem. *Medical Image Segmentation: Multiresolution-based Algorithms*. VDM Verlag, Dr. Mueller, 2011.

[17] Andrzej Pronobis. The cold-stockholm database, 2009.

[18] Jose A. Gomez Jesus Martinez-Gomez, Alejando Jimenez-Picazo and Ismael Garcia-Varea. Combining invariant features and localization techniques for visual place classification: successful experiences in the robotvision@imageclef competition. *JOURNAL OF PHYSICAL AGENTS*, 5(1), JANUARY 2011.

[19] Mohamed A. Tahoun, Khaled A. Nagaty, Taha I. El-Arief, and Mohammed A.-Megeed Salem. A robust content-based image retrieval system using multiple features representations. In *IEEE Int. Conf. on Networking, Sensing and Control*, Arizona, USA, Mar. 19-22 2005.