Nested Graph Representation for Visual SLAM based on Local and Global Feature Processing

Sara Elgayar
Fac. of Comp. & Info. Sciences
Ain Shams University
Abbassia 11566, Cairo, Egypt

Mohammed A.-M. Salem
Fac. of Comp. & Info. Sciences
Ain Shams University
Abbassia 11566, Cairo, Egypt

Mohamed I. Roushdy
Fac. of Comp. & Info. Sciences
Ain Shams University
Abbassia 11566, Cairo, Egypt

ABSTRACT

Self localization of mobile autonomous systems is fundamental step in various applications, such as assistant navigation systems for blind people or smart house appliances. 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. A method to combine local and global features mapping have been proposed in a nested graph representation, where the indoor environment is divided into locations which is then decomposed into different views. The Scale Invariant Feature Transform is used to extract and build up a global map which provide rough estimation of the robot position. Horizontal, vertical and diagonal details of the wavelet coefficients are then used to provide finer estimation of the robot position and pose. The output topological map is validated with the ground truth of the environment. Moreover, the number of decomposition levels of the wavelet transform is analysed. The results show high localization accuracy and low rate of matching time.

General Terms:

I.2.9 Robotic, I.4 Image Processing and Computer Vision

Keywords:

Robot Vision, Local Features, Global Features, SIFT, Wavelet Transform, Topological Map

1. INTRODUCTION

Simultaneous localization and mapping (SLAM) or Concurrent Mapping and Localization (CML) as referred in [27] is one of the most extensively researched field of robotics.
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. 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 [27]. SLAM is significantly more difficult since robot poses and map are both unknown.
Topological V-SLAM is a SLAM process which is based on vision for environment sensing and used topological map for representing the environment. Where robot location is determined by capturing a picture (key frame) and compare it with the previously collected key frames in the evolving map. Then, it is localized at the position of the most similar matched key frame. Otherwise, if no match, then it is a new reference location. While map building can be achieved by adding new node to the evolving map whenever a new reference location is detected [19].
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: First, the related work about vision-based simultaneous localization and mapping is explained in section 2. Then, the mathematical basis of wavelets, and scale invariant feature extraction (SIFT) are introduced in section 3. The architecture of the proposed system is given in section 4. Section 5 describes the data set used and the experimental results followed by the conclusion in section 6.

2. VISION-BASED SIMULTANEOUS LOCALIZATION AND MAPPING

Figure 1 illustrates vision-based SLAM process. It demonstrates the egg-and-chicken relationship between the two main processes localization and mapping. The captured images, landmarks database and the map of the environment are the input for the localization process. On the other hand, the captured images, landmarks database and the robot locations are the input for the mapping process. 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 [16].

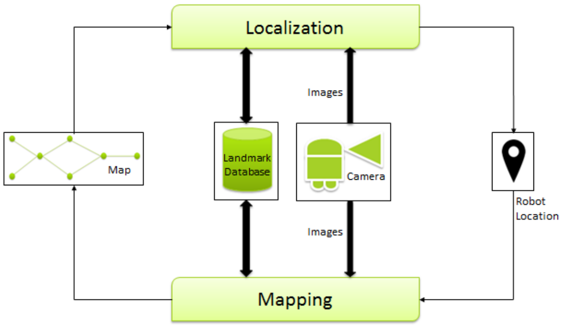


Fig. 1. Overview of Vision-based SLAM process.

Imaging sensors offer a variety of desirable properties. Cameras are low cost, provide a huge amount of information, available and passive. Moreover, vision offers numerous algorithmic solutions to essential problems such as: feature detection and matching, image segmentation and classification, object recognition and scene interpretation and image indexing. There are different types of the imaging devices , e.g. omnidirectional cameras, stereo cameras, multiple camera rigs and depth cameras. However, the use of single cameras have received increasing attention during the past few years. Some visual SLAM implementations using single cameras are [17][4].

Localization is a fundamental problem in mobile robotics. It has three approaches: geometric, topological and hybrid [23]. A 2D grid is usually used as a map representation for geometric-based approaches, where the exact robot location is tracked with respect to map coordinate system. Topological approaches use a simple graph as a map representation. The robot is localized if a node of the graph is recognized as the current location of the robot. If an approach combines the geometric and topological map representations it is said to be a hybrid approach [30]. Localization systems based on landmarks rely on either artificial or natural landmarks to represent the environment and try to determine correspondences between the observed landmarks and a pre-loaded map to estimate the location of the robot [31][14]. Artificial landmarks are easier to detect. However, they require modifications of the environment. Therefore, systems based on natural landmarks are often preferred. Various features could be used as natural landmarks such as corners, doors, windows, or wall colors. Typically, these systems are designed for specific environments and hence they can be hardly applied to different environments [30]. Another way for a robot to determine its location is through image statistics. The most popular image statistic is the histogram. 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 [25][6]. The most popular metric representation is grid-based maps, also known as occupancy grids. They represent the environment being mapped in form of cells. Each cell, is marked as free or occupied. It's very intuitive for humans. However, usually they don't scale well with the actual environment dimension [18]. The main idea of feature-based maps is to extract features from the environment (e.g. lines, corners, doors) and then to represent them by, for instance, colour, length, width, position, etc [29]. Topological maps represent environments as a list of significant places (nodes) that are connected via arcs (edges). Topological maps scale well to large environments, since the amount of information that is stored is limited to the description of the places. Topological maps yield ambiguities when representing the same place more than once, which are difficult to overcome. To overcome such problem, many authors proposed to use the topological map in a combination with metric or feature-based maps which is commonly know as hybrid maps [18][2].

3. 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. There are multiple ways to do feature extraction and it depends largely on what types of features are extracted as well as the used sensors. 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 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 [19]. The following sections explain the feature extractors used in this paper.

3.1 Local Features Processing

Indoor environments usually present high variability in their visual appearance, i.e. same visual feature can be captured from different angles or distances. 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 [23][9][26][24]. A lot of approaches have been proposed for local feature extraction before SIFT, but they aren't invariant to scale and more sensitive to projective distortion and illumination change [10]. SIFT was developed and published by David Lowe in 1999-2004. It aims at representing an image by a set of local interest points (visual features) which are invariant to image translation, scaling, and rotation and partially invariant to illumination changes and affine or 3D projection. SIFT algorithm consists of two main stages which are (1) detection of key points and (2) description of the detected key points.

Key points are detected & localized in two steps. First a pyramid of difference-of-Gaussian (DOG) images is created. The second step is to localize the key points by comparing each sample point in DOG image to its eight neighbours in the current image and the nine neighbours in the scale above and below. The point is selected if and only if it is significantly greater than (Maxima) or less than (Minima) all of them. A vector of 128 elements is built to describe each localized key point. The 16 x 16 neighbours of each key point are divided into 4 x 4 subregions, for each subregion the orientation histogram is computed in 8 bins, this leads to a SIFT feature vector with 4 x 4 x 8 feature element.

After running the SIFT algorithm, 1000-2000 key points are extracted from the image. To find the best match for key point x of image i in image j simply compute the Euclidean distance between key point x and all key points extracted in image j. The best match would be the nearest point found in image j. Image i and image j are said to be matched if a predefined number of the SIFT points of image i is found in image j.

3.2 Global Features Processing

The purpose of any transform is to make the job easier. Wavelet transform has been successfully used for vision based robot localization, vision based SLAM and image retrieval algorithms [19][11][28][15] for their capability in representing images in a compact way without losing information about location of the image discontinuity, shapes and textures. Wavelet has shown a better tool for non-stationary signal analysis than Fourier transform. The wavelet transform is a tool that decomposes signals into different frequency components, and then studies each component with a resolution matched to its scale. This way, wavelet transform provides a tool for time and frequency localization.Mallat [12] has proposed an iterative algorithm to compute the discrete wavelet transform. It is based on the multiresolution analysis. It applies a two-band subband coding procedure in an iterative fashion and builds the wavelet transform from the bottom up, i.e., small coefficients for small scales are computed first.

The algorithm is based on computing iteratively an approximation at a lower resolution level j of the original signal $f(t)$, which is in the original resolution level 0. 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 $\psi_{k,j}(t), k, j \in \mathbb{Z}$.

Daubechies constructed the first wavelet family of scale functions that are orthogonal and have finite vanishing moments, i.e. compact support [3]. This property insures that the number of non-zero coefficients in the associated filter is finite and assures the locality of the analysis. At a certain position k the corresponding coefficients $A_{j,k}(t)$ or $D_{j,k}(t)$ analyze $f(t)$ around k. So this analysis is local. The Haar wavelet ψ_{haar}, is the basis of the simplest wavelet transform. Historically, it is the first mention of what is called now “wavelet” in thesis by Alfred Haar in 1909. It is discontinuous and the only symmetric wavelet in the Daubechies family and the only one of them that has an explicit expression. It is a simple difference function. The associated filter is of length two. This means that the resulting approximation and detail images are all half the number of columns and rows. 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. This transform is performed by two separate 1D transforms along the rows and the columns of the image data constructing one 2D scaling function and three different 2D wavelet functions.

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. It results from averaging the image in both dimensions x and y. The horizontal detail H is obtained by averaging in the x-dimension and differencing in the y-dimension. The vertical detail V is obtained by averaging in the y-dimension and differencing in the x-dimension. The diagonal detail D is obtained by differencing in both dimensions and then averaging [7]. As shown in Figure 2 horizontal edges tend to show up in H and vertical edges in V, while D contains all other details [21].

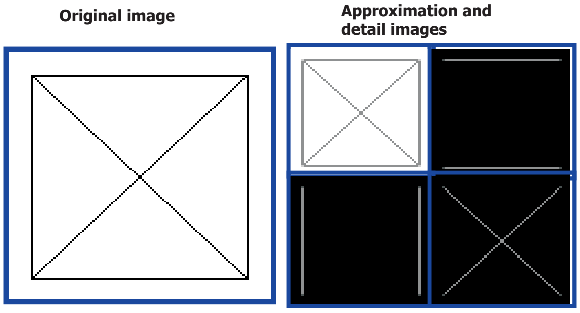


Fig. 2. Approximation and details of an image.

4. THE PROPOSED SYSTEM

This paper proposes a novel system for vision-based Robot SLAM. The input to the system is a sequence of key frames, and the output is a topological map represented by two-levels nested graph, as well as, two-levels local/global localization. The mobile robot starts off without a priori knowledge of the environment replying only on the visual information. A prototype of the proposed system was introduced in [5].

4.1 Sensing

A single camera is used. 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 compared to other capturing devices; besides the robot's stability is not affected due to its light weight. However, single cameras don't provide any information about the feature depth, for this reason there must be an extra processing over multiple frames to track the same feature in order to compute its 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, however, a monocular vision system is used in this paper. The images were continuously acquired at a rate of 5fps, more details about the used dataset can be found in [20].

4.2 Preprocessing

First, the captured image is converted to grey scale, and then the *"Next Increase"* procedure is applied to decide whether

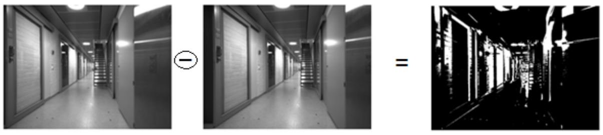


Fig. 3. Two successive captured images are subtracted to estimate the amount of change.

the captured image is a key frame or not. This procedure was successfully used in [13]. To demonstrate, key frames are images with a significant change between it and the previous accepted key frame, thus, eliminating the useless images and reducing the computation complexity without losing information. Next increase procedure simply computes the absolute difference between two images pixel by pixel, an example is shown in Figure 3. It keeps only the images whose difference between it and the last non-removed key frame is greater than a predefined threshold (0.08 for all images).

4.3 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 [28][1]. The authors believe 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-level localized, for example in which room the robot is, and in which corner the robot is. Global image signatures and local features are combined in the same framework as shown in Figure **??**. 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.

4.4 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.

Feature Tracking Initiation. Initially the map is empty. When SIFT features from the first frame arrive, a new map node is created. Each extracted SIFT interest point p_i, is saved in the node as: $[X_i, Y_i, S_i, O_i, D_i, C_i]$ Where (X, Y) is the current 2D position of the SIFT landmark relative to the initial coordinates frame, (S, O) are the scale and orientation of the landmark, D is a set of descriptors (n x 128) associated with each interest point, and C is a count to indicate how many consecutive frames this landmark has been missed. Initially this count is set to 0.

Feature Tracking Maintenance. Over subsequent frames, the map is maintained, 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 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.

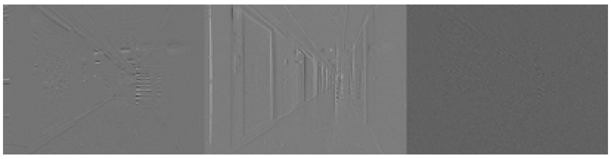


Fig. 4. 4^{th} level of 2D Haar discrete wavelet signature (horizontal, vertical & diagonal details, respectively.)

Feature Tracking Termination. If the missed count C of any feature in the map reaches a predefined limit N , i.e., this feature has not been observed at the location it is supposed to appear, therefore 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. Figure 4 presents an example of the computed wavelet signature.

4.5 Mapping

In the proposed approach the environment map is represented by adjacency graph. Nodes of the graph represent locations, while arcs represent the adjacency relationships between the locations. The general approach of map building is to incrementally integrate new nodes into the map. Each node in the graph of the first level (locations) is a rich node that contains information about a reference location (ID, Label, Matched key frames, The set of SIFT interest points, A count for the missed matches for each SIFT interest point, The total number of SIFT Interest points, and the total number of key frames matched). Similarly, each node in the second level is also a rich node that contains the following information (ID, Label, and the wavelet signatures of the matched key frames).

The complete process for two-level mapping can be summarized in the following steps: First, SIFT points are extracted from the captured key frame. 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. as show in Figure 5, accordingly SIFT features extracted from one key frame can match SIFT features from multiple key frames representing same scene (captured at close locations). Clearly, the idea of matching the test image with a set of SIFT features that represents a reference location, is much preferred than matching the test image with all images representing reference location in the database, it provides wider range of matching and saves extensive search time complexity and memory requirement. The situation is shown in Figure 6. Finally, a ranking with the best n

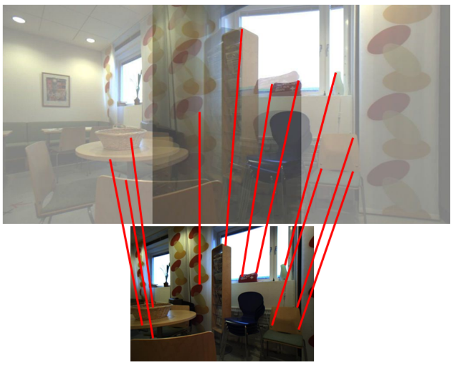


Fig. 5. Sample example of matching a key frame with node consisting of SIFT features extracted from set of key frames representing same scene.

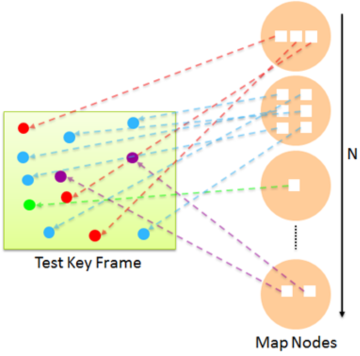


Fig. 6. Similarity computed between an input key frame and global map nodes.

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 used in the experiments) 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 their 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.

4.6 Localization

The key element of the proposed 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. The proposed place recognition module, treat the previously learned set of images for the same reference location as group of features, owing to this, the current input image is compared to features of each reference location which results in a fast matching process. For the global level localization, the place recognition system depends on SIFT features extracted from the new input image and match it with SIFT features of each reference location. The input image is localized and is given a position label that is associated with the matched global node. Furthermore, for the local level localization, the place recognition system utilizes image signatures, created by the standard 2D Haar discrete wavelet decomposition technique. When the system is presented with a new image, the image signature is computed and then compared with all wavelet signatures in the same global node to determine the nearest view. The current image is localized locally and is given the label of the view associated with the nearest matched image signature.

5. EXPERIMENTAL RESULTS AND DISCUSSION

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 [20]. The robot was manually driven through the environment while continuously acquiring images at a rate of 5 fps. The dataset will be explained in the following section.

5.1 Dataset Description

The dataset COLD-Stockholm [20] 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 sequences were acquired using the PowerBot robot platform equipped with a stereo camera system. However, monocular system can be used by ignoring the left or the right images. The environment consists of eight main rooms or areas, Corridor, Kitchen, LargeOffice, MeetingRoom, PrinterArea, RecycleArea, SmallOffice & Toilet.

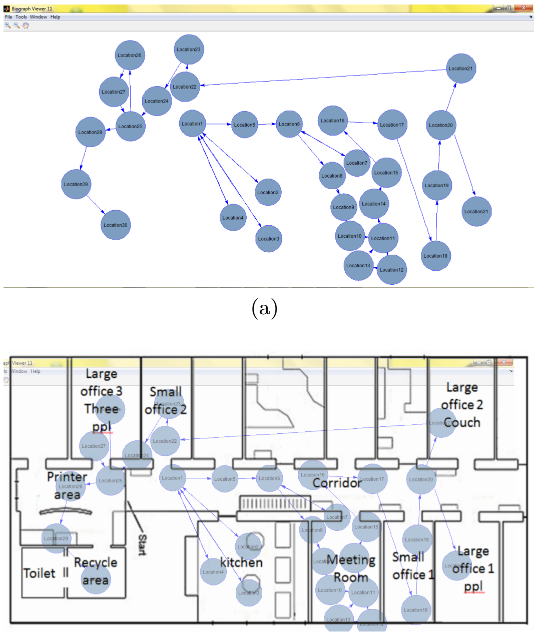
5.2 Evaluation Metrics

The results were evaluated manually with the help of the annotation of the images in the datasets. Each frame was labelled either correctly matched or miss matched based on the estimated classification of the system and the ground truth of the datasets.

The results were also evaluated according to the confusion matrix described in [8]. Where Accuracy AC, Recall TP and Precision P are calculated.
The accuracy AC is defined as, the proportion of the total number of predictions that were correct. The recall or true positive rate TP is defined as, the proportion of positive cases that were correctly identified. The precision P is defined as, the proportion of the predicted positive cases that were correct.

5.3 Results

Figure 7 shows the output of the experiment, in which the high level topological map is estimated. Figure 7(a) shows the output topological map augmented on the ground truth of the environment. Figure 7(b) 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. The set of observed nodes are colored in blue and edges represent the connection between reference locations (Rooms). For example, the room category 'kitchen' is recognized and represented by three nodes in the global map: reference locations 2, 3 & 4.



(a)

(b)

Fig. 7. The resulted global map: (a) The estimated global map of the system and (b) An overlay of the output global map and the ground truth.

Figure 8 shows the output low level topological map. It shows the decomposition of node 3 to produce sub-map, which consists of five views and connection between them. A visual example of matching a query frame to a location and view is shown in Table 1. For the high level topological localization, a ranking with the best 3 nearest locations are obtained associated with the number of similarity interest points, the test key frame is high level estimated at the location with highest matching interest points. On the other hand, for the low level topological localization, when the best matched scenes don't exceed the predefined threshold (96%), a new scene is added to the map, and the current key frame localized locally.
Table 2 describes the distribution of the right and left images 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. For the left images of the dataset, the table shows the distribution of the key frames similar to the right images.
Table 3 shows the evaluation based on the confusion matrix for the right and left images, where the average Accuracy, precision and recall were calculated.
The matching rate is slightly differs between the right and the left images, However, it lays in the range between 90.2% and 93.8%. An interesting observation, as the number of nodes (global and local) increases, the matching rate also increases. This is observed regarding the physical rooms as well as the overall matching.

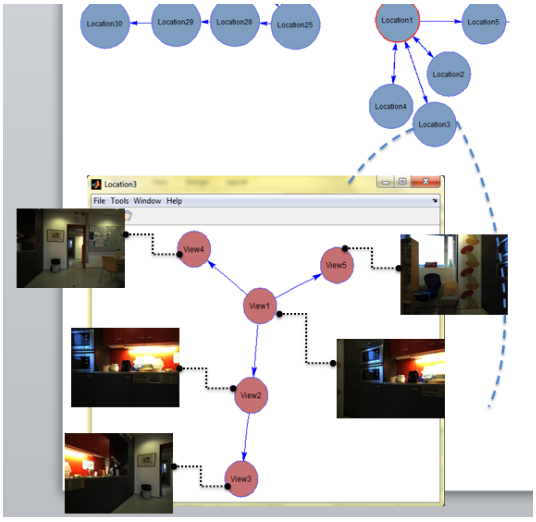


Fig. 8. The decomposition of the global node 3 into nodes for different views.

Table 2. Left & Right datasets distribution and percentage of correctly classified key frames.

Category	No. of Images	No. of Key Frames		No. of Locations G. Nodes		No. of Views L. Nodes		Correct Match	
		Right	Left	Right	Left	Right	Left	Right	Left
Corridor	2,292	240	235	8	6	46	57	209	171
Kitchen	1,200	57	55	3	3	21	18	36	40
Meeting Room	1,156	71	67	10	8	38	33	66	50
Small office	1,392	101	102	8	6	36	36	101	87
Large office	2,366	428	425	3	2	34	23	428	418
Printer area	420	53	57	2	2	22	13	51	55
Recycle area	354	15	15	1	1	9	8	15	15
Toilet	384	0	0	0	0	0	0	0	0
Total	9,564	965	965	35	26	206	185	906	863
Matching Rate								93.8%	90.2%

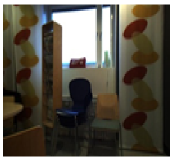
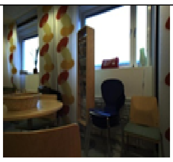
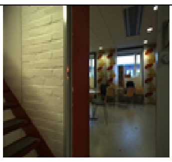
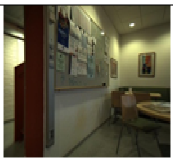
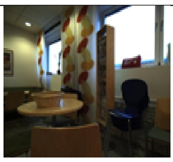
Table 3. The average Accuracy, Recall and Precision according to the confusion matrix of the right & left datasets.

Category	a	b	c	d	(AC)	(TP)	(P)
Corridor	695.5	47.5	43.5	190	0.9	0.8	0.79
Kitchen	865.5	18	4.5	38	0.97	0.9	0.67
Meeting Room	852.5	11	3	58	0.98	0.94	0.83
Small office	820	7.5	5.5	94	0.98	0.94	0.9
Large office	480	3.5	26	423	0.96	0.94	0.99
Printer area	860	2	5.5	53	0.98	0.91	0.96
Recycle area	898	0	5.5	15	0.98	0.75	0.98
Toilet	914	0	0	0	1	0	0
Average	798.1	11.1	11.6	108.8	96.8%	77%	76%

5.4 Discussion

In previous work [22], An analysis have been done on the selection of the mother wavelet transform for feature de-

Table 1. Example of two-level localization of a test key frame.

Test key frame Actual Location: 3			
Nearest Matched Locations (High Level Topological Localization)			
Number of common IP's	Location3: 116	Location1: 32	Location2: 0
Nearest Matched Views in Location 3 (Low Level Topological Localization)			
Similarity	View 1: 89.5%	View 3: 86.4%	View 1: 86.3%
Estimated Two-Level Topological Localization (Location# , View#)	High Level Localization: Location 3. Low Level Localization: New View is added to the map 'View 5'.		

tection and localization. The Haar wavelet has better feature localization property in a comparison with the higher ordered Daubechies wavelets. Based on this analysis it is preferred to use the Haar wavelet for creating image signature based on the 4th decomposition level. Further analysis have been done on the number of the decomposition levels. Figure 9 shows, the relation between level of wavelet decomposition and number of decomposed views is inversely proportion. As the level of decomposition increases, the signature's size is significantly reduced, but higher level decomposition discard important features in the image, like edges and high frequency patterns, useful for environment characterization. On the other hand, as the level of decomposition increases, the number of views (nodes in the submap) formed increases, which limits the matching between frames to a restricted settings, i.e. each captured image will be treated as a separate view, as a result, the decomposed local node will represent a narrow view. Level 4 is chosen for decomposition because the trade-off between a compact representation and a reliability similarity computation. Another reason is the compromise between a reduced size of wavelet signature and a controlled number of the splitting nodes of the local map.

6. CONCLUSION

In this paper, a global image signature together with a local feature extractor module is combined in a framework for mobile robot nested-based topological localization and mapping. 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. Successful experiments are presented using COLD-Stockholm database [20], 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. The proposed system succeeds in achieving an overall matching of 92% and an overall retrieval accuracy of 97%.

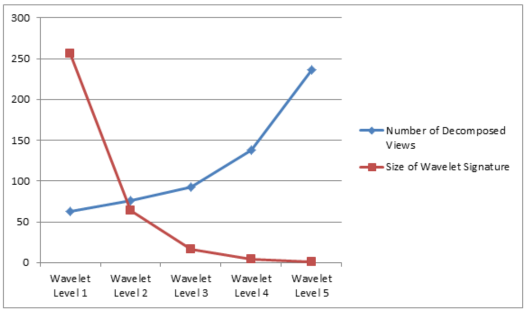


Fig. 9. Analysis of wavelet decomposition levels.

In the future, some optimization techniques are intended to be used to group the different global nodes of the same location in one class. So that the graph is extended by a third level from above, where each node in that level represents a class gathering all nodes of the same location of the ground truth data. The objective is to represent each real world room by a single class.

7. REFERENCES

[1] Mariam Al-Berry, Mohammed A.-Megeed Salem, A S. Hussein, and M. F. Tolba. Spatio-temporal motion de-

tection for intelligent surveillance applications. *submitted to the International Journal of Computer Mathematics*, 2013.

[2] J.-L. Blanco, J.-A. Fernandez-Madrigal, and J. Gonzalez. Toward a unified bayesian approach to hybrid metric–topological slam. *Robotics, IEEE Tr. on*, 24(2):259–270, 2008.

[3] Ingrid Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

[4] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1052–1067, 2007.

[5] Sara Elgayar, Mohammed A.-Megeed Salem, and Mohamed I. Roushdy. Two-level topological mapping and localization based on sift and the wavelet transform. In *2nd International Conference on Circuits, Systems, Communications, Computers and Applications*, Dubrovnik, Croatia, June 25-27 2013.

[6] J. Guivant, J.I. Nieto, and E Nebot. The hybrid metric maps (hymms): A novel map representation for denseslam. In *Proceedings of IEEE 2004 Int. Conf. on Robotics and Automation*, 26 Apr. - 01 May 2004.

[7] Gerald Kaiser. The fast haar transform, gateway to wavelets. *IEEE potentials*, April-May 1998.

[8] Kohavi and Provost. Confusion matrix, 1998.

[9] Thomas Lemaire, Cyrille Berger, Il-Kyun Jung, and Simon Lacroix. Vision-based slam: Stereo and monocular approaches. *International Journal of Computer Vision*, 74(3):343–364, 2007.

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

[11] Wen Lik Dennis Lui and Ray Jarvis. A pure vision-based topological slam system. *The International Journal of Robotics Research*, 31(4):403–428, 2012.

[12] 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.

[13] Jesus Martinez-Gomez, Alejando Jimenez-Picazo, Jose A. Gomez, 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.

[14] M. Mata, J. M. Armingol, A. De La Escalera, and M. A. Salichs. Using learned visual landmarks for intelligent topological navigation of mobile robots. In *IEEE Int. Conf. on Robotics and Automation*, 2003.

[15] Ana Cris Murillo, JJ Guerrero, and C Sagues. Surf features for efficient robot localization with omnidirectional images. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 3901–3907. IEEE, 2007.

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

[17] Richard A Newcombe and Andrew J Davison. Live dense reconstruction with a single moving camera. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1498–1505. IEEE, 2010.

[18] 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.

[19] 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.

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

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

[22] 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*, Cairo, Egypt, November 29-30, December 1 2011.

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

[24] Stephen Se, David G Lowe, and James J Little. Vision-based global localization and mapping for mobile robots. *Robotics, IEEE Transactions on*, 21(3):364–375, 2005.

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

[26] Robert Sim, Pantelis Elinas, Matt Griffin, and James J Little. Vision-based slam using the rao-blackwellised particle filter. In *IJCAI Workshop on Reasoning with Uncertainty in Robotics*, volume 14, pages 9–16, 2005.

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

[28] 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.

[29] 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.

[30] Iwan Ulrich and Illah Nourbakhsh. Appearance-based place recognition for topological localization. In *Int. Conf. on Robotics and Automation, ICRA'00*, pages 1023–1029, San Francisco, USA, 2000.

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