Occupancy Grid Mapping for Mobile Robot Using Scale Invariant Feature Transform

Utpal Kant
Department of Instrumentation and Control Engineering
Dr. B.R.A. National Institute of Technology,
Jalandhar, India
utpal0401@gmail.com

Kuldeep Singh Nagla
Department of Instrumentation and Control Engineering
Dr. B.R.A. National Institute of Technology,
Jalandhar, India
naglaks@nitj.ac.in

Abstract— **In order to achieve full autonomy for a mobile robot, it is essential to sense the environment accurately. Sensor plays a fundamental role in the autonomy process of mobile robot. Various types of sensor such as sonar sensor, infrared sensor, laser sensor, vision sensor etc. are used for environment mapping. Each of the above mentioned sensors has its own drawbacks. The sonar range finders for instance suffer from wide beam cone, specular reflection, crosstalk etc. On the other hand lasers available in market are expensive and output of the sensor depends upon the surface brightness. The stereo vision systems are very sensitive to changes in illumination and the algorithms used for vision are computationally expensive. Hence it is clear that it is difficult for one sensor alone to give satisfactory information for mapping of mobile robot's environment. Researchers used many ways to improve sensory information such as by using sensor fusion techniques etc. The proposed works is focused upon improvement of vision accessibility in mobile robot. The first section deals with the selection of appropriate sensor for particular environment. The second section deals with optimization of threshold ratio used in Scale Invariant Feature Transform (SIFT). The occupancy grid mapping of environment is computed by using SIFT feature modeling and fusion of grid (using recursive Bays rule) with prior information is given in the last section of the dissertation.**

Keywords- mobile robot; sensor fusion; occupancy grid mapping; scale invariant feature transform (SIFT); machine vision

I. INTRODUCTION

There are a large number of algorithms available for disparity map using stereo vision system. The correspondence problem in stereo vision system is the problem of finding which element in the right image corresponds to the left image. The two main approaches in finding the solution to the correspondence problem in a stereo pair of images are correlation-based and feature based. The correlation-based correspondence algorithms usually find the correspondence between pixels in both images creating a dense disparity map were the 3D position of each pixel can be computed. The feature based correspondence algorithms find correspondence between features from both images. These features are edges, lines, points etc. Certain criteria must be satisfied in order to find the correspondence between features; for instance criteria for matching a line could be the length of the line, its orientation, the coordinates of a midpoint and the average contrast along the line edge. The disparity maps created by the feature-correspondence algorithms are more sparse than the ones created by the correlation-based, which makes the correspondence search less time consuming. The approach used in this paper is the scale invariant feature transform, because SIFT- feature based correspondence reduces the number of matches between left and right images. And examined effects of threshold value on disparity matching using SIFT.

This chapter deals with image acquisition and various steps of SIFT algorithm for disparity matching. It includes an empirical technique for optimization of threshold ratio and a distance measurement method to measure the distance to an object which is located outside the axes of stereo cameras. The simulation results of the measured distance of the objects in the dynamic environment are shown by using Vision Lab Feature Library (VL feat) toolbox in MATLAB.

In this chapter, SIFT feature model has been implemented for mapping the occupancy grid of indoor environment the mobile robot has to deal with. Acquisition of priory information to the environment is performed by moving the mobile robot having a sonar sensor mounted on it. The range information given by sonar sensor and its orientation gives the polar coordinate of the occupied cell. The coordinates of occupied cell is calculated by the transformation of polar coordinates given by sonar sensor to the Cartesian coordinates.

Moravec- Elfes model has been used for generating the occupancy grid by SIFT feature model. And the final occupancy grid has been generated by updating the occupancy grid generated by SIFT by using Recursive Bays Rule. The priory to the environment is used for updating the occupancy

grid. The chapter is concluded by discussing the results obtained and applications in this field.

II. WORK ARCHITECTURE

The environment the mobile robot has to deal with is an indoor environment. There is a variety of approaches to the integration and fusion of information from the combination of different types of sensors. Sonar sensor has been used for acquiring the prior information from the static environment.

The vision system is one of the most powerful sources of information. The vision algorithms used in the field of mobile robots are mainly based on feature extraction, where the features are edges or corners. This fact led to the exploration of new alternatives for extracting features from the scene to use them in the field of sensor fusion in mobile robot. It does provide a good basis for matching images of different sizes. SIFT algorithm is used to extract the features from the stereo snapshots taken from the robot during its path.

Acquisition of priory information to the environment is performed by moving the mobile robot having a sonar sensor mounted on it. The range information given by sonar sensor and its orientation gives the polar coordinate of the occupied cell. The coordinates of occupied cell is calculated by the transformation of polar coordinates given by sonar sensor to the Cartesian coordinates.

SIFT algorithm for disparity matching has been implemented for the measurement of distance of the objects. This is performed by a proposed empirical technique for optimization of threshold ratio and a distance measurement method to measure the distance to an object which is located outside the axes of stereo cameras.

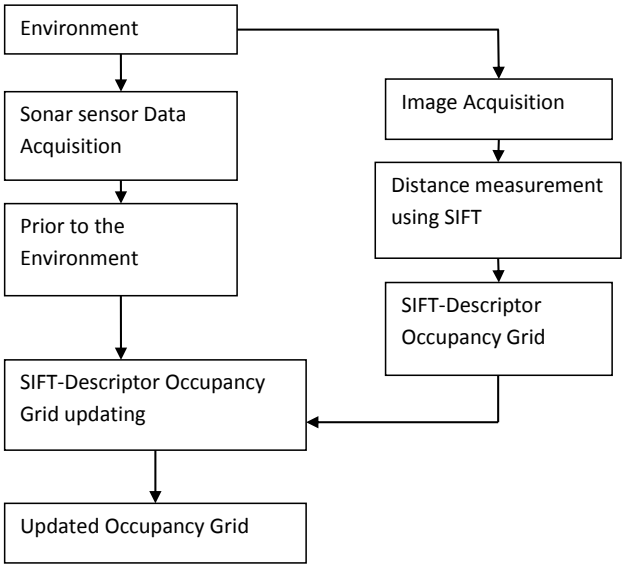


Figure 1. Work Architecture

Mathies and Shafer [1] successfully modeled the triangulation error in stereo matches with normal distributions. This approach has been taken into account on modeling the occupied area of the SIFT-descriptor. The Elfes approach has been considered to model the empty areas from the cameras to the boundary of the 3D normal distribution. These two models led to a SIFT- descriptor probabilistic model. The SIFT-descriptors identified by the SIFT algorithm in a snapshot is modeled and registered into occupancy. The priory information of the environment is used to integrate the SIFT-descriptor occupancy grid by using Recursive Bays formula.

III. IMAGE ACQUISITION

Stereo pair of images is being taken by using single CCD camera, Sony cyber shot (DSC-S650). The camera is placed on a slider at two different position separated by a distance of 5cm keeping the axis of the cameras parallel. Parallel camera setup has been shown in ―Fig. 2".

The features of the camera are given in ―Table I" and the ―Fig. 3", shows the acquired stereo pair of images.

TABLE I. FEATURES OF CAMERA (DSC-S650) FOR SIMULATION

Item			Feature
Setup Method			Parallel Setup
Distance between two cameras			50mm
Focal Length			5mm
CCD Camera	Size	1 Pixel	8.6µm X 6.6µm
		Whole	1280pixel X 960pixel
	Resolution	Horizontal	72dpi
		Vertical	72dpi

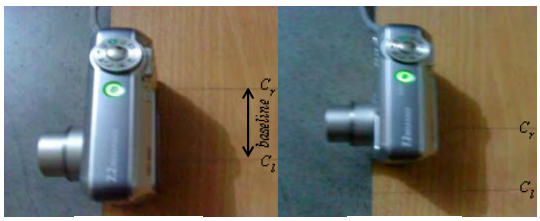
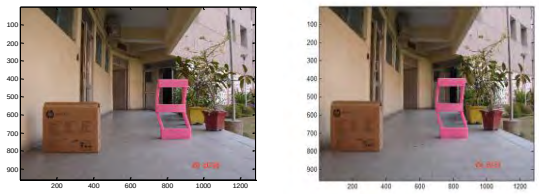


Figure 2. Parallel Camera Setup



Left Image Right Image

Figure 3. Stereo pair of images

IV. EXPERIMENTAL SETUP AND OCCUPANCY GRID MAPPING RESULTS

The whole experimental work has been performed in four different experiments. First experiment is about acquisition of prior information from the static environment the mobile robot has to deal with. Second experiment discusses the method for distance measurements of the objects in the dynamic environment using SIFT. The third experiment deals with modeling the occupied area and empty area using SIFT feature descriptor model. And the fourth experiment is integrating the occupancy grid with prior information about the environment by Recursive Bays Rule.

V. ACQUISITION OF PRIOR INFORMATION

A. *Experimental setup*

The static environment in which the mobile robot has to deal with is shown in "Fig.4". The environment is known and the position of static objects in the environment is as shown in figure. The area beyond right of the transparent line is at the depth of 0.6m from the level of plane where mobile robot has to move. Some assumption has been taken while mapping the environment according to the priory. While mapping the occupied area we assumed that the cells lying on position of the transparent line are occupied.

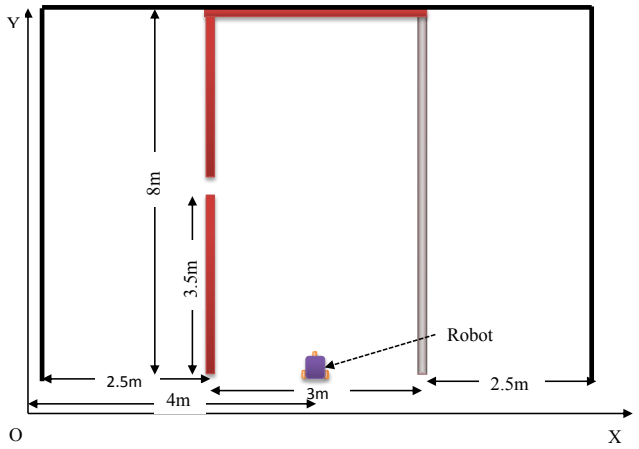


Figure 4. Geometric map of the static environment

Acquisition of priory information to the static environment is performed by moving the mobile robot having a sonar sensor mounted on it. The range information given by sonar sensor and its orientation gives the polar coordinate of the occupied cell. The coordinates of occupied cell is calculated by the transformation of polar coordinates given by sonar sensor to the Cartesian coordinates.

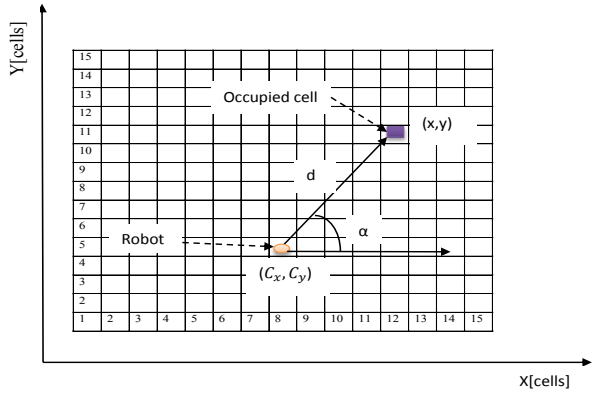


Figure 5. Occupancy grid of 15x15 cells with robot position (C_x , C_y) and occupied cell having polar co-ordinate (d, α)

The Cartesian coordinates of the occupied cell has been calculated by using equation "1" and "2".

$$x = C_x + d\cos\alpha \quad (1)$$

$$y = C_y + d\sin\alpha \quad (2)$$

B. *Simulation results*

The range information of the static environment has been acquired by moving the mobile robot throughout the environment that the mobile robot has to follow. The mobile robot having a sonar sensor mounted on it, sonar sensor is connected with a laptop by using Doctor DAQ data acquisition device as shown in "Fig. 6".

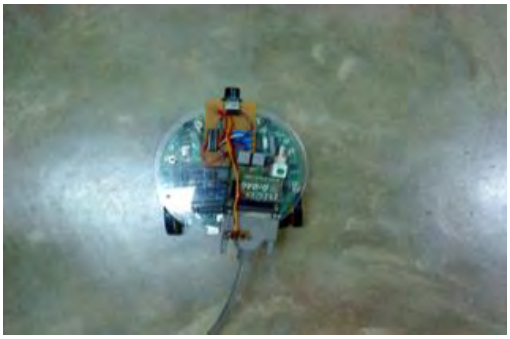


Figure 6. Mobile robot having sonar sensor mounted on it

The range information given by sonar sensor is stored in a MAT file. The range information is set as input to a Simulink programme for the transformation of polar coordinates of the occupied cell to the Cartesian coordinates. Whole static environment has been mapped by using an occupancy grid of 100x100 cells, each cell having size of 8cmx8cm. According to acquired prior information the occupancy grid map of the static environment is shown in "Fig. 7".

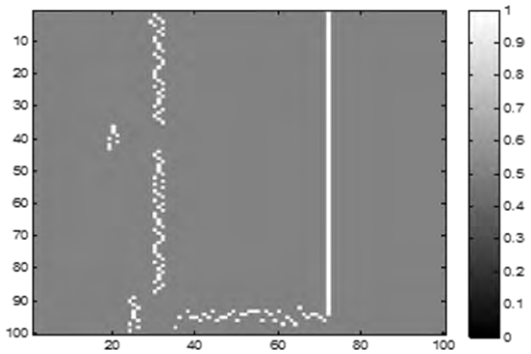


Figure 7. Occupancy grid map of the environment according to the priory

VI. DISTANCE MEASUREMENT USING SCALE INVARIANCE FEATURE TRANSFORM (SIFT)

A. *The Scale Invariance Feature Transform*

The Scale Invariance Feature Transform (SIFT) method was originally introduced by Lowe [2] and refined by the same author [3].

It is a method for extracting distinctive invariant features from digital images, in which the features are invariant to scale and rotation. They also provide a robust matching across a substantial range of affine distortion, change in 3D view point, addition of noise and change in illumination.

The SIFT algorithm consists of the following major steps:

- Scale-space peak detection
- Accurate key-point localization.
- Majority orientation assignment
- Computation of the local image descriptor.
- SIFT descriptor matching.

1) Scale-space peak detection: This is the stage where interest points, which are called key-points in the SIFT, are detected. For this, the image is convolved with Gaussian filters at different scales, and then the differences of successive Gaussian-blurred images are taken. Key-points are then taken as maxima/minima of the Difference of Gaussian (DoG) that occur at multiple scales.

A Scale-space function is defined as;

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \tag{3}$$

Where: $G(x,y,\sigma)$ be the Gaussian kernel and $I(x,y)$ be the input image.

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{4}$$

The scale parameter σ just indicates which scale level is being defined.

The Difference of Gaussian function is defined as;

$$DoG(x,y,\sigma) = G(x,y,k\sigma) - G(x,y,\sigma) \tag{5}$$

The Difference of Gaussian image $D(x,y,\sigma)$ is defined as;

$$D(x,y,\sigma) = L(x,y,k\sigma) - L(x,y,\sigma) \tag{6}$$

D.G. Lowe [3] Proposes to find potentially interesting key-point locations by detecting extrema in the scale-space representation obtained from convolving the image with the DoG function.

In the implementation, before applying the formula –(6)", the original image, $I(x,y)$, is up sampled by a factor of two using linear interpolation. [3] Mentions that, by up sampling, the image has the effect of increasing the number of stable key-points by almost a factor of 4. Afterwards the up sample image is incrementally blurred by a Gaussian kernel. Blurring the original image by a Gaussian kernel has the equivalent effect of low-pass filtering, discarding the higher frequencies. The creation of images separated by a constant factor k in scale space, $L(x,y,k\sigma)$, brings the concept of an octave.

An octave O_c is defined as the creation of incrementally convolved images with a Gaussian kernel in scale space. An octave in scale space is divided into a fix integer number s_u of intervals. The maximum of intervals are chosen as doubling σ, so $k = 2^{\frac{1}{s_u}}$.

$$O_c = L(x,y,k^i\sigma) = G(x,y,k^i\sigma) * I(x,y), \quad i = 0,\ldots,s_u+1 \tag{7}$$

$s_u + 1$ Images must be produced in the stack of blurred images for each octave, so the final extrema detection covers a complete octave. Lowe's implementation uses $\sigma = \frac{1}{2}$

After the first octave has been created, the Gaussian image $I(x,y)$ that has twice the initial value of σ (it is the second image) is down sampled by a factor of two i.e. by taking every second pixel in each row and each column. A new octave is created at half resolution. This process is repeated until the image reaches a predetermined minimum size. This process produces an octave pyramid, OP, consisting of different octaves, as seen in –(8)".

$$OP = \begin{bmatrix} L_1(x,y,k^i\sigma) & i = 0\text{K}\ ,s_u+1 \\ L_2(x,y,k^i\sigma & i = 0\text{K}\ ,s_u+1 \\ & \text{M} \\ L_n(x,y,k^i\sigma) & i = 0\text{K}\ ,s_u+1 \end{bmatrix} \tag{8}$$

Adjacent images in each octave in the pyramid are subtracted to produce the DoG functions' pyramid of octaves at different scales, as seen in –(9)".

$$DP = \begin{bmatrix} L_1(x,y,k^{i+1}\sigma - L_1(x,y,k^i\sigma) & i = 0\text{K}\ ,s_u \\ L_2(x,y,k^{i+1}\sigma) - L_2(x,y,k^i\sigma) & i = 0\text{K}\ ,s_u \\ & \text{M} \\ L_n(x,y,k^{i+1}\sigma) - L_n(x,y,k^i\sigma) & i = 0\text{K}\ ,s_u \end{bmatrix} \tag{9}$$

Local extrema detection is achieved by comparing each pixel in the current image with its eight neighbours and nine neighbours in the image above and below in the current DoG octave. More specifically each pixel is compared to its eight neighbours at the scale k^i and nine neighbours at the scales k^{i-1} and k^{i+1} respectively. A pixel is selected only if it is larger than all of its neighbours or smaller than all of them.

2) Accurate key-point localisation: In the previous section, key-point candidates were found by a process called scale-space extrema detection. In this process too many key-point candidate were produced, some of which were unstable. The next step in the algorithm is to perform a detailed fit to the nearby data for accurate location, scale, and ratio of principal curvatures. This information allows points to be rejected that have low contrast or poorly localised along an edge.

As in [4], where it is mentioned that the location of the extrema to a sub-pixel Laplacian. This approach uses Taylor expansion (up to quadratic terms) of the scale space function $D(x,y,\sigma)$ which is shifted so the origin is at the interest point. The Taylor expansion of $D(x,y,\sigma)$ around the detected key-point location can be seen in –(10)".

$$D(z) \approx D(z_0) + \frac{\partial}{\partial z} D^T(z)\Big|_{z_0} (z - z_0) + \frac{1}{2}(z - z_0)^T \frac{\partial}{\partial z} D(z)\Big|_{z_0} (z - z_0) \tag{10}$$

The function $D(x,y,\sigma)$ and its derivatives are evaluated at the sample point $z_0=[x_0,y_0,\sigma_0]^T$ and $\Delta z=[\Delta x,\Delta y,\Delta\sigma]^T$ is the sub-pixel offset from this point.

The approximate sub-pixel location of the exterma is computed by taking the derivative of –(10)" with respect to z and setting it to zero, and then evaluated at z* which is the location at extremum.

$$\frac{\partial^2}{\partial z^2}D(z)\Big|_{z_0}z^*=\frac{\partial^2}{\partial z^2}D(z)\Big|_{z_0}z_0-\frac{\partial}{\partial z}D(z_0) \quad (11)$$

–Equation (9)" is a linear system of the form A $z^*=b$, where $\frac{\partial^2}{\partial z^2}D(z)$ is the Hessian matrix.

[4] Suggests approximating the derivatives of $D(z)$ by using differences of neighbouring sample points.

If the final offset z*- z_0 is larger than 0.5 in any dimension. This means that the extremum lies closer to a different sample point than z_0. In that situation, the sample point is changed to the closest one, and the process of interpolation is repeated for that sample point.

[4] Mentions that locating interest points at sub-pixel accuracy is especially important close to the top of the pyramid, where the resolution of the image is very low. This is because the samples spaces at levels in the pyramid correspond to large distances relative to the base image.

3) Eliminating unstable key-points: The extrema locations that exhibit a low function value of $|D(z)|$ are sensitive to additive noise and therefore unstable. In order to reject interest points with low contrast, the function D(z*) at extremum is evaluated. This is done by substituting –(11)" into –(10)", giving –(12)". In [3] all extrema values with a value of $|D(z^*)|<0.03$ are discarded, assuming image pixel values in the range [0, 1].

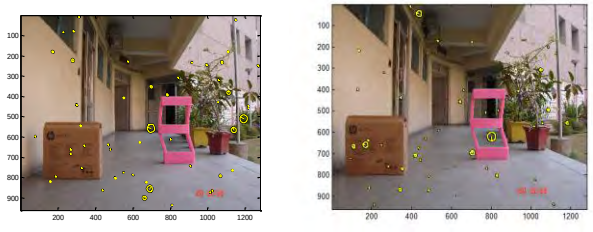
$$D(z^*)=D(z_0)+\frac{1}{2}\frac{\partial}{\partial z}D^T(z)\Big|_{z_0}(z^*-z_0) \quad (12)$$

Another set of locations that may have stability problems are those located along edges of the DoG function. According to [3], the DoG function has the property that it produces strong responses along edges, even if the location along the edge is poorly determined. This property makes the locations sensitive to small amounts of noise and therefore unstable. For this reason, it is not sufficient to eliminate interest points with low contrast. In order to get better stability further precautions must be taken. A poorly defined peak in the Dog will have the property that it will have a larger principal curvature across the edge and a small one in the perpendicular direction. Finding these principal curvature amounts to solving for the eigenvalues of the second order Hessian matrix (H),

If λ_1 and λ_2 be the eigenvalues with the largest and smallest magnitude respectively. Then, defining $r_a=\lambda_1/\lambda_2$ .Thus,

$$R_a=\frac{(r_a+1)^2}{r_a} \quad (13)$$

R_a is minimum when the Eigenvalues are equal to each other. Therefore, the higher the absolute difference between the two eigenvalues, which is equivalent to a higher absolute difference between the two principal curvatures of D(z), the higher the value of R_a. It follows that, for some threshold eigenvalue ratio r_{th} , if R_a for a candidate key-point is larger than $\frac{(r_{th}+1)^2}{r_{th}}$, that key-point is poorly localised and hence rejected.



Key points for Left Image Key-points for Right Image
Figure 8. Accurate key-points localisation

4) Majority orientation assignment: This step is done by assigning a consistent orientation to each key-point. This step of the algorithm involves the scale at which the key-point was detected.

First, a gradient is defined which provides two pieces of information, magnitude and direction; it is natural to encode this information in a vector. The length of this vector provides the magnitude of the gradient, while its direction gives the gradient direction. Because the gradient may be different at every location, it is represented with a different vector at every image location.

The gradient magnitude m_g and the gradient orientation θ_d of a Gaussian-smoothed image $L(x,y,\sigma)$ at the key-point's scale σ are computed as follows:

$$m_g(x,y)=|L(x,y,\sigma)|=\sqrt{\left(\frac{\partial L(x,y,\sigma)}{\partial x}\right)^2+\left(\frac{\partial L(x,y,\sigma)}{\partial y}\right)^2} \quad (14)$$

$$\theta_d=\angle L(x,y,\sigma)=\tan^{-1}\left(\frac{\partial L(x,y,\sigma)/\partial y}{\partial L(x,y,\sigma)/\partial x}\right) \quad (15)$$

Second, an orientation histogram is described in the following:

Histogram; from the gradient orientations of the sample points around a key-point, an orientation histogram is formed.

Quantisation; the orientation histogram is quantised into an even-sized bin N_b to reduce the data that has to be processed. A higher value of bins results in a higher resolution, but also a higher computation load, [3] suggests N_b= 36.

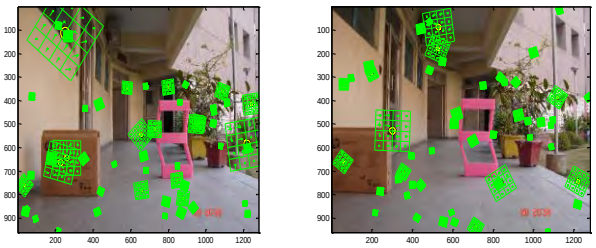
Weighted; each sample is weighted by a gradient magnitude and a Gaussian-weighted circular kernel that is placed on the centre of the sample point, with σ,1.5times the scale of the interest point. This has the effect of giving higher weight to the samples near the centre of the window.

Creation of new interest points; the peaks in the orientation histogram correspond to dominant directions of the local gradients surrounding the interest point. The highest peak in the histogram is detected and if there is any other peak which falls within 80% of the highest peak, a new interest point is created with that orientation. So, there will be multiple key-points with the same location and scale, but different orientations. To obtain a better precision, a parabola is fitted to the three histogram values closest to each peak, in order to interpolate the peak position.

5) Computation of the local image descriptor: Previous steps found key point locations at particular scales and assigned orientations to them. This ensured invariance to image location, scale and rotation. The next step is to compute descriptor vectors for these key-points such that the descriptors are highly distinctive and partially invariant to the remaining variations, like illumination, 3D viewpoint, etc. This step is pretty similar to the orientation assignment step.

An interest point descriptor is created by first sampling the gradients magnitudes $m_g(x,y)$ and orientation $\theta_d(x,y)$ in the surroundings of the key-point location. The gradient magnitude and orientation are computed using –(14)"and –(15)". Then, a rotation matrix is applied to the descriptor in order to get orientation invariance. Afterwards, the key-point descriptor is weighted and boundary effects are reduced. Finally the SIFT-descriptor is created. The effect of change in image contrast and nonlinear illumination can affect the SIFT-key-point descriptor.

A SIFT-descriptor vector is defined as a vector with 128 elements that uniquely identifies a key-point. The key-point descriptor is arranged into 16 histograms. Each histogram shows eight direction bins. Each bin in each histogram is formed by adding up the gradient magnitudes in the current direction. The value of the orientation histogram constitutes the 128-dimentional vector (8 orientations X 16 histograms). The 128 dimensional vector is the SIFT feature vector or SIFT descriptor as shown in –Fig. (9)".



SIFT- discriptor for left image SIFT- discriptor for right image

Figure 9. SIFT- descriptor

B. *SIFT-Descriptor Matching*

Once the descriptors from the left and right images are found, the next step is to find the best candidate match for each key-point in the left image which corresponds with the right image.

[3] Suggests a method for comparing the descriptors. The method consists of looking at the relation between the two shortest Euclidean distances, as seen in –(16)".

If the ratio (r_{tt}) between the second shortest (d_{scn}) and the shortest (d_{cn}) Euclidean distances is close to 1, it implies that two key-points in the reference set match the image key point equally well. This ambiguous match significantly reduces the probability that the match is correct.

This ratio is use to get rid of incorrect matches. Setting a proper threshold (t_s) for this ratio does it, thus all matches above this threshold are considered false matches (f_m) and are rejected, and all matches below this threshold are considered correct matches (cm) and are accepted. That threshold has been determined empirically to 0.8. It is, of course, inevitable that this procedure will discard some of the correct matches, as it is stated in –(17)".

$$r_{tt} = \frac{d_{cn}}{d_{scn}} \tag{16}$$

Correct matches if $r_{tt} < t_s$

False matches if

$$r_{tt} \geq t_s \tag{17}$$

The Euclidean distance d between two descriptors is computed as stated in –(18)"

$$d = \sqrt{\sum_{i=1}^{128} (p_{1_i} - p_{2_i})^2} \tag{18}$$

C. *RANSAC (RANdom SAmple Consensus)*

The RANSAC algorithm is proposed by Fischler and Bolles [5] is a general parameter estimation approach designed to cope with large proportion of outliers in the input data. It is a resampling technique that generates candidate solutions by using minimum number of observations (data points) required to estimate the underlying model parameters.

After the feature matching pairs are determined, we apply RANSAC to estimate the homography. The RANSAC algorithm is being used to perform this computation under the possible presence of outlier feature matches. –Fig. 10" shows the results of matching interest points in an image stereo pair, tentative matches are the total matches determined by matching SIFT descriptor and inlier matches are the more accurate matches generated after applying RANSAC.

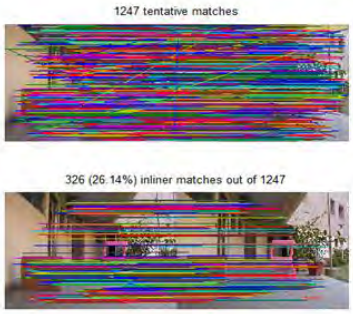


Figure 10. Descriptor matches are connected by coloured lines

D. *Optimization of Threshold Ratio*

Selection of threshold ratio for matching the descriptors is an important step while applying SIFT algorithm. Because if the selection of the threshold ratio is less than a certain limit then there is a large number of tentative matches found and that will lead to increase in the possibility of wrong matches. And if the selection of threshold ratio is more than certain limit then the number of tentative matches are very less. So there is a possibility to miss some objects in the environment.

TABLE II. VARIATION IN MATCHING RESULTS WITH THRESHOLD RATIO

Threshold value	Tentative matches	In liner matches	% of In liner matches
0.10	1247	326	26.14
0.20	972	295	30.35
0.30	716	235	32.82
0.40	551	196	35.57
0.50	460	172	37.39
0.60	390	160	41.03
0.70	355	149	41.69
0.80	323	147	45.51
0.85	308	145	47.07
0.90	298	135	45.30
0.95	106	21	20.1

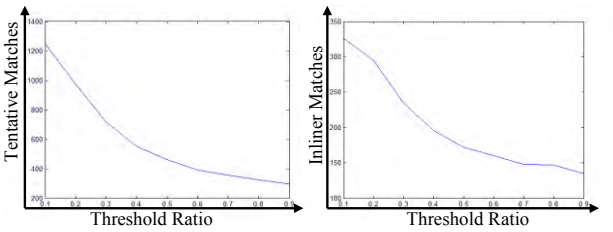
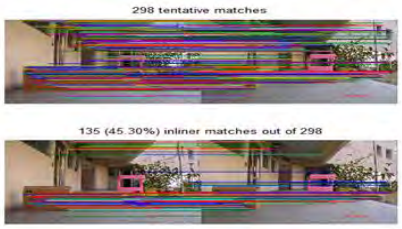


Figure 11. Variation of tentative matches and inlier matches with t_s

12 a. Matching result at $t_s = 0.7$

12 b. Matching result at $t_s = 0.8$



12 c. matching result at $t_s = 0.9$

Figure 12. Descriptor matching with different t_s

For, selection of optimal value of threshold ratio we applied RANSAC after applying SIFT with varying the value of threshold ratio. After this experiment we found that the number of tentative matches given by SIFT decreases and percentage of inlier matches given by RANSAC increases by increasing the value of threshold ratio up to $t_s = 8.5$ but when we further increased the value of t_s, the percentage of inlier matches decreased rapidly. Some results of tentative matches and inlier matches are shown in "Fig.12", and complete experimental results are given in "Table II". Variation of tentative matches and inlier matches are shown in "Fig.11".

By this experiment we conclude that optimal value of threshold ratio t_s can be selected by optimization of percentage of inlier matches.

E. *Distance measurement*

Distance measurement method used in this work was suggested by Hai-Sung Baek [6]. The whole stereo vision system area consists of three areas: area *A* between the two optical axes, area *B* outside the two optical axes, and a block section to be seen by only one camera which consist of area *RC* which is seen by the right camera only and area *LC* which is seen by the left camera only. Other proposed ways to measure distance to an object by SVS is only applicable if object is in area *A*.

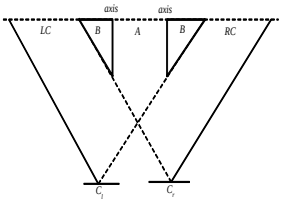


Figure 13. Stereo Vision Systems area

Distance measurement method proposed in [6] can be used when the object is located even in the outside area of the two optical axis as long as the object is in overlapping area of the two cameras. The "Fig. 14" is picturing that object expresses case in outside area of optical axis P_l and P_r of "Fig. 14"amounts to negative number and D_1, D_2 is calculated in "(19)"and "(20)".

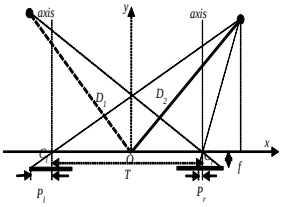


Figure 14. Object beyond the optical axis of right and left camera

$$D_1 = \frac{T}{2(P_l - P_r)}\sqrt{(P_l + P_r)^2 + 4f^2} \quad (19)$$

$$D_2 = \frac{T}{2(P_r - P_l)}\sqrt{(P_l + P_r)^2 + 4f^2} \quad (20)$$

F. *Simulation Results*

The simulation of these systems is all performed in MATLAB and SIMULINK. To find the correspondence relation between left and right images SIFT and RANSAC has been implemented by using VLfeat toolbox in MATLAB developed by Vedalid at UCLA [7]. This gives the disparity value in a MAT file and the MAT file is being set as the input to the Simulink programme as suggested by Hai-Sung Baek [6] for the distance measurement. Output of this Simulink programme has been saved in workspace that gives the distance of each point whose descriptors matches. And finally we compared the distance given by the system for the point around each objects with actual measured distance of the objects. ―Table III" gives the specific experimental data.

TABLE III. MEASURED DISTANCES BY USING SIFT

The actual distance (m)	The measured distance (m)	The error (%)
1.9	1.956	2.94
2.8	2.91	3.92
3.12	3.07	1.60
3.5	3.42	2.28
3.62	3.73	3.03

VII. SIFT FEATURE MODEL

A. *Experimental setup*

The experiment is performed for the mapping of an indoor environment using SIFT feature descriptor model. Four objects have been placed on the path that mobile robot has to follow. Stereo pair of images has been taken by using single CCD camera, Sony cyber shot (DSC-S650). The camera is placed on a slider at two different position separated by a distance of 5cm keeping the axis of the cameras parallel. The ―Fig. 15" shows the geometric map of the experimental setup. Whole environment has been mapped by using an occupancy grid of 100x100 cells, each cell having size of 8cmx8cm.

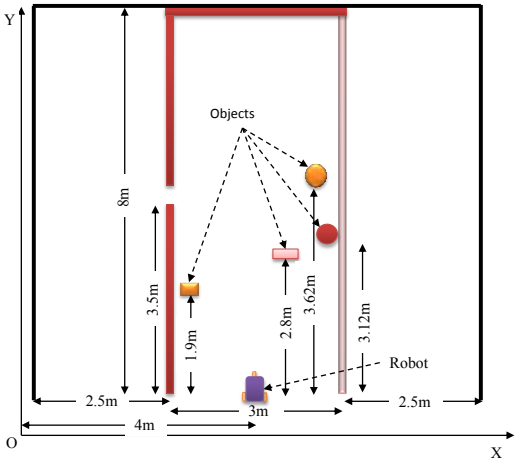


Figure 15. Geometric map of the dynamic environment

Position of robot from where stereo pair of images has been taken is (50, 1) and the positions objects are (34, 23), (66, 35), (69, 39) and (68, 45).

B. *SIFT- Feature Model*

One of the tasks of a stereo vision system in mobile robot is the extraction of features from a pair of digital images. These features can be used for map making. Feature based mapping has been implemented in the past by researchers. [1] For instance used the canny edge detector to extract edges from the environment, matching edges in the pair of digital images that are used to construct the map. [8] Suggests the use of visual techniques such as the SIFT algorithm to identify features common to each camera view, and thereby infer the geometry.

Sensor readings are uncertain by nature and the stereo vision system is not the exception. In this chapter a probabilistic sensor model is proposed. The model takes into account the uncertainties inherited by the stereo vision system readings. The model is divided in two areas; the occupied area which is in fact the model of the quantification error and the modelling of empty area. Combination of the two models into one gives the final SIFT-descriptor model.

C. *Modelling the Occupied Area*

Stereo triangulation is needed in order to get the depth from the stereo system to the features extracted from the SIFT. And, due to the factors of quantification and calibration errors, a certain degree of uncertainty must be expected in the triangulation. This uncertainty must be modeled. In [1], Mathies and Shafer show three approaches to model such uncertainty. They are: discrete tolerance limits, scalar weights and multidimensional probability distribution.

The multidimensional probability distribution will be explained in the following. The geometry of stereo triangulation is shown in ―Fig.16(a)". The trick marks on the image planes denote pixel boundaries as well as the lines which are radiating to the space from this tick marks.

Assuming a point M in the space which is projected onto the left and right image planes at x_l and x_r respectively. Because of errors in quantification, the stereo system will determine x_l and x_r with some error, which in turn causes error in the estimated location of M. Thus causing the point M lying on the region surrounds the true location. It can also be seen in "Fig. 16(a)" that the uncertainty is skewed and orientated (diamond around the point M). This uncertainty can be captured by using 3D Gaussian distributions. This distribution can be depicted as in "Fig. 16(b)",where the ellipse represents the contour of the error model and the diamond represents the quantization error.

The mean values and the covariance matrix of each point in the space are needed in order to get a proper shape of the Gaussian distribution. For that, the triangulation error for the general case of 3D points projecting onto 2D images is described in the following.

Consider 3D point M=(X, Y, Z) or a vector= $[X,Y,Z]^T$, which is projected onto the left and right image planes respectively as $m_l=[x_i,y_l]^T$ and $m_r=[x_r,y_r]^T$ as depicted in "Fig. 17".

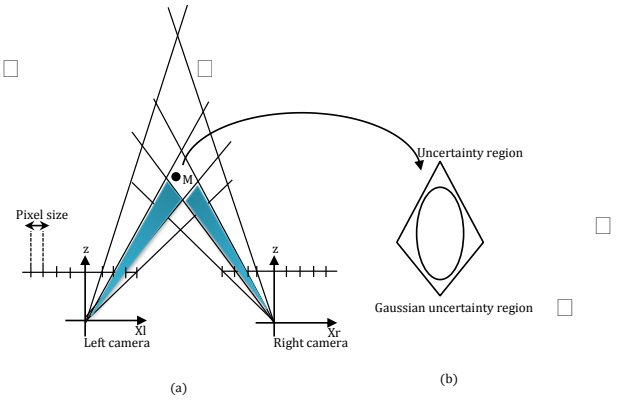


Figure 16. (a) Stereo Geometry showing triangulation uncertainty as a diamond around a point *M*. It also shows the empty region uncertainty from the pair of cameras to the uncertainty region of the point *M*.

(b) 2D Gaussian distribution uncertainty region.

The vectors m_l and m_r are considered to be normally distributed with means μ_l and μ_r and covariance matrices V_l and V_r [1]. This mean and the covariance matrix of the point *M* are μ_M and V_M. μ_M nd *M* are functions of m_l and m_r meaning that $\mu_M=f(x_l,y_l,x_r,y_r)$ and $M=[X,Y,Z]^T=f(x_l,y_l,x_r,y_r)$ as stated in [1]. Equation 2.6 and 2.7 from the triangulation process explained in chapter2 can be used to estimate the coordinates of the vectors $M=[X,Y,Z]^T$. These equations need to be changed in order to use them in the estimation of the point *M*. these equations deviate from the one used in this section in that, MATLAB calibration tool box uses the right camera as a reference frame instead of the left reference stated in equation "(21)" and "(22)".

$$s_M = am_r + \tfrac{1}{2}cn \tag{21}$$

$$am_r - bR^T m_l + c(m_r \times R^T m_l) = T$$

$$\left[m_r - R^T m_l \quad m_l + c(m_r \times R^T m_l)\right]\begin{bmatrix} a \\ b \\ c \end{bmatrix} = T \tag{22}$$

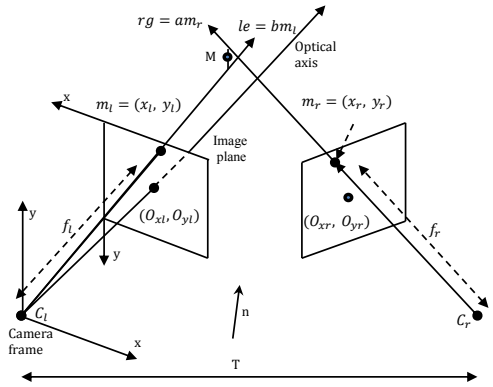


Figure 17. Figure shows in the left part the perspective or pinhole model and also the triangulation.

The vectors s_M in "(21)" defines the coordinates of the point *M* as seen in "(23)", where $n=m_r\times R^T m_l$.

$$M=[X,Y,Z]^T = am_r + \tfrac{1}{2}c(m_r \times R^T m_l) \tag{23}$$

Making the following definitions $m_l(x_l)=m_{lx}, m_l(y_l)=m_{ly}, m_r(x_r)=m_{rx}$ and $m_r(y_r)=m_{ry}$. The vectors m_l and m_r can be defined as vectors with 3 components each one such as $[m_{lx},m_{ly},f_l]^T$ and $[m_{rx},m_{ry},f_r]^T$ respectively, where f_l and f_r are the focal length of the left and right cameras respectively. $m_{lx}=(O_{xl}-x_l)S_x, m_{ly}=(O_{yl}-y_l)S_y$, $m_{rx}=(O_{xr}-x_r)S_x$, and $m_{lx}=(O_{yr}-y_r)S_y$, where S_x is the pixel size in mm in horizontal and vertical direction respectively. O_{xl}, O_{yl}, O_{xr}, and O_{yr} are the coordinates in pixels of the left and right image centres (The principal point). R is a (3x3) rotation matrix.

A proper substitution of the definition mentioned above into "(23)".

$$[X,Y,Z]^T = a[m_{rx},m_{ry},f_r]^T + \tfrac{1}{2}c\left([m_{rx},m_{ry},f_r]^T \times R^T[m_{lx},m_{ly},f_l]^T\right) \tag{24}$$

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = a\begin{bmatrix} m_{rx} \\ m_{ry} \\ f_r \end{bmatrix} + \tfrac{1}{2}c\begin{bmatrix} m_{ry}R^T f_l - R^T m_{ly} f_r \\ m_{rx}R^T f_l - R^T m_{lx} f_r \\ m_{rx}R^T m_{ly} - R^T m_{lx} m_{ry} \end{bmatrix}$$

$$X = a(O_{xr}-y_r)S_x + \tfrac{1}{2}c\left[(O_{yr}-y_r)S_y R^T f_l + R^T(O_{yl}-y_l)S_y f_r\right]$$
$$= f_1(x_l,y_l,x_r,y_r)$$

$$Y = a(O_{yr} - y_r)S_y + \frac{1}{2}c\left[-(O_{xr} - x_r)S_x R^T f_l + R^T(O_{xl} - y_l)S_x f_r\right]$$
$$= f_2(x_l, y_l, x_r, y_r)$$
$$Z = af_r + \frac{1}{2}c\left[(O_{xr} - x_r)S_x R^T(O_{yl} - y_l)S_y - R^T(O_{xl} - x_l)S_x(O_{yr} - y_r)S_y\right]$$
$$= f_3(x_l, y_l, x_r, y_r) \quad (25)$$

The parameters a and c in equations stated in "(25)" can be obtained by solving the linear system as seen in "(22)" where a, b and c are scalars. This linear system is defined as follows $(A_x = b_l)$ where $A = \left[m_r, -R^T m_r, m_l \times R^T m_r\right]$, $x = [a,b,c]^T$ and $b_l = T$.

The mean value μ_M and the covariance matrix V_M of the point M can be computed as follows. V_M is calculated as stated in [1]as seen "(26)",

$$V_M = J\begin{bmatrix} V_l & 0 \\ 0 & V_r \end{bmatrix}J^T \quad (26)$$

Where J is the Jacobian of first partial derivatives of $f_{1n}(x_l, y_l, x_r, y_r)$ with n= 1, 2, 3 as seen in expression "(25)". The jacobian, which is a matrix of all first-order partial derivatives of a vector-valued function, can be computed as in "(27)".

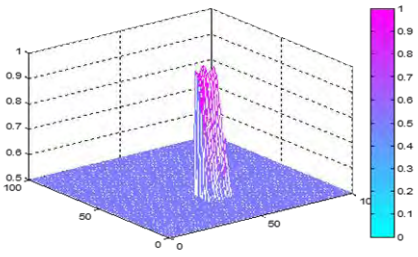
$$J = \begin{bmatrix} \frac{\partial f_1}{\partial x_l} & \frac{\partial f_1}{\partial y_l} & \frac{\partial f_1}{\partial x_r} & \frac{\partial f_1}{\partial y_r} \\ \frac{\partial f_2}{\partial x_l} & \frac{\partial f_2}{\partial y_l} & \frac{\partial f_2}{\partial x_r} & \frac{\partial f_2}{\partial y_r} \\ \frac{\partial f_3}{\partial x_l} & \frac{\partial f_3}{\partial y_l} & \frac{\partial f_3}{\partial x_r} & \frac{\partial f_3}{\partial y_r} \end{bmatrix} \quad (27)$$

The Jacobian matrix stated in "(27)" can be solved by means of numerical analysis, ans an example of such numerical approach is shown in "(28)"

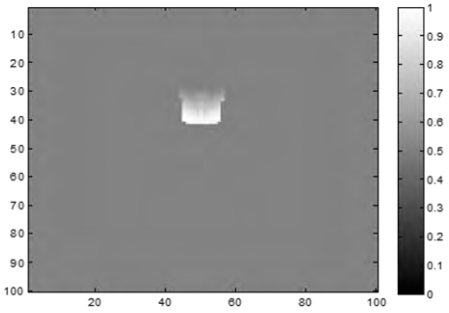
$$\frac{\partial f}{\partial x_1} \approx \frac{1}{2}\left(f(x_l - 1, y_l, x_r, y_r) - f(x_l + 1, y_l, x_r, y_r)\right) \quad (28)$$

Mathies and Shafer [1] suggest to approximate the means $(\mu_l \text{ and } \mu_r)$ with the coordinates returned by the stereo matcher and the covariance with the identity matrices. This is equivalent to treating the image coordinates as uncorrelated with variances of one pixel.

"Fig. 18 (a)" shows a three dimensional probabilistic model of the area by the SIFT – descriptor. The colour- bar indicates the lowest and the highest probabilities of being occupied. "Fig. 18 (b)" is a two dimensional grey-scale representation of the occupied area. The 0 in the grey-scale bar indicates the lowest probability of occupancy, e.g. $P^o_{i,j} = 0$. Whereas the 1 indicates the highest probability of occupancy, e.g. $P^o_{i,j} = 1$.



18 (a)



18 (b)

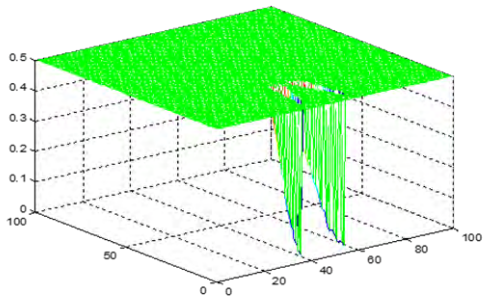
Figure 18. (a) is a 3D representation of the occupied area by the SIFT-descriptor.

(b) Is a 2D representation of the occupied area by the SIFT- descriptor.

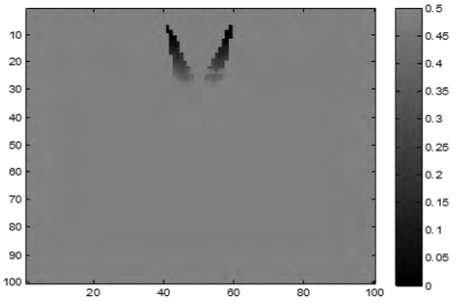
D. *Modelling the Empty Area*

The empty regions from the left and right cameras, as shown as shadow areas in "Fig. 16 (a)", also need to be modelled. A search concerning the modelling of the uncertainties of these two regions has been carried out in the literature with unsuccessful results.

The approach taken by Elfes [2] to model the empty region of the sonar beam has been taken into consideration to solve the problem of modelling these two empty regions. The implementation came up with satisfactory results as depicted in "Fig. 19". "Fig.19 (a)"shows the three dimensional probabilistic model of the two empty areas of the SIFT-descriptor, whereas "Fig. 19 (b)" shows a two dimensional grey scale representation of the empty area, the more black the region is the more probability of emptiness.



19 (a)



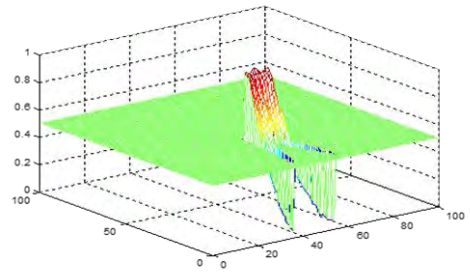
19 (b)

Figure 19. (a) A 3D probabilistic representation of the empty area of the SIFT- descriptor.

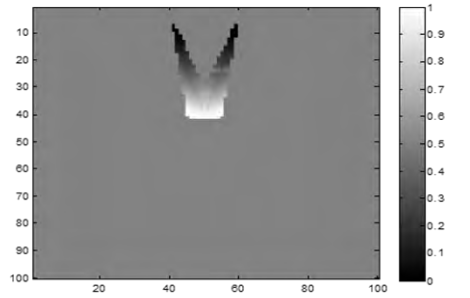
(b) A 2D grey- scale representation of the empty area of the SIFT-descriptor.

E. *Experimental results*

The occupied area as well as the empty one of the SIFT-descriptor model is put together bringing up the final probabilistic SIFT- descriptor model. "Fig. 20 (a)" shows a 3D model of the uncertainty triangulation together with the empty uncertainty region of the empty areas, which in fact is the 3D probabilistic model of the SIFT descriptor. "Fig. 20 (b)" shows a 2D view of the former 3D view of the uncertainties.



20 (a)



20 (b)

Figure 20. The SIFT- descriptor probability model. (a) 3D view of the SIFT- descriptor probability model. (b) 2D view of the SIFT- descriptor probability model.

VIII. INTEGRATING THE OCCUPANCY GRID

A. *Recursive Bayes Update Rule*

The attraction of the Bayesian inference approach to map building stems from the fact that Bayes updating rule is recursive. When it is used to support sensor fusion, Bayes rule provides a way of computing a posteriori probability of a hypothesis being true giving supporting of evidence. [9], [10], [1], and [11] have successfully used Bayes rule to update the occupancy grid for multiple sensor readings $(s_1, \ldots, s_n)$.

"Equation (29)" and "(30)" are obtained when Bayes rule is transferred to the occupancy grid framework for multiple sensor readings.

$$P_{i,j}^{o/s} = \frac{P_{i,j}^{s/o} P_{i,j}^{o}}{P_{i,j}^{s/o} P_{i,j}^{o} + (1 - P_{i,j}^{s/o})(1 - P_{i,j}^{o})} \quad (29)$$

$$P_{i,j}^{e/s} = \frac{P_{i,j}^{s/e} P_{i,j}^{e}}{P_{i,j}^{s/e} P_{i,j}^{e} + (1 - P_{i,j}^{s/e})(1 - P_{i,j}^{e})} \quad (30)$$

The following statements are defined.

- The relevent evidence A is given by the sensor reading s.
- The certainty of the true parameter B_i is given by $P_{i,j}^{o}$ and $P_{i,j}^{e}$, meaning that they are the prior probabilities of the cell $C_{i,j}$ being occupied or empty. They are taken from the existing map
- The conditional probability $P(B_i|A)$ is given by $P_{i,j}^{s|o}$ and $P_{i,j}^{s|e}$, which are the conditional probabilities that a sensor reading will exist given the state of the cell $C_{i,j}$, being occupied or empty. This conditional probability is given by the probabilistic sensor model.
- The condotional probability $P(B_i|A)$ is given by $P_{i,j}^{o|s}$ and $P_{i,j}^{e|s}$, which is the conditional probability that a cell is occupied based on the past sensor readings. It is the new estimate.

A new sensor reading, introduces additional informational information about the state of the cell $C_{i,j}$. This information is done by the sensor model $P_{i,j}^{s|o}$, and it is combined with most recent probability estimate stored in the cell. This combination is done by the recursive Bayes'rule ($P_{i,j}^{o|s}$) based on the current set of readings to give a new estimate $P_{i,j}^{o|s}$. It is worth nothing that when initialising the map an equal probability to each call $C_{i,j}$ must be assigned. In other words, the initial map cell prior probabilities are $P_{i,j}^{o} = P_{i,j}^{e} = \frac{1}{2}$.

A sensor reading s in "(29)"and "(30)" changes to r when there is a sensor reading and to v when there is a vision reading, e.g. $s \to r$ and $s \to v$.

The graphical interpretation of "(29)" and "(30)" for a signal cell $C_{i,j}$ within a sonar model can be depicted as in "Fig. 21". In order to update the cell $C_{i,j}$, with equations "(29)" and "(30)", a prior probabilities $P_{i,j}^{0_r}$ and $P_{i,j}^{e_r}$ given by the existing grid are needed. These probabilities are combined with the probabilities given by the sonar model ($P_{i,j}^{r|o}$ and $P_{i,j}^{r|e}$) by means of Bayes'rule of combination to give a new estimate or posteriori probabilities ($P_{i,j}^{o|r}$ and $P_{i,j}^{e|r}$)

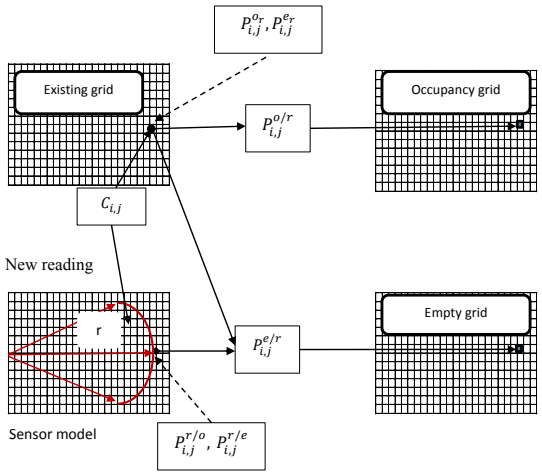


Figure 21. A graphical interpretation of equation 4.1 and 4.2 for a single cell $C_{i,j}$ within the main lobe of the sonar model. An existing grid contains the old probabilities (prior) of the sonar model $P_{i,j}^{o_r}$ and $P_{i,j}^{e_r}$ of a single cell $C_{i,j}$ being occupied or empty respectively. A new sensor data interpreted by a sonar model ($P_{i,j}^{r|o}, P_{i,j}^{r|e}$) is used together with the existing probabilities in the grid to estimate the new state of the cell ($P_{i,j}^{o|r}, P_{i,j}^{e|r}$).

B. *Experimental results*

The occupancy grid obtained from the SIFT descriptor model is updated by using Recursive Bays Rule. For integration of the occupancy grid obtained from the SIFT descriptor model we have used the occupancy grid of the prior information about the static environment as an existing grid. The occupancy grid obtained from the SIFT descriptor model is updated with existing occupancy grid by using Recursive Bays rule. The integrated occupancy grid map of the indoor environment is shown in "Fig. 22".

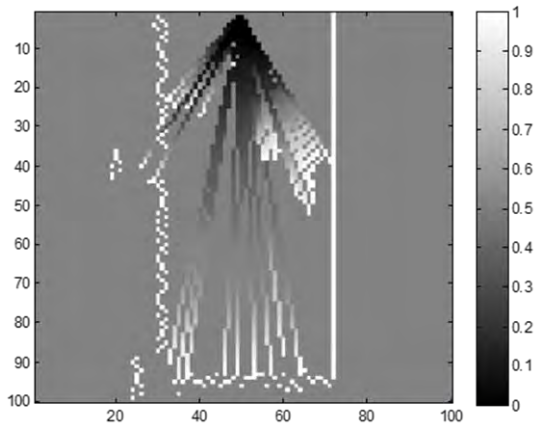


Figure 22. Integrated occupancy grid map of the environment

White cells represents occupied cell and the scale on the side bar shows the probabilistic value of occupancy according to the color of a particular cell. The more the color is black of a cell the less is the occupancy of the cell.

IX. CONCLUSION AND FUTURE SCOPE OF WORK

A. *Conclusions*

In this chapter, we have studied the distance measurement system by using Scale Invariant Feature Transform. Analysed the key algorithm, brought out the image match method based on SIFT matching and introduced an empirical technique to optimize the system performance. And implemented the discussed algorithm for a distance measurement method to measure the distance of an object which is located outside the axes of stereo cameras, obtained the ideal results. In the following system, distance to an object in any location of overlapping area of stereo cameras can be measured. It is expected that the system will be widely applied in engineering measurement and automatic control.

This chapter has also proposed a SIFT-descriptor probabilistic sensor model. The model takes into account the uncertainty inherited in the stereo vision system when computing the SIFT-descriptors. To this end, the model has been divided in two regions; occupied and empty areas. The occupied area was modeled according to [1], where the multidimensional probability distribution approach is used to model the quantification error. The empty region was modelled according to the approach proposed in [1] for modelling the empty area of the sonar beam. The modeling came up with satisfactory results. The model has been used for internal representation of sensor fusion tasks. The occupancy grid obtained from SIFT-descriptor model is integrated by using Recursive Bays rule with occupancy grid of the prior information about the static environment obtained by using sonar sensor. This reduces the computational burden of the system for mapping the indoor environment by providing static information about the environment that cannot be changed. This system can be used for the multi sensor data fusion tasks.

In this work, we have studied different sensors and selected vision sensors for mapping the environment because of its high accuracy for distant objects, sonar sensors for acquisition of prior information about the environment. Analysed the key algorithm, brought out the image match method based on SIFT matching and introduced an empirical technique to optimize the system performance. The number of tentative matches given by SIFT, number of inliner matches given by RANSAC and percentage of inliner matches are given in "Table II". By using this experiment we have selected 0.85 as a threshold ratio where the percentage of inliner matches is highest. It is found that percentage of accurate matches in inliner matches given by RANSAC is 89% which much better than other type of sensor like Sonar Sensor.

We improved the result for distance measurement by using a new method [6]. Using this method an object in any location of overlapping area of stereo cameras can be measured after

implementing the discussed SIFT algorithm. The improved results are given in table III.

The occupancy grid obtained from SIFT-descriptor model is shown in "Fig. 20" and integrated by using Recursive Bays rule with occupancy grid of the prior information about the static environment obtained by using sonar sensor. The modelling came up with satisfactory results for occupancy grid, almost similar to the original environment as shown in "Fig. 22".

B. Future Scope of the Work

In this work, we have pre assumed while occupancy grid mapping according to prior information about the environment that the area lying at some depth from the level of the floor of mobile robot are occupied. Further research can be done for modelling the region lying at some depth from the level of the floor of mobile robot by using stereo vision camera or any other sensors.

From results of sensor fusion it is clear that only vision sensor cannot detect objects accurately so for increasing accuracy of the system, combination of same sensor as well as combination of different sensor may be used.

The dissertation presented the Bayesian theorems for occupancy grid mapping. Experiments with fuzzy logic algorithms, Dempster-Shafer grid and Neural Network algorithms are a possible and interesting research area to work further and may be getting even better estimations.

ACKNOWLEDGMENT

I express my gratitude to Former Director of NIT Jalandhar Dr. Moin Uddin and Dr. Dilbag Singh Professor in the department of Instrumentation and Control engineering, NIT Jalandhar, for the facilities and support provided in the Robotics Research Lab.

REFERENCES

[1] A.Elfes. Sonar-based real-world mapping and navigation. IEEE Journal on Robotics and Automatation, 3:249-265, June 1987.

[2] D.G. Lowe. Object recognition from local scale-invariant features. In Proceedings of the International Conference on Computer Vision ICCV, Corfu, pages 1150-1157, 1999.

[3] D.G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91-110, 2004.

[4] M. Brown and D.G. Lowe. Invariant features from interest point groups. In Proceedings of BMVC, Caerdiff, Wales, pages 656-665, 2002.

[5] M.A. Fischler and R.C. Bolles. Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM, 24(6):381-395, 1981.

[6] Hai-Sung Baek and Jung-Min Choi. Improvement of didtance measurement algorithm on stereo vision system (SVS).

[7] A. Vedaldi and B. Fulkerson. VLFeat library. http://www.vlfeat.org/, 2008.

[8] A. Dankers, N. Barnes, and A. Zelinsky. Active vision-rectification and depth mapping. In Proceedings of the 2004 Australasian Conference on Robotics and Automation, 2004.

[9] A.Elfes. Atesselated probabilistic representation for spatial perception and navigation. In Proceedings of the NASA Conference on Space Telerobotics, volume N90, pages 341-350, 1989.

[10] A.Elfes and L.Matthies. Integration of sonar and stereo range data using a grid-based representation. In Proceedings of the IEEE International Conference on Robotics and Automation, pages 727-733, 1988.

[11] A.Elfes. Using occupancy grids for mobile robot perception and navigation. IEEE Computer, 22 (6), 1989, pp. 46-57.

[12] Hans P. Moravec, Alberto Elfes, "High Resolution Maps from Wide Angle Sonar,"Proc. IEEE Int'l Conf. on Robotics and Automation, CS Press, Los Alamitos, Calif, March 1985, pp. 116-121.

