Reduced SIFT Features For Image Retrieval and Indoor Localisation

Luke Ledwich, Stefan Williams
ARC Centre of Excellence for Autonomous Systems
School of Aerospace Mechanical and Mechatronic Engineering
University of Sydney, NSW, 2006, Australia
lledwich@mail.usyd.edu.au, stefanw@acfr.usyd.edu.au

Abstract

SIFT features are distinctive invariant features used to robustly describe and match digital image content between different views of a scene. While invariant to scale and rotation, and robust to other image transforms, the SIFT feature description of an image is typically large and slow to compute. This paper presents a method to reduce the size, complexity and matching time of SIFT feature sets for use in indoor image retrieval and robot localisation. Our method takes advantage of the structure of typical indoor environments to reduce the complexity of each SIFT feature and the number of SIFT features required to describe a scene. Our results show that there is a minimal loss of accuracy in feature retrieval while achieving a significant reduction in image descriptor size and matching time. We also outline how the scale information of the SIFT features can be used to improve the accuracy of a localisation filter. The results were obtained using digital images from interior home and office environments.

1 Introduction and Related Work

Matching images based on visual content is a fundamental problem in computer vision. The image matching problem occurs in many computer vision systems from a variety of fields including image retrieval and robot localisation. Content Based Image Retrieval (CBIR) addresses the matching and retrieval of images which share similar visual content to a search concept from a large database of unannotated images [Goodrum, 2000] [Eakins and Graham, 1999] [Koskela *et al* , 2001]. The most common CBIR methods use combinations of primitive features such as colour, texture and structure to describe an image. These image descriptors are used with a similarity measure to retrieve images that are alike [Iqbal and Aggarwal, 1999] [Liu and Picard, 1996] [Mirmehdi and Perissamy, 2001]. More advanced systems retrieve images by statistically attaching linguistic indexes and retrieving by index association [Li and Wang, 2003] or using learned mappings in feature space to group similar images [Laaksonen *et al* , 2000]. The techniques used in CBIR have successfully been applied to robot localisation in both topological [Ulrich and Nourbakhsh, 2000] and metric applications [Wolf *et al* , 2002].

Robot localisation applications have more stringent matching requirements than traditional CBIR applications because they rely on retrieving only images which share a common view. Indoor environments pose the additional problem of perceptual aliasing where different locations are visually similar. This makes discrimination and accuracy important characteristics for image features in indoor localisation applications. A common approach to accurate image matching is known as 'Keypoint' or 'Interest point' extraction. It involves identifying points that can be reliably extracted from different images of the same scene. Effective keypoint extraction has been achieved with the addition of invariant features [Schmid and Mohr, 1997].

Invariant features are features that do not change when exposed to a set of image transformations. Earlier research into invariant features focused on invariance to rotation and translation [Siggelkow, 2002] [Schulz-Mirbach, 1995]. These methods have achieved relative success with 2D object extraction and image matching. Later work in invariant features has focused on expanding their invariance to illumination, scale and affine transforms. There has been research into the development of fully invariant features [Brown and Lowe, 2002] [Mikolajczyk and Schmid, 2001]. However full affine invariance has not been achieved due partly to the impractically large computational cost.

While complete invariance has yet to be achieved, features which are robustly resilient to most image transforms have been proposed by [Lowe, 2004]. Scale Invariant Feature Transforms (SIFT) are invariant to ro-

tation, translation and scale variation between images and partially invariant to affine distortion, illumination variance and noise. These features have been applied to object recognition [Lowe, 1999], topological localisation [Koescka and Li, 2004] and SLAM [Se *et al* , 2002]. However a significant drawback with the SIFT features is the significant amount of data generated and the computational cost involved.

In this paper we propose a reduction to the SIFT features taking advantage of the structure of the indoor environment. The goal of this reduction is to make them more efficient and practical in the context of image retrieval and indoor robot localisation without a significant reduction of accuracy and discrimination. In section 2 the method used to generate SIFT features is described and in section 3 the reduction modifications are detailed. Section 4 covers the results of the reduced SIFT features for image retrieval and sections 5 and 6 describe their application to indoor robot localisation. Section 6 also details an improvement to localisation by using the scale space information from matched SIFT features. Finally section 7 summarises the results of the reduced SIFT features and details further work on a more compact SIFT feature set.

2 SIFT Features

SIFT features were proposed in [Lowe, 2004] as a method of extracting and describing keypoints which are robustly invariant to common image transforms. The Scale Invariant Feature Transform (SIFT) algorithm has 4 major stages.

- *Scale-space extrema detection:* The first stage searches over scale space using a Difference of Gaussian function to identify potential interest points.
- *Keypoint localisation:* The location and scale of each candidate point is determined and keypoints are selected based on measures of stability.
- *Orientation assignment:* One or more orientations are assigned to each keypoint based on local image gradients.
- *Keypoint descriptor:* A descriptor is generated for each keypoint from local image gradients information at the scale found in stage 2.

An important aspect of the algorithm is that it generates a large number of features over a broad range of scales and locations. The number of features generated is dependent on image size and content, as well as algorithm parameters. A typical image of 500x500 pixels will generate approximately 2000 features however in our indoor examples a similar size image will typically only generate 300 features.

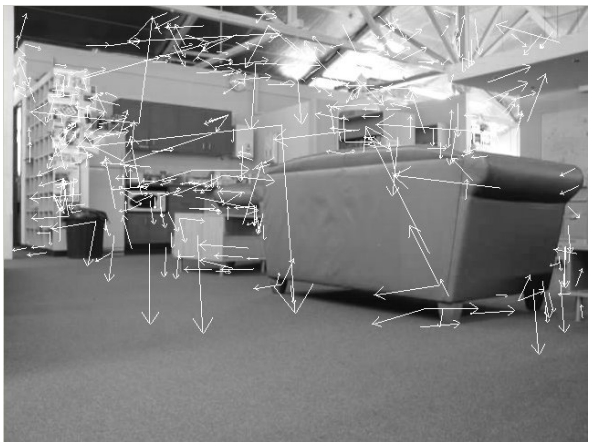


Figure 1: Typical indoor office environment and the extracted SIFT features with their locations represented by arrows. The length of the arrow represent the scale of the extracted keypoint and the direction represents the orientation of the descriptor.

The SIFT feature algorithm is based upon finding locations within the scale space of an image which can be reliably extracted. The first stage finds scale-space extrema located in $D(x, y, \theta)$, the Difference of Gaussians (DOG) function, which can be computed from the difference of two nearby scaled images separated by a multiplicative factor k:

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

where $L(x, y, \sigma)$ is the scale space of an image, built by convolving the image $I(x, y)$ with the Gaussian kernel $G(x, y, \sigma)$. Points in the DOG function which are local extrema in their own scale and one scale above and below are extracted as keypoints. Generation of extrema in this stage is dependent on the frequency of sampling in the scale space k and the initial smoothing σ_0. The keypoints are then filtered for more stable matches, and more accurately localised to scale and subpixel image location using methods described in [Brown and Lowe, 2002].

Before a descriptor for the keypoint is constructed, the keypoint is assigned an orientation to make the descriptor invariant to rotation. This keypoint orientation is calculated from an orientation histogram of local gradients from the closest smoothed image $L(x, y, \sigma)$. For each image sample $L(x, y)$ at this scale, the gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ is computed using pixel differences:

$$m(x,y) = ((L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2)^{\frac{1}{2}} \quad (2)$$

$$\theta(x,y) = tan^{-1}\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)} \quad (3)$$

The orientation histogram has 36 bins covering the 360 degree range of orientations. Each point is added to the histogram weighted by the gradient magnitude, $m(x,y)$, and by a circular gaussian with σ variance that is 1.5 times the scale of the keypoint. Additional keypoints are generated for keypoint locations with multiple dominant peaks whose magnitude is within 80% of each other. The dominant peaks in the histogram are interpolated with their neighbours for a more accurate orientation assignment.

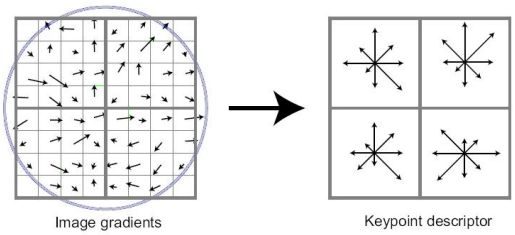


Figure 2: A keypoint descriptor is created using the gradient magnitude, $m(x,y)$ and orientation, $\theta(x,y)$ around the keypoint. These are weighted by a circular gaussian window indicated by the overlaid circle. Each orientation histogram is calculated from a 4x4 pixel support window and divided over 8 orientation bins. Figure from [Lowe, 2004]

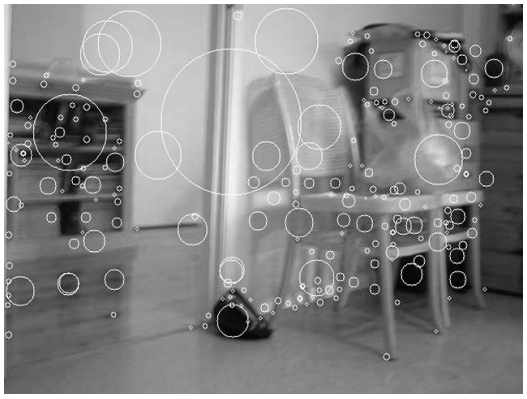
The local gradient data from the closest smoothed image $L(x,y,\sigma)$ is also used to create the keypoint descriptor. This gradient information is first rotated to align it with the assigned orientation of the keypoint and then weighted by a gaussian with σ variance that is 1.5 times the scale of the keypoint. The weighted data is used to create a nominated number of histograms over a set window around the keypoint. Typical keypoint descriptors [Koescka and Li, 2004] [Lowe, 2004] use 16 orientation histograms aligned in a 4x4 grid. Each histogram has 8 orientation bins each created over a support window of 4x4 pixels. The resulting feature vectors are 128 elements with a total support window of 16x16 scaled pixels. For a more detailed discussion of the keypoint generation and factors involved see [Lowe, 2004].

3 Reduced SIFT Features

The main drawback of SIFT features compared to other image descriptors is their high computational cost. The SIFT feature extraction process also generates a large volume of information which is redundant in most image retrieval applications. In indoor environments up to 80 percent of the common keypoints generated are not matched between images which share a common view. It is therefore desirable to reduce the quantity of keypoints generated without affecting the number of matching keypoints.

Our approach uses the structure of the indoor environment in relation to a stable viewpoint to reduce the complexity of the SIFT features to make them more efficient to calculate and match. An advantage of an indoor environment is the short average view depth of most images. This makes vertical planes such as walls the majority of an image's composition. Another advantage is that floor and roof surfaces rarely generate many SIFT features because they are composed of low contrast uniform textures. This can be seen in Figure 1 where approximately 95 percent of the SIFT features are generated around

(a)



(b)

Figure 3: Two example images from our environment database with the reduced SIFT features locations displayed as circles. The size of the circle represents the scale that the keypoint was extracted from. Figure (a) comes from Location A and figure (b) comes from Location B.

corners and edges mounted on vertical planes.

If it can be assumed that the view point for the images will be relatively stable to rotation around the view axis, the orientation of the keypoint descriptors located on vertical surfaces will not rotate. To remove the rotational invariance of the SIFT features the following three steps of the algorithm are removed.

- The calculation and assignment of keypoint orientations
- The generation of additional keypoints at locations with multiple dominant orientations
- The alignment of the keypoint descriptor to the keypoints orientation.

Removing these three steps improves the efficiency of the SIFT features in several ways. The complexity of calculating each keypoint is reduced when the orientation assignment step and the descriptor alignment step are removed. Also, less keypoints are generated because there is no need to generate multiple keypoints at a single location due to multiple peaks in the orientation assignment.

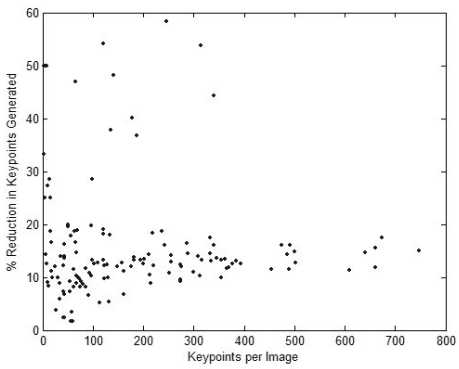


Figure 4: The percentage reduction in keypoints generated when rotational invariance is removed.

The removal of the rotation invariance steps results in only a relatively small reduction in computation complexity when compared to the more significant computational costs of the DOG and keypoint descriptor generation. However the removal of the multiple keypoints and the size of the reduced keypoint set is significant as it reduces not only the time taken to generate an image decription, it also reduces the complexity of the matching process. The reduction of keypoint generation can be seen in Figure 4.

The reduced SIFT feature relies on a stable view point to be able to remove its rotational invariance without affecting retrieval. This assumption requires the agent that generates the query and environment image databases to be rotationally stable and have a camera angle orthogonal with the vertical walls. The second requirement is the most significant as the SIFT keypoint descriptor is robust to minor rotations. If the agent is a mobile robot the second requirement restricts the robot to a camera mounted parallel with the ground in an environment which is flat. Minor bumps and tilts in the robots odometry should not affect the localisation system, however slopes or ramps with an angle greater than approximately 20 degrees (see Section 4) will greatly reduce the chance of accurately retrieving images.

4 Image Retrieval

The database of images used in this paper includes two distinct indoor environments. Location A is an Office/Lab environment situated on the top floor of the Rose Street Building in the grounds of the University of Sydney. Location B is a home environment consisting of a single floor of a private residence. There are a total of 358 images in the database, 104 from Location A (Figure 3(a)) and 156 from Location B(Figure 3(b)). There are also 98 images which represent 2 agent paths through these locations and are used as query images for the retrieval and localisation experiments. The reduced and full SIFT features for different image sizes and keypoint descriptors are extracted for each image in the database and stored.

To accurately test the effectiveness of the reduced SIFT features they need to be compared to the full SIFT features for the desired application. As the localisation application is essentially an image retrieval problem, we will use the CBIR performance measure used in [Mirmehdi and Perissamy, 2001]. This measure rates success based on percentage of images retrieved which share common content. Each image is compared to the entire database to measure the feature's retrieval rate. For this application an image was retrieved if it matched at least 5 keypoints with the query, and was considered to share common content if the view of the two images overlapped. The results of the image retrieval of the full and reduced features are shown in Table 1.

Feature	Image Size	Descriptor	No. of Retrieved Images	Average No. Of Keypoints	No. of False Positives
Full SIFT Features	640x480	4x4x8	23	18.7	0
		4x4x4	18	24.4	0
		2x2x8	26	5.83	38
	320x240	4x4x8	17	13.1	0
		4x4x4	15	13.3	0
		2x2x8	10	10.6	12
Reduced SIFT Features	640x480	4x4x8	27	20.0	0
		4x4x4	25	22.1	0
		2x2x8	40	8.93	27
	320x240	4x4x8	20	14.9	0
		4x4x4	19	13.7	0
		2x2x8	17	11.4	0

Table 1: This figure depicts the image retrieval rate of full vs reduced SIFT features for different image sizes and keypoint descriptor size

The keypoint descriptor of the SIFT features is inherently resilient to rotation due to the underlying histogram structure and gradient information. To reduce the effects of histogram boundary conditions, trilinear interpolation is used to spread the gradient information over adjacent bins. This reduces the effect of rotation on the keypoint matching and makes the features resilient to small rotational differences between matching images. Figure 5 shows the keypoints matched between two images separated by a camera rotation.

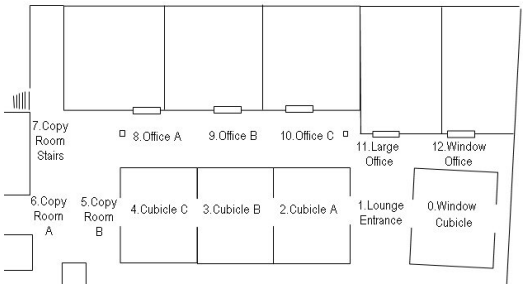
Figure 5: Keypoint matching between images rotated 17 degrees. Even under significant rotation 66 key points were matched, approximately 20 percent of the keypoints generated for the image.

5 Topological Map

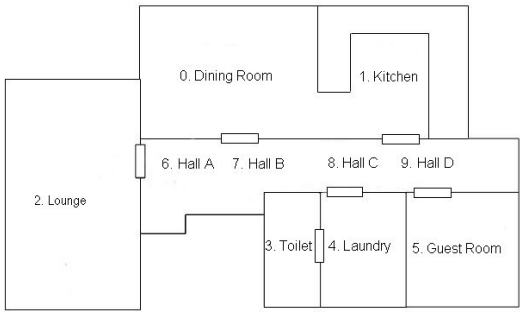
A topological map will be used to represent the indoor environments used for localisation, based on the general theory proposed by [Remolina and Kuipers, 2004]. The theory proposes that a mobile agent's experiences can be discretely segmented into distinctive states (*dstates*) with associated sensor views. The *dstates* can be further grouped into *places* which are groups of *dstates* that share a common spatial location and *paths* which represent continuous linked *places*. A *place* represents a physical location in the agent's environment at which a logical decision about the agent's actions could be made. The most common location for a *place* is at the intersection of two corridors, where an agent can choose to change its topological *path*, or a significantly important location based on the application.

Location A as shown in Figure 6(a) is segmented into 13 *places* using a ordered technique. *Places* are represented by 8 images taken at 45 degree intervals from a single physical point, which have direct line of sight to adjacent neighbouring *places*. The more rigid segmentation technique requires a large number of *places* to adequately cover possible paths through the environment or it will encounter problems if the agent leaves the vicinity of the visible *places* and *paths*.

Location B as shown in Figure 6(b) is segmented into 10 *places* based on the logical setup of the house. Each *place* is represented by a set of 10 to 20 images taken from various positions and orientations at the location. The problem with segmenting an environment based on the logical separations within it is the possibility of physical gaps in the view coverage. Unless the images used to represent the *places* have sufficient coverage, an image based localization method will have difficulty tracking an agent. Examples of gaps in view coverage resulting from missing images are present in Location B between *places* 'Hall D' and 'Kitchen', and 'Hall D' and 'Guest Room'.



(a)



(b)

Figure 6: The physical layout of the two test environments. Figure (a) shows the office/lab environment, Location B, segmented into 13 labeled *places*. Figure (b) shows the home environment, Location A, segmented into the 10 labeled *places*.

Two sequences (I, II) were used to test the reduced SIFT features, one from Location A and one from Location B. The test sequences are composed of still digital images taken at approximately 50 cm or 45 degrees intervals. The sequences span the majority of location and transitions in both environments.

6 Localisation using Retrieved Images

Localisation using retrieved images has a greater need for features with good image discrimination. The removal of the rotational invariance stage has reduced the dimensions of the features and potentially their effectiveness for localisation. A vision based localisation application

will be used to verify that the reduced SIFT features are still effective. There are several previous examples of retrieved image and SIFT feature localisation systems. In [Koescka and Li, 2004] and [Ulrich and Nourbakhsh, 2000] voting schemes were used to determine the most likely location of an agent given the set of retrieved images. In [Wolf *et al* , 2002] they calculate the visibility regions of the retrieved images combined with a Monte-Carlo filter. We use a probabilistic method based on a recursive Bayesian filter [Durrant-Whyte, 2001]. The probability of the current state x given the sequence of observations Z^k up to time k is

$$P(\mathbf{x}|\mathbf{Z}^k) = \frac{(P(\mathbf{z}_k|\mathbf{x})P(\mathbf{x}|Z^{k-1})}{P(\mathbf{z}_k|Z^{k-1})} \quad (4)$$

where the *sensor model* or *Likelihood Function*, $P(\mathbf{z}_k|\mathbf{x})$, is calculated using a method similar to the one described in [Koescka and Li, 2004]. In this method each image in the database is represented by a set of SIFT features $\{S_n(I_j^i)\}$, where n is the number of keypoints for the jth image at the ith location. For each query image Q and its associated keypoints $\{S_m(Q)\}$ a set of corresponding keypoints between Q and each image in the database I_j^i, $\{C(Q, I_j^i)\}$, is calculated. $\{C(Q, I_j^i)\}$ is calculated using an Euclidian distance measure as described in [Lowe, 2004].

This method then calculates the conditional probability, $p(\mathbf{z}_k|\mathbf{x}_k = \mathbf{x}_i)$, that a query image $\mathbf{Q}_k$ at step k characterised by an observation $\mathbf{z}_k$ $=\{S_m(Q_k)\}$ came from location i. This is calculated using the correspondence set $C(i)$, normalised by the total number of matched keypoints across all images.

$$p(\mathbf{z}_k|\mathbf{x}_k = \mathbf{x}_i) = \frac{C(i)}{\sum_j C(j)} \quad (5)$$

6.1 Hidden Markov Model

The Bayesian filter can be significantly improved with the addition of spatial information between image views. The relationships between locations can be modeled by a Hidden Markov Model (HMM) as shown in [Torralba *et al* , 2003]. In our model the states correspond to individual *places* and the transition function determines the probability of moving between *places*. The conditional prior is as follows

$$p(\mathbf{x}_k = \mathbf{x}_i|Z^{k-1}) = \sum_j^n A(i,j)p(\mathbf{x}_{k-1} = \mathbf{x}_i|Z^{k-1}) \quad (6)$$

where N is the number of *places* and $A(i, j)$ is the transition matrix. A is a N x N matrix where $A(i, j) = P(\mathbf{x}_k = \mathbf{x}_i|\mathbf{x}_{k-1} = \mathbf{x}_j)$. Entries in the matrix corresponding to adjacent locations are assigned a value of one and the final matrix is normalised across each row. The results of the localisation using the full and reduced SIFT features are shown in Figure 7. Localisation probabilities were obtained by running the Bayesian filter against full and reduced SIFT feature sets at 640x480 resolution with a 4x4x8 keypoint descriptor.

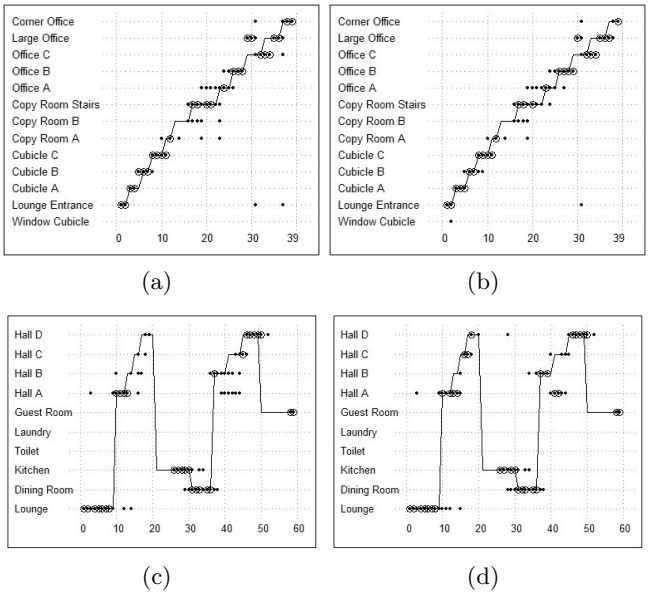


Figure 7: These 4 figures show localisation results for the full and reduced features. The line represents ground truth for the images, with a dot representing partial confidence and a circle representing full confidence in the agents location. Figures (a) & (b) are from location A and Figures (c) & (d) are from location B. Full SIFT features are used in Figures (a) & (c) and reduced SIFT features in Figures (b) & (d)

In these results a partial confidence classification is given to *places* with a probability of greater than 0.2 and a full confidence classification is given to *places* with a probability of greater than 0.5. The difference is marginal between the effectiveness of the full and reduced SIFT features. The filter using either feature has low false positive rate, approximately 5%, and never strays from the true paths view area. The reduced SIFT features on average takes 25% less time to match and retrieve.

As described in Section 5 an environment segmented into logical *places* can encounter gaps in the view coverage. This is shown in Figure 7(c) & (d) when the agent transitions between *places* 'Hall D' and 'Kitchen' in steps 20 to 25, and 'Hall D' and 'Guest Room' in steps 51 - 58. During these times there are no images that can be retrieved from the environment forcing the probability of the location towards an even distribution. In Path I, the agents uses an invisible *path* in Location A because its view does not directly overlap with images from the

environment. This results in a loss of confidence until the agents view covers a known *place*. This is shown in Figure 7(a) & (b) during steps 14 to 17, when the agent moves through the copy room area. This loss of confidence can be solved by using a more comprehensive image representation of the environment.

6.2 Scale Space Refinement

The current Bayesian filter has problems distinguishing between locations which are physically close neighbours, because they will share high keypoint matching to images in overlapping views. The resulting ambiguity can be seen in Figure 7(c) when the agent enters the hall. A possible solution to this problem is weighting the likelihood of the sensor model with the scale space information extracted from the SIFT features. The average ratio of scale between the matched keypoint pairs is used with a gaussian to weight the number of matched keypoints for a *place*. The results of using a gaussian with different values of σ is shown in Figure 8.

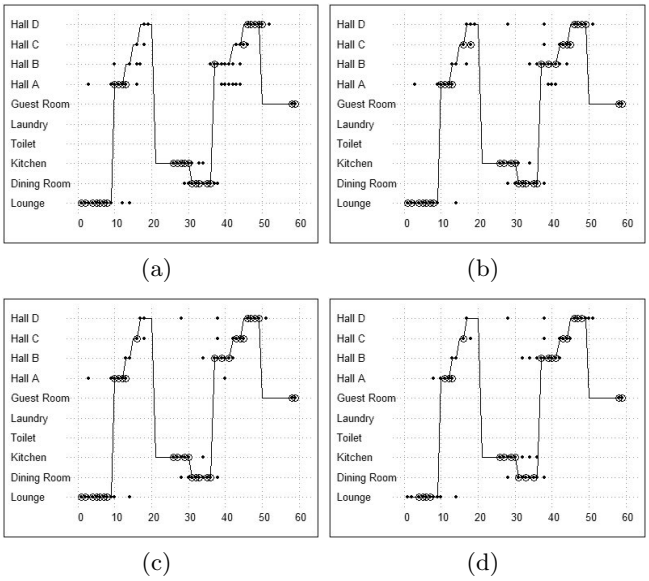


Figure 8: These 4 figures show the scale space refinement for Path II from Location B. The line represents ground truth for the images, with a dot representing partial confidence and a circle representing full confidence in the robots location. Figure (a) shows the unweighted results and Figures (b)-(d) have had image retrieval weighted by a gaussian with σ equal to 0.30, 0.15 & 0.1 respectively.

The addition of scale space refinement significantly improves the accuracy of the filter. Results were obtained by running the Bayesian filter with and without scale space refinement against full and reduced SIFT feature sets at 640x480 resolution with a 4x4x8 keypoint descriptor. Using scale space refinement the filter generated 35% less false partial confident classifications, 59% less false positive classifications and 15% more correct confident classifications.

7 Conclusion and Future Work

SIFT features provide a distinct and accurate means of matching digital images for image retrieval and vision based localisation. In this paper we have presented a reduction to the traditional SIFT feature to improve their performance. This reduction uses the structure of an indoor environment to remove the need for rotational invariance of the features. From the results obtained we have shown that this reduction has a minimal affect on the retrieval rate of images and significantly reduces the size of the image descriptors and the time to needed to generate and match them. Furthermore we have used the scale information of the SIFT features to improve location discrimination.

Future work with the reduced SIFT feature will look at further reducing the size of the image description by filtering keypoints with low matching likelihood. Partially unknown environments with on-line learning of the topological map and underlying HMM is another area of current investigation.

8 Acknowledgements

We would like to thank David Lowe for the SIFT feature code. This work is supported by the ARC Center of Excellence programme, funded by the Australian Research Council.

References

[Brown and Lowe, 2002] Matthew Brown and David G. Lowe. Invariant features from interest point groups. In *British Machine Vision Conference*, BMVC 2002, Cardiff, Wales (September 2002), pp. 656-665.

[Durrant-Whyte, 2001] Hugh Durrant-Whyte. Multi Sensor Data Fusion. Course Notes, University of Sydney, January 2001.

[Goodrum, 2000] A. A. Goodrum. Image Information Retrieval: An Overview of Current Research. In *Informing Science*, Volume 3 No 2, 2000.

[Eakins and Graham, 1999] J. Eakins and M. Graham. Content-Based Image Retrieval. tech. rep., *JISC Technology Applications Programs*, University of Northumbria, http://www.unn.uk/iidr/report.html, 1999.

[Iqbal and Aggarwal, 1999] Q. Iqbal and L. K. Aggarwal. Using Structure in Content-based Image Retrieval. In *Proceedings of the IASTED International Conference on Signal and Image Processing*, pp 123-133, October 1999.

[Koescka and Li, 2004] Jana Kosecka and Fayin Li. Vision Based Topological Markov Localization". In *IEEE International Conference on Robotics and Automation*, May 2004.

[Koskela *et al* , 2001] M. Koskela, J. Laaksonen and E. Oja. Comparison of Techniques for Content-Based Image Retrieval. *Proccedings of SCIA*, Bergen, Norway, June 2001.

[Laaksonen *et al* , 2000] J. Laaksonen, E. Oja, M. Koskela and S. Brandt. Analysing Low-Level Visual Features Using Content-Based Image Retrieval. *Proceedings of ICONIP*, 2000.

[Li and Wang, 2003] Jia Li and James Z. Wang. Automatic Linguistic Indexing of Pictures by a Statistical Modelling Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1075-1088, 2003.

[Liu and Picard, 1996] F. Liu and R. W. Picard. Periodocity, Directionality and Randomness: Wold Features for Image Modelling and Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intellegence* , vol 18, no 7, July 1996.

[Lowe, 2004] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* , 60, 2 (2004), pp. 91-110.

[Lowe, 1999] David G. Lowe. Object recognition from local scale-invariant features. *International Conference on Computer Vision* , Corfu, Greece (September 1999), pp. 1150-1157.

[Mikolajczyk and Schmid, 2001] Krystian Mikolajczyk and Cordelia Schmid. An Affine Invariant Interest Point Detector. *European Conference on Computer Vision*, vol 1, pp. 128 - 142, 2002.

[Mirmehdi and Perissamy, 2001] M. Mirmehdi and R. Perissamy. CBIR with perceptual region features. In *Proceedings of the 12th British Machine Vision Conference* , pages 511-520. BMVA Press, September 2001.

[Remolina and Kuipers, 2004] Emilio Remolina and Benjamin Kuipers. Towards a General Theory of Topological Maps. *Artificial Intelligence* , vol. 152, no. 1, pp. 47-104, January 2004.

[Schmid and Mohr, 1997] Cordelia Schmid and Roger Mohr. Local Greyvalue Invariants for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, May 1997.

[Schulz-Mirbach, 1995] Hanns Schulz-Mirbach. Invariant Features for Gray Scale Images. *DAGM-Symposium I*, 1995: 1-14

[Se *et al* , 2002] Stephen Se, David G. Lowe and Jim Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *International Journal of Robotics Research* , 21, 8 (2002), pp. 735-758.

[Siggelkow, 2002] Sven Siggelkow. Feature Historgrams for Content-Based Image Retrieval. PhD Thesis, *Albert-Ludwigs-University Frieiburg* , December 2002.

[Torralba *et al* , 2003] A. Torralba, K. Murphy, W. Freeman and M. Rubin. Context-based vision system for place and object recognition. *International Conference on Computer Vision* , 2003.

[Ulrich and Nourbakhsh, 2000] I. Ulrich and I.Nourbakhsh. Appearance-Based Place Recognition for Topological Localisation. In *IEEE International Conference on Robotics and Automation* , San Francisco, pp 1023-1029, April 2000.

[Wolf *et al* , 2002] J. Wolf, W. Burgard, H. Burkhardt. Robust Vision-based Localization for Mobile Robots using an Image Retrieval System based on Invariant Features. In *Proceedings of the IEEE International Conference on Robotics and Automation* (ICRA), 2002.