Representation and Learning of Visual Information for Pose Recognition

David Prasser and Gordon Wyeth
School of Information Technology and Electrical Engineering
University of Queensland
St. Lucia, Queensland 4069
Australia
prasserd@itee.uq.edu.au, wyeth@itee.uq.edu.au

Abstract

Recovering position from sensor information is an important problem in mobile robotics, known as localisation. Localisation requires a map or some other description of the environment to provide the robot with a context to interpret sensor data. The mobile robot system under discussion is using an artificial neural representation of position. Building a geometrical map of the environment with a single camera and artificial neural networks is difficult. Instead it would be simpler to learn position as a function of the visual input. Usually when learning images, an intermediate representation is employed. An appropriate starting point for biologically plausible image representation is the complex cells of the visual cortex, which have invariance properties that appear useful for localisation. The effectiveness for localisation of two different complex cell models are evaluated. Finally the ability of a simple neural network with single shot learning to recognise these representations and localise a robot is examined.

1 Introduction

Vision based localisation is the problem of determining the position of a robot from the information provided by its visual sensor. This paper discusses early work on a localisation mechanism for the RatSLAM robot[1]. The RatSLAM project aims to use neuro-physiological methods to perform Simultaneous Localisation And Mapping (SLAM), a larger problem which involves localisation while learning an environment under uncertainty. For this application it would be appropriate to use an artificial neural network to localise the robot. A neural network could be used to learn a mapping between the visual input and the robot's position. This approach which learns the appearance of places avoids the problems of map building and geometric reasoning.

In this paper the learning process will be performed by a neural network known as the Extended Conjunction of Localised Features network (ECLF). This network can perform single shot learning of visual inputs and relate them to cells that represent particular positions.

Using unprocessed image data with a learning system is a difficult proposition and it is usual to perform some level of pre-processing and represent the data in some simplified fashion. The nature of the pre-processing will affect the performance of the learning system. This paper examines the performance and learnability of two variants of a biologically inspired representation of visual impression.

1.1 RatSLAM and Place Representation

The aim of this research is to develop place recognition techniques for the RatSLAM project (also in these proceedings [Milford and Wyeth, 2003]). This project involves the implementation of a biologically plausible navigation system based on studies of brain activity within the hippocampus of rats. In this representation position is encoded in 'place cells' and 'head direction' cells. Each place cell responds maximally when the rat is in a particular position and each head direction cells when the rat is orientated at a particular bearing. In the RatSLAM model the activation of these cells increases gradually as the robot approaches the cells preferred location, because of this many cells will be activated at one time creating a hill of activity centred on the robot's estimated position. Mechanisms exist within this neural model to change the distribution of cell activity as the robot moves around, accounting for changes in the robots internal sensors.

The place cells are connected to a representation of the visual input known as the local view (LV). At the present time the RatSLAM project is using an artificial landmark system to provide it with its LV information [Prasser and Wyeth, 2003]. Replacing this with a system that can learn natural scenes is a major objective of the RatSLAM project.

1.2 Neural Network for Learning Position

A simple neural network structure known as the Conjunction of Localised Features network (CLF) was proposed to demonstrate object recognition [Edelman, 1991; Edelman and Weinshall, 1991]. This was adapted into the Extended CLF (ECLF) network specifically for robot navigation [Chan and Wyeth, 1999]. This network is capable of single shot learning which makes it very

1 This research is sponsored in part by an Australian Research Council grant.

attractive for a robot that is exploring and learning an environment. The output layer (O-Layer) of the network contains cells that respond to a particular position, these cells have a correspondence with the hippocampal place cells used in the RatSLAM system. There is a similar correspondence between the representation layer of the ECLF network and the local view of the hippocampal model. The network therefore seems to fit neatly with the RatSLAM structure (Figure 1).

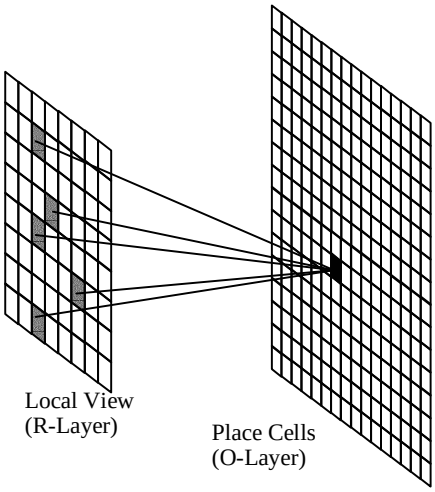

Figure 1: The RatSLAM system contains a layer of cells each of which represents a particular location (place cells). Visual input is represented as a group of cells corresponding to visual features (local view). In the ECLF network the place cells are the output layer (O-Layer) and the local view corresponds to the representation layer or R-Layer. The R-Layer has internal connections between individual units.

1.3 Summary of Paper

Firstly the approach used in this paper to deal with the problem of vision based localisation is outlined in section 2. Some techniques for representing images are discussed as well as the general operation of a basic neural network that performs image learning. The experimental procedure is outlined in section 3, along with a description of the mathematical models and learning rules used. The performance of both the image representations and the learning technique are shown in section 4. Some observations about the effectiveness of the representations and the neural network are made in the following section.

2 Approach

The first problem of appearance based mapping or learning is image representation. The second problem is the actual learning process. In this section both of these problems will be addressed and the solutions investigated in this paper will be described.

2.1 Input Representation

Rather than learn the image directly it is necessary to represent the image in a manner that is easy to learn. This representation should be slightly spatially invariant, that is small changes in the camera position should not lead to significant changes in the output representation. Large changes in camera position of course must cause a change in the representation otherwise determining camera position is impossible.

Histograms [Gonzalez-Barbosa and Lacroix, 2002; Ulrich and Nourbakhsh, 2000] and Principle Component Analysis (PCA) [Kröse and Bunschoten, 1999; Nayar *et al.*, 1994; Pourraz and Crowley, 1998] have been proposed as image representations for robot localisation. PCA cannot be performed incrementally which prevents the sort of exploratory learning needed for RatSLAM. Histograms are usually employed with omnidirectional cameras, where by their nature they provide a completely rotationally invariant representation. This invariance is accomplished by ignoring the position of image features when constructing the histogram. Without an omnidirectional camera total rotational invariance cannot be achieved and it may become difficult to recover orientation reliably.

Simple Input Representation

In the original ECLF implementation a primitive visual representation was described [Chan and Wyeth, 1999]. In this representation a 64 × 64 pixel greyscale image was reduced to a 10 × 10 binary matrix, whose individual elements correspond to small overlapping regions of the original image. The matrix encodes the amount of contrast, as measured by a Laplacian operator, in each region of the original image (Figure 2).

Figure 2: A low dimensional representation of a greyscale image. The input image on the left is reduced to a binary image on the right. High values (white) of the output image correspond to regions of high contrast in the input.

Complex Cells

Another method that has been examined for low dimensional representations of images is based on the complex cells of the visual cortex. Complex cells are generally considered to detect or respond to edges or bars at a particular orientation within a region of the retina, which is known as the cell's receptive field [Hubel, 1988]. As complex cells are insensitive to local changes in feature position they are suitable for spatially sub-sampling an image and also generalisation. For example, if a complex cell has a receptive field of 6° × 6° then panning or tilting the camera by 3° will not change the cell's activity significantly. Complex cells have been investigated in image learning tasks [Edelman *et al.*, 1997] and are a common feature in hierarchical visual networks [Fukushima, 2001; Riesenhuber and Poggio], they have also been used as image primitives in robot pose recognition [Arleo *et al.*, 2001].

Artificial complex cells are often constructed from Gabor or Derivative of Gaussian wavelets with some form of nonlinearity. One simple model of these cells is that of an energy mechanism or quadrature pair [Spitzer and Hochstein]. An odd and even pair of Gabor filters is convolved with the data making the sum of the square of their responses locally phase independent. Heeger [1992] extends this by adding a nonlinearity where the output of nearby cells normalises the cells output, which inhibits

cells with weak activation that are near highly activated cells.

An alternative to the energy mechanism approach is to pool the outputs of Gabor filters in a non-linear manner, for example taking the maximum over a small region of an image [Riesenhuber and Poggio] or a non-linear summation [Wersing and Körner, 2003] (Figure 3).

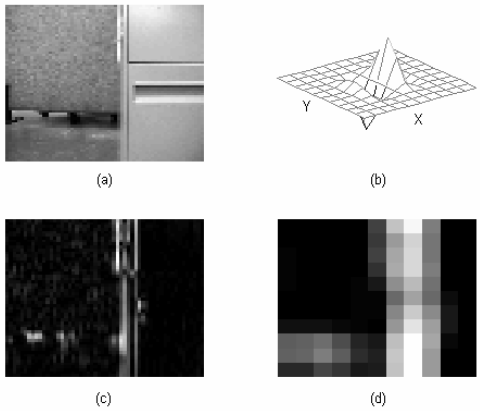

Figure 3: Spatial pooling complex cells. (a) Original image. (b) Gabor filter tuned to detect changes along the X axis. (c) The absolute output of the Gabor filter. (d) Complex cell outputs. The implementations in this paper have other sets of cells tuned to three more edge orientations

2.2 ECLF network

The ECLF network is provided with a vector of features as an input, during training it learns to recognise different sets of features and also to associate sets of features together. The network consists of two layers, the representation layer (R-Layer) and the output layer (O-Layer). Both layers have specific learning rules and functions.

The R-Layer, which has the same number of units as the length of the input vector, is responsible for associating features together. Each unit on the R-Layer is connected to every other unit with an initial weight of zero. During training, the connection weights are updated in a Hebbian manner when their corresponding inputs in the feature vector are activated. In other words, if the two features are presented as being activated at the same time the connection weights between their two corresponding units on the R-Layer will be increased.

The O-Layer is used to represent the output of the network. In the context of localisation or path recognition the units represent particular locations. Each unit in the O-Layer is connected via vertical connections (V-Connections) to all of the units in the R-Layer.

3 Experimental Setup

The various image representations and the ECLF network will be evaluated on common data sets. In this section the implementation details of both the image representations and the ECLF network are described, as well as the nature of the data sets.

3.1 Simple Representation

The simple representation begins by convolving the input image with a 3 × 3 contrast detecting filter and then thresholding:

$$t = I * \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} > T \qquad (1)$$

Further resolution reduction is then applied reducing the 66 × 66 pixel image to a 10 × 10 map. This reduction is performed by dividing the image into overlapping squares each of which is 7 pixels on a side. The number of highlighted pixels in each squares area is computed and compared to a threshold. This final thresholding results in a 10 × 10 map of binary cells. Each of these cells has a receptive field of about 4.8° horizontally by 3.8° vertically.

3.2 Complex Cells

Two complex cell models were evaluated for use as a representation of visual information. The images provided were 64 × 64 pixel greyscale images which represented about 48° horizontally × 38° vertically. For both evaluations the fundamental Gabor filters had the following properties $\sigma_x = \sigma_y = 2$ and $\omega = 2$.

$$g^l(x,y) = e^{j\omega x} e^{\frac{\omega^2}{2}\left(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2}\right)} \qquad (2)$$

For both models four orientations of Gabor filters were used, resulting in a three dimensional array of cells. The first two dimensions *i* and *j* correspond to the cell's physical position on the retina, while the third $l = 0, \pi/4, \pi/2, 3\pi/4$ denotes the orientation of the stimulus that the cell responds to.

The performance of the cells was examined by using their output as the input for a nearest neighbour classifier (NNC) which attempted to recover orientation information. The error from the NNC was used to evaluate the cell's performance.

Energy Model

The orientated energy model that was examined used a normalisation based on Heeger [1992]. In this implementation the energy $E^l_{i,j}$ is computed by vector dot multiplying the part of the image that is in the cells receptive field, $I_{i,j}$ by the vectorised complex Gabor filter and then taking the sum of the squares of the real and imaginary components. The energy outputted by the complex cell at position *i*, *j* and orientation *l*, is then normalised by the total output in its neighbourhood:

$$E^l_{i,j} = \mathrm{Re}\left(g^l \cdot I_{i,j}\right)^2 + \mathrm{Im}\left(g^l \cdot I_{i,j}\right)^2 \qquad (3)$$

$$E^l_{ni,j} = \frac{E^l_{i,j}}{\delta + \sum_{i',j'} \sum_{l'=0}^{l'=3\pi/4} E^{l'}_{i'j'}} \qquad (4)$$

The semi-saturation constant δ for the cells prevents very low levels of cell energy from being normalised to a large value when there is only a small normalising signal. Finally the normalised energy is passed through a sigmoid nonlinearity with gain *A*:

$$c_{i,j}^{l} = \frac{1}{1+\exp\left(-A\left(E_{ni,j}^{l}-1\right)\right)} \quad (5)$$

Spatial Pooling

The spatial pooling method from the first layer of [Wersing and Körner, 2003] is computed in a different manner. In this model the image is convolved first with four odd Gabor filters to produce four resultant images q^l, with l = 0, $\pi/4$, $\pi/2$, $3\pi/4$. Each of these is then passed through a winner-takes-most mechanism across the orientation dimension:

$$r^{l}(x,y) = \begin{cases} 0, & \dfrac{\left|q^{l}(x,y)\right|}{M} < \gamma \\ \dfrac{\left|q^{l}(x,y)\right| - M\gamma}{1-\gamma}, & \dfrac{\left|q^{l}(x,y)\right|}{M} \geq \gamma \end{cases} \quad (6)$$

Where $M = \max_k|q^k(x,y)|$ is the maximum absolute response at a particular pixel location for any of the four orientations and γ is a competition parameter. The filter outputs are pooled with a local summation:

$$c^{l}(i,j) = \tanh \sum_{x,y} p(x,y) H\left(r^{l}(x,y) - \theta\right) \quad (7)$$

Where H is a unit step function and $p(x,y)$ is a two dimensional Gaussian distribution which defines the spatial extent of the complex cell's receptive field. The centre of the distribution x_i', y_j' specifies the centre of the receptive field while the σ parameter controls the size. The hyperbolic tangent is used to limit the result. In this work σ is constant and x_i', y_j' vary to create a square grid of evenly sized complex cells on the image surface.

3.3 Implementation of ECLF Network

The ECLF network was trained by first pre-processing several reference images into either the simple or complex cell representations. Each training image is applied once to the R-Layer which is the same size as the image representation. An O-Layer unit is created for each training image and associated with the position the image was acquired at.

In training the weight $w_{i,j}$ between two R-layer units i, and j is increased from zero depending upon the activation of the two units (A_i and A_j) up to a maximum strength of $w_{\max}$.

$$\Delta w_{i,j}(t+1) = \frac{w_{\max} - w_{i,j}(t)}{2} A_i A_j \quad (8)$$

In recall the final activation A_i of R-unit i is the sum of the initial activations of all other R-Layer units multiplied by the appropriate connection weight. The initial activation a_j of unit j is the activation from the input layer without the effect of lateral connections.

$$A_i = \sum_j w_{i,j} a_j \quad (9)$$

When the network is being trained a learning rule increases the connection strength $w_{r,v}$ between activated R-Layer unit r, and the output unit v, on the O-Layer that has been activated by the training process. Initially all of the R-Layer to O-layer weights are set to unity.

$$\Delta w_{r,v} = \min\left(\alpha w_{r,v} A_r A_v, w^m - w_{r,v}\right) \frac{w^m - w_{r,v}}{w^m} \quad (10)$$

Where α controls the learning rate and w^m is the maximum vertical weight. The unit's threshold increases from zero proportionally to the increase in the sum of the weights connected to it:

$$\Delta T_v = \sigma \sum_r \Delta w_{r,v} \quad (11)$$

In recall the activation of an O-Layer unit is the linear sum of the activation from the R-Layer through the weights:

$$A_v = \sum_r w_{r,v} A_r - T_v \quad (12)$$

The non-maximally activated O-layer units are then suppressed.

3.4 Test Data

These experiments all used sets of data captured from the forward looking camera of a Pioneer 2DX mobile robot. This data was synchronised with the robot's position estimate obtained through path integration of wheel displacements. This path integration information is used as absolute data, although it should be recognised that over time path integration becomes increasingly inaccurate.

4 Results

The goal is to measure the performance of two interconnected systems: the input representation; and the effectiveness of the ECLF network in learning these input representations.

4.1 Image representation

One way of characterising the effectiveness of an input representation would be to measure the error of a simple classifier performing tasks using the representation. A Nearest Neighbour Classifier (NNC) was used as a basic classifier. The error was measured in two ways: the mean error and the percentage of correct classifications. The mean error is calculated as the average absolute physical distance between the location the test images was acquired at and the location the closest matching training image was acquired at. Due to the fact that errors in odometry accumulate over time the mean error will have a systematic component even if recognition is perfect. The output resolution will also have a systematic contribution to the mean error. The classification rate is the fraction of test images that were classified as the physically closest training image.

Simple Image Representation

The 10 × 10 binary representation obtains a classification rate of 62.2% when trained with one NNC template every 7.4°, unfortunately the mean error in degrees of the output is then 21.1°. If the angle between training images is reduced to 1° then the mean error is only 2.5°.

Energy Based Complex Cells

Four parameters of the energy based complex cell were

examined: the semi-saturation constant, δ; the size of the normalising region; the gain of the sigmoid nonlinearity; and the spacing of the complex cells on the retina. A nearest neighbour classifier was used to perform an orientation learning task with 29 reference images taken every 12.5°. A large variety of parameter combinations were trialled and the mean error in degrees of the classifier and the fraction of classifications that selected the closest output recorded. The parameters that produced the best performance by both metrics are shown in Table 1.

Mean Error (°)	Class-ification Rate (%)	δ	Normalising Region Size	Sig-moid Gain A	Cell Spacing
6.85*	75.8	.015	7	4.5	4
7.54	78.6*	.015	11	8.0	4

Table 1: The parameters that produced the best classification rate and mean error for orientation recognition using energy model complex cells to represent the visual input. The entry denoted by an * is the criterion that was optimal over the search space. There were 29 training images and therefore output resolution is 12.5°. The expected mean error is one quarter of the output resolution or 3.1°.

The mean error criterion produces more attractive results than the number of correct classifications. This is because the mean error is sensitive to outlying errors, while the number of incorrect classifications does not indicate the severity of the error. The fact that the best parameters are very similar though indicates that the distinction is not a very important one.

To examine the ability of the cells to learn position information a second NNC experiment was conducted. In this experiment the classifier was given the problem of recovering robot position with a fixed orientation. 48 training images were provided at 100 mm intervals in X and 500 mm intervals in Y over a 1 m × 3 m area. The results are shown in Table 2.

Optimised for	Mean error in X (mm)	Mean Error in Y (mm)	Mean Abs Error (mm)
Mean Angular Err	83.5	426.4	451.7
Angular Classification Rate	82.5	426.2	451.2

Table 2: The energy model representations used for orientation learning were also tested with the task of recovering 2 dimensional position information. The camera was fixed along the Y axis. Both sets of parameters from Table 1 were evaluated.

Spatial Pooling Complex Cells

Again a rough search of the parameter space was performed and one set of parameters was found that produced both a good classification rate and a low mean error. Like the energy based cells the minimum threshold is low and the spatial extent large. The best result was found with a cell spacing of 3 as compared to 4 for the energy model cells. This means that for the same sized image there will be more spatial pooling cells used to represent the image than energy model cells. To see how important inter-cell spacing is, the cell spacing was fixed at 4 and the performance of the cells revaluated.

Mean Error (°)	Classification Rate (%)	θ	γ	σ	Cell Spacing
5.38 *	76.7 *	0.05	0.3	7.0	3
5.83 *	75.6	0.1	0.1	7.5	4(fixed)
5.86	76.7 *	0.05	0.3	7.5	4(fixed)

Table 3: The spatial pooling cells. The best result of those searched had a spacing of three pixels or 2.3° between receptive field centres. When the cell spacing is constrained to four pixels the performance decreases but not significantly. Again, * denotes results that were optimal for the parameter search space and the expected mean error is 3.1°.

The performance was not significantly degraded by increasing the cell spacing to 4 pixels between receptive field centres. Since increasing this parameter reduces the size of the input representation and therefore the entire network this is a helpful result.

The ability to resolve the position instead of orientation was evaluated with the same data as for the energy model cells.

Cell Spacing Pixels	Mean error in X (mm)	Mean Error in Y (mm)	Mean Abs Error (mm)
3	70.5	371.8	396.1
4	79.4	391.0	418.1

Table 4: The spatial pooling complex cell representations used for orientation learning were used in the task of recovering 2 dimensional position information. The camera was fixed along the Y axis. The first two sets of parameters from Table 3 were evaluated.

4.2 Results for ECLF Network

Orientation Learning

The ECLF layer was trained using the output of both types of complex cells with several different sets of parameters, as well as the simple representation. The ECLF network was evaluated for learning orientation from visual input and the network parameters were adjusted to give good performance for each representation. The training and test images came from the same data set that was used to evaluate the different image representations in section 4.1. In the complex cell experiments the image representation is considerably larger than for the simple representation. As the size of the R-Layer is the square of the number of input features the network became unwieldy with such a large representation, so the R-Layer was removed from the system. The results for the different input representations are shown in Table 5. Notice that the results for the spatial pooling cells are slightly better than with the NNC in Table 3, this is because the ECLF network suppresses particularly bad matches.

Representation	Mean Error (°)	Classification Rate (%)
Simple Representation	29.7	22.6
Energy Based (Mean Error)	11.3	71.9
Thresholded Energy Based (Mean Error)	9.8	75.8
Energy Based (Classification Rate)	11.3	71.9
Thresholded Energy Based (Classification Rate)	9.9	75.8
Spatial Pooling (3 pixel cell spacing)	5.6	75.8
Spatial Pooling (4 pixel cell spacing)	5.7	77.5

Table 5: All of the different types image representations discussed in Section 4.1 were used as input representations to

train an ECLF network which was given the task of learning orientation from The simple representation had an interval of 6° between training images, less than half that of the complex cell representations.

With the simple visual representation the robot was able to use the ECLF network to learn orientation in the real world. The network used 29 O-layer units and had an output resolution of 12.5°. A histogram of the error in this test is shown in Figure 4.

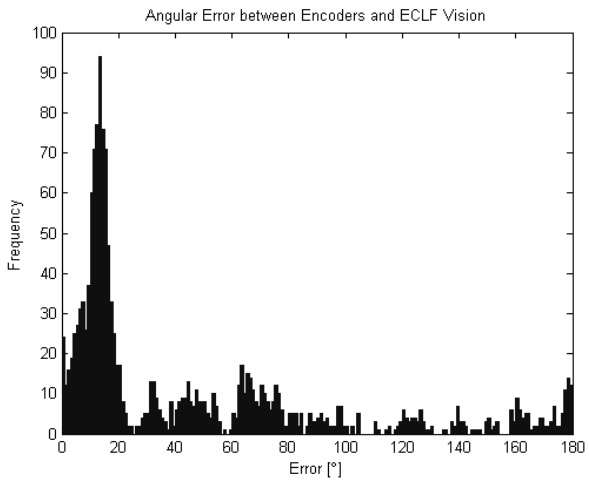

Figure 4: Histogram of error in angular position of the robot after performing an angular localisation task in the real world. This error is calculated by comparing the orientation calculated from odometry to that provided by the ECLF network. The median error is 15.3° which possibly indicates some encoder slip between the robots mapping and localisation phases.

Position Learning

From the results in section 4.1 it plausible that a robot's two dimensional position could be learnt from either of the complex cell representations. Using the same network and complex cell parameters as for orientation learning the results in Table 6 were obtained.

Representation	Mean error in X (mm)	Mean Error in Y (mm)	Mean Abs Error (mm)
Simple Representation	200.1	836.4	889.4
Energy Based (Mean Error)	89.2	450.1	476.8
Thresholded Energy Based (Mean Error)	83.4	433.5	458.9
Spatial Pooling (3 pixel cell spacing)	73.2	373.1	398.4
Spatial Pooling (4 pixel cell spacing)	77.6	403.3	428.7

Table 6: The performance of the ECLF network recognising two dimensional position. The complex cell and network parameters are left unaltered from before, that is they are still optimised for learning orientation.

5 Discussion

Several points can be made about this work particularly the way the complex cells are tuned and the way the ECLF network learns.

5.1 Position Learning

It is worth noting that the position learning experiments occurred in a different environment without any changes to either the complex cell parameters or the network parameters. A new set of spatial pooling complex cells were designed using the task of NNC based position rather than orientation recognition. These cells had better results, with a mean error of 358.8 mm which is 15% better than the spatial pooling cells designed for orientation recognition. These cells had a σ of 3 and a cell spacing of three pixels which makes the cells receptive field much smaller and have less overlap than the orientation ones. These cells have more error when used for orientation (mean error is 7.4° compared to 5.4°). How much of this problem is caused by differences in the environment as opposed to differences in the task is under investigation. It seems likely though that the level of invariance required for the position learning task is lower than that needed for the orientation learning task.

5.2 Cell Activity

From the results in Table 5 it is clear that the representation can have significant effects on the localisation accuracy, although both classes of complex cell models appear fairly similar – both representations have cells that are tuned to one of four orientations and the cells make a grid approximately 11 x 11. To see what is different a histogram can be made of cell activity over all orientations and images in the training set for each type of cell. These are shown in Figure 5, where it can be seen that there is a greater likelihood of the cell in the energy model representation having an activation of unity compared to the spatial pooling model. The spatial pooling models appear to be sparser and to have a more gradual change in value.

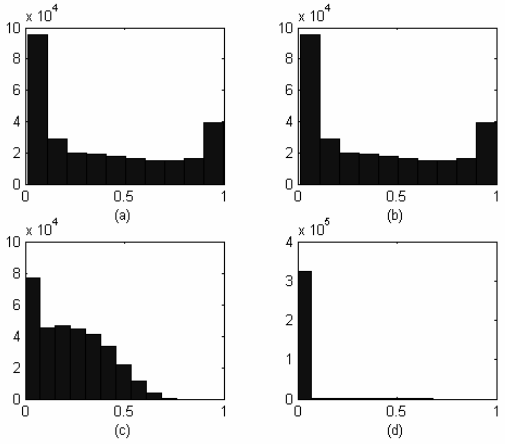

Figure 5: Histograms of cell activity for four different complex cell representations: (a) mean error optimised energy model cells; (b) classification error optimised energy model cells; (c) spatial pooling cells with three pixel cell spacing; and (d) spatial pooling cells with four pixel cell spacing.

5.3 Denser Training Sets

The number of training images is fairly sparse in these experiments. This was to see how far the network could be pushed as it was believed that requiring many training images would make the system unsuitable for autonomous operation on a robot. This requires the robot to have a capacity for interpolation. The NNC classifier is fairly good at interpolation as the match between the input and a

template will decline gracefully as the input changes slightly from the template. The ECLF network on the other hand will actually increase in fitness if the value of an input cell increases. This explains why the thresholded energy cells had better performance in Table 5 than the unthresholded energy cells.

5.4 Improved Complex Cell Optimisation

The rough search for good parameters for the complex cell models can't really be considered to be anything like a learning process. Possibly using an optimisation procedure could find a better set of parameters. If this is done it would be wise to acquire training images from a large number of environments, otherwise there is a risk of the cells becoming specialised for one environment. Also allowing variation in the underlying Gabor filters may improve performance further.

5.5 Learning Uncertainty in Position Data

So far it has been assumed that the visual learning process must form associations between visual inputs and an activated place cell. In the context of the RatSLAM project however the place cell activity is actually distributed in a Gaussian lump, which can be considered to represent the uncertainty in the position estimate. It would be logical to associate the lump with the visual input in a manner similar to that performed in the artificial landmark RatSLAM system [Milford and Wyeth, 2003]. A simple network like the ECLF network is not able to learn this data, mostly because it functions at least in part as a linear feed forward neural network and as such will have problems with linear separability. When only associating a pattern to one place cell the network functions more as a correlator or pattern matcher. Also the current implementation is unable to represent ambiguous outputs and indicate for instance that the robot could be in two different positions.

One solution would be to create 'view tuned cells', which respond to a view at a particular orientation. This could be accomplished for instance by a Gaussian basis function which have their distribution centred on a particular complex cell pattern, alternatively ECLF O-layer units could serve as view tuned cells. The position learning network could then associate the current position information with the activated view tuned cell.

6 Conclusion

The results with complex cells being recognised by nearest neighbour classifiers are encouraging. The spatial pooling cells had a minimum mean error of 5.4° measured against odometry which would have an error of about 1° when acquiring the data set. Using an output resolution of 12.6° limits the mean error to a minimum of 3.15°, so the majority of the error comes from the data and not from the classifier. This and the results for position classification indicate that robot position can be learnt from complex cell activity representations.

The ECLF network results demonstrate that neural networks can be used learn position from images represented by complex cells. The next stage is to create a network that is able to localise orientation and position, while also representing uncertainty in the localisation.

References

[Arleo *et al.*, 2001] A. Arleo, F. Smeraldi, et al. "Place Cells and Spatial Navigation based on Vision, Path Integration, and Reinforcement Learning." *Advances in Neural Information Processing Systems*, 2001.

[Chan and Wyeth, 1999] Phillip Chan and Gordon Wyeth. Self-Learning Visual Path Recognition. *Proceedings of the Australian Conference on Robotics and Automation (ACRA '99)*. Brisbane, 44-49, 1999.

[Edelman, 1991] Shimon Edelman. "A network model of object recognition in human vision." *Neural networks for perception*, vol. 1: 25-40, 1991.

[Edelman *et al.*, 1997] Shimon Edelman, Nathan Intrator, et al. *Complex Cells and Object Recognition*. NIPS 97, 1997.

[Edelman and Weinshall, 1991] Shimon Edelman and Daphna Weinshall. "A self-organizing multiple-view representation of 3D objects." *Biological Cybernetics*, vol. 64: 209-219, 1991.

[Fukushima, 2001] K. Fukushima. *Neocognitron of a new version: handwritten digit recognition*. Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on, 2001.

[Gonzalez-Barbosa and Lacroix, 2002] J.-J. Gonzalez-Barbosa and S. Lacroix. *Rover localization in natural environments by indexing panoramic images*. Robotics and Automation, 2002. Proceedings. ICRA '02. IEEE International Conference on, 2002.

[Heeger, 1992] David J. Heeger. "Normalization of cell responses in cat striate cortex." *Visual Neuroscience*, vol. 9: 181-198, 1992.

[Hubel, 1988] David H. Hubel. *Eye, Brain, and Vision*. New York, 1988.

[Kröse and Bunschoten, 1999] B. J. A. Kröse and R. Bunschoten. Probabilistic Localaization by Appearance Models and Active Vision. *IEEE International Conference on Robotics and Automation*. Detroit, vol 3: 2255-2260, 1999.

[Milford and Wyeth, 2003] Michael Milford and Gordon Wyeth. *Hippocampal Inspired models for Simultaneous Localisation and Mapping on an Autonomous Robot*. submitted to the Australasian Conference on Robotics and Automation (ACRA) 2003, Queensland Centre for Advanced Technology, Brisbane, Australia, 2003.

[Nayar *et al.*, 1994] S.K. Nayar, H. Murase, et al. *Learning, positioning, and tracking visual appearance*. Robotics and Automation, Proceedings IEEE International Conference on, 1994.

[Pourraz and Crowley, 1998] F. Pourraz and J.L. Crowley. *Use of eigenspace techniques for position estimation*. Advanced Motion Control, AMC '98-Coimbra, 5th International Workshop on, 1998.

[Prasser and Wyeth, 2003] David Prasser and Gordon Wyeth. Probabilistic Visual Recognition of Artificial Landmarks for Simultaneous Localization and Mapping. *Proceedings of the 2003 IEEE International Conference on Robotics and Automation*. Taipei, 2003.

[Riesenhuber and Poggio, 1999] Maximilian Riesenhuber and Tomaso Poggio. "Hierarchical Models of Object

Recognition in Cortex." *Nature Neuroscience*, vol. 2: 1019-1025, 1999.

[Spitzer and Hochstein, 1988] Hedva Spitzer and Shaul Hochstein. "Complex-Cell Receptive Field Models." *Progress in Neurobiology*, vol. 31: 285-309, 1988.

[Ulrich and Nourbakhsh, 2000] I. Ulrich and I. Nourbakhsh. *Appearance-based place recognition for topological localization*. Robotics and Automation, 2000. Proceedings. ICRA '00. IEEE International Conference on, 2000.

[Wersing and Körner, 2003] Heiko Wersing and Edgar Körner. "Learning Optimized Features for Hierarchical Models of Invariant Object Recognition." *Neural Comp.*, vol. 15(7): 1559-1588, 2003.