A Developing Sensory Mapping for Robots

Nan Zhang and *John Weng*

Dept. of Computer Sci. and Engr.
Michigan State University
East Lansing, MI 48823

Zhengyou Zhang

Microsoft Research
Microsoft Corp.
Redmond, WA 98052

Abstract

A sensory mapping method, called "Staggered Hierarchical Mapping (SHM)," and its developmental algorithm are described in this paper. SHM is a model motivated by human early visual pathways including processing performed by the retina, Lateral Geniculate Nucleus (LGN) and the primary visual cortex. The work reported here concerns not only the design of such a series of processors but also their autonomous development. A new Incremental Principal Component Analysis (IPCA) method is used to automatically develop orientation sensitive and other needed filters. A set of staggered receptive fields model the pattern of positioning of processing cells. From sequentially sensed video frames, the proposed algorithm develops a hierarchy of filters, whose outputs are uncorrelated within each layer, but with increasing scale of receptive fields from low to high layers. To study the completeness of the representation generated by the SHM, we experimentally show that the response produced at any layer is sufficient to reconstruct the corresponding "retinal" image to a great degree. In our experiment, we show that SHM can be used to perform local analysis.

1 Introduction

Computational visual perception faces several major challenges arising from the high dimension of pixel array and the complex relationship between the environmental factors and the pixel value. Attention in computer vision aims to select just the relevant aspects from the broad visual input, which is essential for generalization. A system cannot generalize well if it makes decision using a part of irrelevant information. The sensory mapping proposed here for developmental robots is motivated by biological sensory cortices. Although not all the functions of a sensory cortex are known, one possible use is performing local information processing based on classical and non-classical receptive field.

Identification of a local view from a field of view requires processing units whose receptive fields are located at the corresponding position of the partial view with a proper size. Since the local view can potentially be located at any position inside the field of view with any size, a visual processing system need processing elements dedicated to any position with all possible sizes. In practice, however, only computation for a finite number of positions and sizes is possible, resulting in a set of finite samples in positions and sizes of receptive fields. It is well known that the biological visual pathway consists of processing cells for receptive fields at different positions and with different sizes.

Since the measurement and study of the response of neural cells in early visual pathways by D.H.Hubel and T.N.Wiesel [1], more detailed studies have been made in modelling the complex cells and architecture organization. A single neuron by itself is not sufficient to code necessary information. An ensemble of cell population is needed. In their recent study, Stanley et al. reconstructed cat's retina sensory input from recorded LGN responses [2]. Atick and Redlich [3] proposed that the retina and LGN are dedicated to recoding and whiten the input signals. It is now known that neural cortex is organized in location and scale, but exact detail of such organization is unknown. In cortex, each neuron has a receptive field centered at a specific location, while earlier cortex has neurons with a smaller receptive field than those in a later cortex. An overview of sensory maps is available in an excellent work by Kandel, Schwartz and Jessell [4].

Is sensory mapping completely determined by human genes? The answer is negative. As early as 1970, Blakemore and Cooper [5] reported that the kittens' visual cortex does not have cells sensitive to edged orientations that they did not observe, if they lived in a controlled environment after birth, in which only edges of a certain orientation were presented. Recent studies by Sur and coworkers [6] have shown that input signal can fundamentally determine the filters generated under development. The auditory cortex of ferrets show orientation sensitive cells if they receives visual signals early in life.

Many studies have been done to model the cerebral cortex with hierarchical models and artificial neural networks. Fukushima et al., proposed a neural network named "Neocognitron" [7]. The Neocognitron is a hand designed multilayered network consisting of cascaded connections of many layers of cells. The information of the stimulus pattern given to the input layer is processed step-by-step through stages of the network. The synapses between the neurons are updated by a supervised learning method. After training, the neurons in the higher layer have the tendency to respond selectively to some complicated features despite the variance in the same feature samples. Weng et al., in 1996 proposed a dynamic neural network model called "Cresceptron" [8], which could automatically grow a hierarchical of maps directly from image input by a learning-with-teacher process. The network grows by creating new neurons, connection and architecture which memorize new image structures and context as they are detected. Although these studies address global structure of sensory mapping, the efficiency is a major challenge since those methods do not explicitly use statistical properties of the signals in the filter generation.

A number of researchers have studied the issue of dimension reduction or neural coding for a fixed receptive field. Their studies of dimension reduction or neural coding focus on automatic development of filters for a fixed receptive field. The existing work includes the random noise based method by Linsker [9], the entropy reduction method by Atick and Redlich [10], the sparse coding method by Olshausen and Field [11] and the independent component method by Bell and Sejnowski [12]. However, as we will show in the work presented here, it is beneficial to study the neural coding issues globally for a family of receptive fields, instead of one. For example, the sparse coding goal is automatically addressed in our SHM when it automatically develops a global map for a family of receptive fields.

This paper Proposes an architecture of sensory map and its developmental algorithm for active vision system where visual attention is required to analyze local image structures. The algorithm develops filters for a family of receptive fields. The responses of all filter are available for further analysis. Therefore the proposed sensory map develops a hierarchical representation for a large set of receptive fields but it does not complete vision by itself. The autonomous development of cells in the visual cortex is performed by our Complementary Candidate Incremental Principle Component Analysis (CCIPCA) method [15].

The organization of this paper is as follows. In Section 2 to Section 4 we present the structure of the proposed SHM model and its developmental algorithm. In Section 5 we show the results of some experiments with SHM. Section 6 presents a classification system in which the SHM is used for local analysis.

2 Staggered Hierarchical Mapping

When a family of receptive fields are considered, we have observed that the Principal Component Analysis is a suited mechanism for filter development. Computationally it gives a set of filters that minimize the mean square error between input space and output space, giving a fixed dimensionality under a linear projection. The response from the PCA filters are mutually uncorrelated, thus, increasing the degree of dimension reduction and facilitating later processing. Further, IPCA has some support from biological modelling of the cortex. It has been proven [13] [14] that artificial neurons updated by the Hebbian rule while being inhibited by nearby neurons (biologically called lateral inhibition) produce synaptic weights that are principal components produced by PCA.

However, the conventional PCA also has its limitation. It is applied to the whole input vector. Thus, it is not capable of extracting local features in the input space. We must design an architecture in which results of response are computed for a family of receptive fields.

Let us call the stack of filters developed for a single receptive field "neural cylinder" as shown in Fig. 1(a). We have two problems: the spatial resolution is low if receptive fields do not overlap. If they overlap, the computational cost is high. In the proposed staggered scheme, we trade off spatial resolution with a dimension of features as shown in Fig. 1(b).

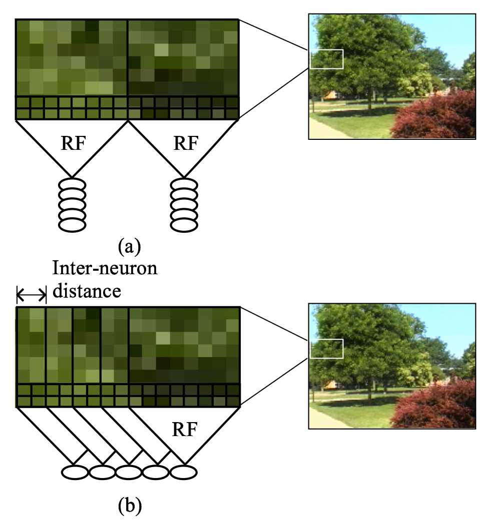

Figure 1: A comparison of the neural cylinder (a) and the staggered receptive fields (b). "RF" denotes the receptive field of a neuron. In (a) a group of filters share the same receptive fields, resulting in a low spatial resolution. In (b) each filter is centered at a different position, resulting in a higher density of spatial sampling.

The structure of the proposed SHM is similar to a multi-

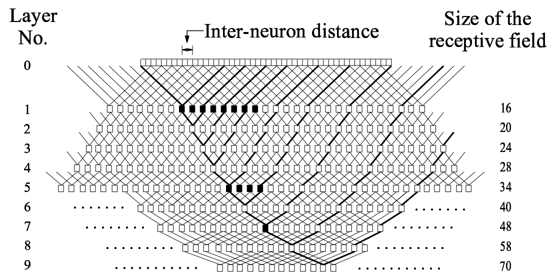

Figure 2: The architecture of SHM. Each square denotes a neuron. Layer 0 is the input image. The neurons marked as black in layer 1 belong to the same eigen-group. Bold lines that are derived from a single neuron and expanded to the original image mark the receptive field of that neuron. The size of the receptive field in a particular layer is 20% larger than its previous layer in this diagram, which is shown at the right. The size of the receptive field is rounded to the nearest integer.

layer perceptron (MLP) as shown in Fig. 2. However unlike MLP, there are feed back processing known as recursive average computation and inter-neural inhibition. Further, the developmental program of SHM is totally different from the learning algorithm of MLP. There are also other differences. The SHM is a sensory mapping method not a classification method. The former is mainly for the feature derivation and feature value computation. The SHM itself does not classify inputs. It is mainly for generating receptive fields dedicating to feature representation. Classification and regression in the SAIL robot is performed by HDR engine, using output from SHM as input.

Usually, the result of MLP is produced from the output layer. But in the SHM, every layer produces output, early ones for outputs for smaller receptive fields and later ones for larger receptive fields.

The basic computation unit, or neurons in a MLP are organized along a single dimension. The order of neurons is not taken much into consideration. However, "neurons" in the SHM are organized in a 2-D array, simulation the flat sheet organization of the retina. This order sensitive organization is very important to the receptive field concept described later.

In Multi-layer Perceptron, neurons in adjacent layer are fully connected. Every neuron is connected to all neurons in its previous and next layer. Whereas in the SHM, a localized connection is applied. Each neuron only gets its inputs from a restricted region in the previous layer. We can see this effect in Fig. 2. Fig. 2 is a simplified 1-D diagram of the proposed mapping. It is a hierarchical network in which the original input is at the layer No. 0. A small square represents a neuron. The two lines that converge from the previous layer (upper layer) to a neuron mark the small region that the neuron gets input from. This small region is referred to the input region of the corresponding neuron. The input region can be traced back to the original input and will include a partial original image, as shown by bold lines in Fig. 2. This partial image is called the (classical) receptive field of the neuron. As we will see later, lateral inhibition implemented by our residual image makes the response from a neuron to be affected by a much larger region in the input image. This later larger region is called non-classical receptive field. By receptive field, we mean classical receptive field by default. Apparently, the sizes of the receptive fields of different neurons are identical within a layer. The concept of receptive field is important in discussing neuron development in the visual cortex.

The use of staggered receptive fields allow for a neural network to be built that spans the entire input region. The input can get an alternative representation that covers different scales and positions. As we can see from the diagram, the position of receptive fields is kept organized and the size of the receptive fields grow incrementally over each layer until the entire image input is covered with a single receptive field (The receptive field of the black neuron in the 7th layer covers the whole input). Furthermore, for the input region at any position with any size, a neuron can be found whose receptive field approximately covers the region. Then intuitionally, this representation could provide information for different scales and locations.

Besides the differences in network structure, the SHM and MLP also differ in learning methods. Unlike the multi-layer Perceptron's supervised learning method, the learning method of the SHM is a self-organized unsupervised learning. The purpose of learning is to gradually update connection weights or filters to get a satisfying response output. Filters are self-developed (updated) by the CCIPCA method [15], and eventually will converge to an eigenvector of the input region without a need to compute the correlation matrix of the input vector. The neurons are first partitioned into what we call eigen-groups. An eigen-group is a square whose size is $n * n$. The value n is determined by the size of input region of this layer, so that n is the maximum possible distance between any two neurons in the eigen-group to have their receptive field overlapped. Thus, neurons at a layer that is beyond n distance apart do not need to inhibit each other, since their receptive fields do not overlap. The main purpose of inhibition in biological network is to enforce nearby neurons to detect different, preferably statistically uncorrelated features. Thus, the input region of all the neurons that are at the same position of different eigen-groups can be tiled to an entire previous layer. Table 1 shows an example of an eigen-group whose size is 4×4. Referring to layer No. 5 in Fig. 2, the nearby black neurons are in the same eigen-group, in which neu-

15	4	1	6
3	8	9	10
12	13	7	14
2	5	0	11

Table 1: Order of computation in an eigen-group. The size of this eigen-group is 4×4. The numbers in the table denote the order of updating using residual image.

rons have at least one overlapping input. The neurons at the same position in all different eigen-groups, say the first neruon of all the eigen-groups, covered the entire input region using their receptive fields.

The neurons in an eigen-group are given a randomly predefined order shown by a number in Table 1. The same order is applied to all eigen-groups in the same layer, for simplicity. The first order neuron is the one who gains an "upper hand" during inhibitive competition; the second, next, and so on.

3 Developmental Algorithm

The developmental algorithm for the SHM must be incremental and fast for real time application. By "incremental", we mean that each input image must be discarded after been used for updating the SHM. It is simply not practical to store all images for batch processing due to the extreme large amount of space required. By "fast", we mean that the updating for each image must be completed within a second. For example, iteration within each updating is not allowed. These requirements limit the complexity of the computation that can be performed by a biological cortex. They also make the design of a developmental program for SHM very challenging.

We have developed CCIPCA [15] that was designed with these requirements as design constrains.

The filter development procedure of a single layer is described as follows:

1. Set i=0; copy original input to a buffer.

2. For all eigen-groups update ith filters as the ith principal component using the CCIPCA algorithm with input from buffer; compute the responses of the ith filter; compute residual image by subtracting the projection of the ith filter from the input buffer.

3. Tile residual images back into the buffer, one tile for each eigen-group.

4. i=i+1; goto step 2, until all filters are updated

The output response images are then passed on to the next layer of the network. Each layer is similar in that it takes its input and runs the CCIPCA algorithm on it, producing both filters for this layer and output response image for the next layer. The first layer of the network uses real input, and each subsequent layer uses output response images from the previous layer.

The receptive field in the same level is not only concentrated in several center positions. Here we have used a staggered receptive field method to spread the receptive fields all over the input layer. Then, the staggered positions along with the increasing size of the receptive field make it possible that input regions with any size at any position (up to a sampled resolution) have a local representation in the network.

Inhibitions between neurons play a vital role in the function of the visual cortex. We use residual image to represent the effect of inhibitive competition between neurons. This SHM network is not simply feed-forward, because it deals with the re-entrant process of neurons. Each time a filter is computed, it is subtracted out from the original image, leaving only information that has not yet been extracted by the CCPCA algorithm (residual image). The next filter can only "see" input part that the previous filters can not extract, thus, the first can inhibit the latter making the process asymmetric. Each neuron "sees" unique input data that has not been "seen" by previous neurons. This is more efficient than symmetrical inhibition where it takes time to fight out the rank of competition, and achieves the same result. Fig. 3 shows a sequence of residual images. It can be seen that the variance of the series residual images is decreased sequentially as more orthogonal components are extracted. Because the residual image is orthogonal to the previous eigenvectors (filters), the later eigenvector is orthogonal to all previous ones. Therefore the responses of filters are statistically uncorrelated.

Figure 3: Sequence of residual images. From left to right and top to bottom, starting from the original input image, the sequence of images displays the process that a component is subtracted from the previous one.

4 Attention Effectors

The response of every neuron (or processing cell) at any level is available to be used for later processing. The goal of attention effectors is to suppress the response (value) from areas that are beyond the attended visual field.

However, with staggered receptive fields, shown in Fig. 2, the neurons share input plane among them. This means that any region of neurons in a layer of SHM does not correspond a clear-cut region in the input plane. A special case of this situation happens when this region is so small that it contains only one neuron, the corresponding larger non-classical receptive field and the smaller classical receptive field demonstrate that no neuron corresponds to a clear cut receptive field. Interactions between neurons expand the extent of each pixel beyond that marked by direct inputs to neurons. Therefore, we should not expect that selecting the response from a region in any layer will result in a clear-cut attended region in the input.

We define attended region as a 3-D ellipsoid centered at a position (x,y,l) in a layer of SHM, where, x, y, denotes the position of the neuron and l denotes the layer. The length of each axis of the ellipsoid corresponds to the scale along the dimension. All the neurons falling into the ellipsoid have their output passed on and all others blocked.

5 Experiment with SHM

Testing began with a collection of over 5000 natural images. They were taken using a Sony digital camcorder on the campus of Michigan State University. A variety of recordings were taken, including buildings, trees and shrubbery, closeups of bark and wood chips, flowers, and other various scenes found in nature and on campus. The frames from the video were then captured and digitized into 118×158 images, each has 118×158 pixels. Fig. 4 shows a few sample images collected.

Figure 4: Several samples of 5000 natural images collected.

There are two methods to develop filters, sharing or non-sharing. The sharing method assumes that the filters that are located at the same location in eigen-groups have the same form. This is a good estimation only if the images have the same statistical properties across the entire image frame. Thus, the sharing method only updates a single set of filters for all eigen-groups. The non-sharing method updates a separate set of filters for each eigen-group.

The filters of the first layer trained by the sharing method is shown in Fig. 5. The eigen-group, in this case, has a size of 8×8, and the size of each filter at this layer is 16×16. The filters are reordered according to its computation order, so that the dominant principal components are shown first. As can be seen from the figure, the first several filters appear similar to the receptive field excitation-inhibition patterns reported by the Hubel and Wiesel's experiment [1].

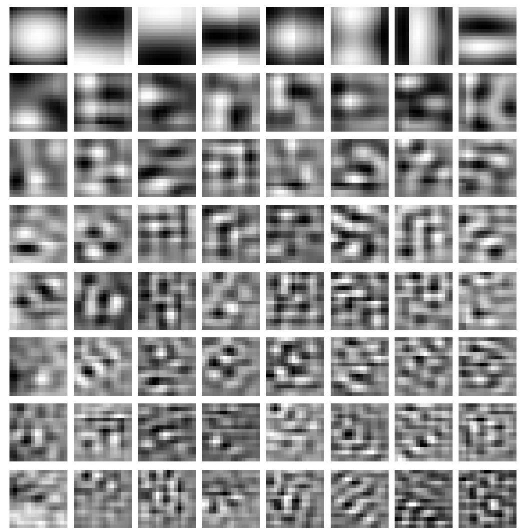

Figure 5: The filters developed in the first layer by the sharing method. The order of dominance is from left to right and from top to bottom.

Fig. 6 displays the first four filters in the first layer developed by the non-sharing method. The filters with the same order are tiled one by one according to its position. It is worth mentioning that the filters near the boundary of the image are trimmed off for better visualization. When a filter falls out the border of the input image, the corresponding pixel is assumed zero.

When input was put into the network, a scalar output response value was generated for each filter at its respective receptive field. The output response is based on the dot product between an eigenvector and its input vector with possibly an post nonlinearity such as sigmoidal function. This measures the amount of energy of input along the direction of the filter and represents the firing strength of a neuron. The filter responses were then organized into groupings based on their order in the eigen-group and then combined into a structured image vector and passed on as input to the next layer of the network. An example input image and its corresponding output response image are shown in Fig. 7. A strong response is represented by either white or black high contrasting pixels depending on the sign. Areas of low response are represented by gray

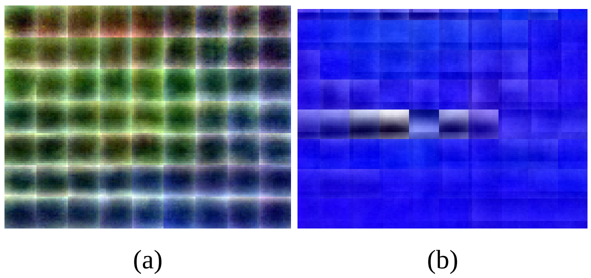

Figure 6: The first two filters developed by the non-sharing method. (a) The most dominant filter for different eigen-groups. (b)The next dominant filters. The filters that are at the same position of different eigen-groups are put together.

pixels.

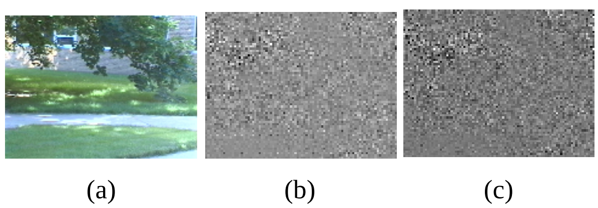

Figure 7: Responses in different layers. (a) The original input image. (b)-(c) The response outputs from layer 2 to layer 3 respectively.

With SHM the input image is coded as responses at several layers. To show how complete the representation is at each single layer, we reconstruct the original image from the response of that layer alone. The reconstruction is a reverse processing of the mapping. The responses of a certain layer, which are actually the sample's projection along the principal components, are used to recover its component in the sample space, then add back to the residual image. The responses of layer are reconstructed by its following layer, finally reaching the original input layer. It is interesting that the quality of the reconstructed image does not decrease significantly from layer to layer. It can be seen from Fig. 8 that all reconstructed images are close to the original image. Mathematical analysis of this issue is currently underway. Our experimental results showed that the representation at each layer of the SHM is almost complete, from lower layer for smaller receptive fields and higher ones for large receptive fields.

6 Experiment with Occlusion

The purpose of the experiment summarized here is not to show a practical application system, but rather to study a use of SHM as a local representation. In a future system that uses SHM as a sensory "cortex," the parallel outputs of

Figure 8: The reconstructed images from different levels. The first image is the original input image. Others are reconstructed images form Layer 1 to Layer 5 respectively.

all neurons in all layers are used by higher level processing that generates attention signals for the SHM. In the current experiment, we use a regressor C3 in Fig. 9 that learns the mapping from input to the attention control signals needed by the SHM at anytime. It is trained first.

The experimental setting was organized as follows: In the training phase, a series of global view of face images is presented to the system with class labels. In the testing phase, different face images of the same set of people are presented to the system, but all of them are partially occluded: the upper view (U view) of a face has the lower third area replaced by a uniform intensity (gray), and the lower view (L view) has the upper third replaced by gray.

If a system is passive, it learns the global view and tries to match the input U or L view with the learned global prototype. If we use nearest neighbor method (NN) to find the prototype, we call it monolithic + NN testing.

We would like to study the use of active internal action in visual perception. For this purpose, we constructed a active system shown in Fig. 9. Three classifiers C1, C2 and C3 are integrated in the system. Here, we use Hierarchical Discriminant Regression Tree (HDR Tree) [16]. To learn the mapping, C3 is to determine what specific occlusion case the image is, U or L. In our test, the classifier C3 has reached 100% classification rate.

The S1 and S2 are SHMs. S1 is for U views and S2 is for L views. C1 and C2 are two HDR Trees to classify images. Which SHM + HDR should be used is determined by the attention signal from C3.

The experiment used face set from Weizmann Institute at Israel. The set was taken from 28 human subjects, each having 30 images with all possible combinations of two different expressions, three lighting conditions and five different facial orientations.

We used the leave-one-out cross-validation method. 30 experiments were conducted, in which 1 out of 30 samples of each person is picked up as testing sample, and all

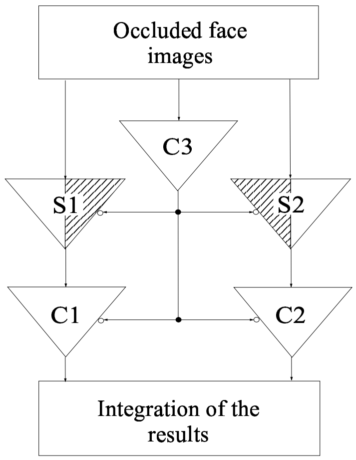

Figure 9: The schematic illustration of operation. C3 is the attention signal generator. C1 and C2 are the classifiers for each occlusion case. S1 and S2 are SHMs, S1 for U views and S2 for L views.

Method	Recognition Rate		
	U	L	U+L
Monolithic+NN	51.43%	75.83%	82.38%
SHM+HDR	92.86%	95.95%	98.57%
Method	Testing Time (ms)		
	U	L	U+L
Monolithic+NN	765.5	765.5	2263.0
SHM+HDR	702.4	702.4	2016.6

Table 2: Summaries of the occlusion experiment.

remaining ones as training samples. The average of the recognition result is reported in Table 2, along with the time run on a dual Pentium III 600 MHz processor PC. U, L columns are for testing with U, L views only. U+L column is for an integration of 2 test views, U and L to give a single classification (person). The U+L integration is done in the following way. For each view, top K best matches form a top labels list. Thus, each pair of U view and L view gives a U label list and a L label list. The label that has the minimum rank sum is the output label.

The Table 2 shows that the proposed SHM with HDR classifier is effective in integrating active partial views, U and L. Without active attention, monolithic + NN performed poorly, even with U+L integration. This is because it lacks a mechanism to actively extract local views when presented with a global view. The speed of the SHM + NN is fast enough to reach about 2 Hz refreshing rate.

7 Conclusions

This paper presents the design principles, the architecture and the developmental algorithm of a general purpose sensory mapping called the Staggered Hierarchical Mapping. The experimental data have shown that the output response is effectively decorrelated for receptive fields at a family of positions and scales. The representation at each layer has a high completeness. The test result of the occluded image recognition system shows that the SHM has a local analysis property, thus, it can be used to achieve attention control in an active vision system. As far as we know, this is the first incremental computational model for a developing hierarchical sensory map with a complete family of receptive fields.

8 Acknowledgements

The work is supported in part by National Science Foundation under grant No. IIS 9815191, DARPA ETO under contract No. DAAN02-98-C-4025, DARPA ITO under grant No. DABT63-99-1-0014, and a research gift from Microsoft Research.

References

[1] D. Hubel and T. Wiesel, “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex,” *J. of Physiology* **160**, pp. 106–154, 1962.

[2] G. B. Stanley, F. F. Li, and Y. Dan, “Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus,” *Journal of Neuroscience* **19(18)**, pp. 8036–8042, 1999.

[3] J. J. Atick and A. N. Redlich, “What does the retina know about natural scenes?,” *Neural Computation* **4**, pp. 196–210, 1992.

[4] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, eds., *Principles of Neural Science*, Appleton and Lance, Norwalk, Connecticut, third ed., 1991.

[5] C. Blakemore and G. F. Cooper, “Development of the brain depends on the visual environment,” *Nature* **228**, pp. 477–478, Oct. 1970.

[6] M. Sur, A. Angelucci, and J. Sharm, “Rewiring cortex: The role of patterned activity in development and plasticity of neocortical circuits,” *Journal of Neurobiology* **41**, pp. 33–43, 1999.

[7] K. Fukushima, S. Miyake, and T. Ito, “Neocognitron: A neural network model for a mechanism of visual pattern recognition,” *IEEE. Transactions on SMC* **13(5)**, pp. 826–834, September/October 1983.

[8] J. Weng, N. Ahuja, and T. Huang, "Learning recognition and segmentation using the cresceptron," *International Journal of Computer Vision* **25(2)**, pp. 109–43, 1997.

[9] R. Linsker, "From basic network principles to neural architecture: emergence of spatial-oponent cells," *Proc. Natl. Acad. Sci.* **83**, pp. 7508–7512, 8390–8394, 8779–8783, Oct. 1986.

[10] J. J. Atick, "Convergent algorithm for sensory receptive field development," *Neural Computation* **5**, pp. 45–60, 1993.

[11] B. A. Olshaushen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy used by v1?," *Vision Research* **37**(23), pp. 3311–3325, 1997.

[12] A. J. Bell and T. J. Sejnowski, "The "independent components" of natural scenes are edge filters," *Vision Research* **37**(23), pp. 3327–3338, 1997.

[13] J. Rubner and K. Schulten, "Development of feature detector by self-organization," *Biological Cybernetics* **62**, pp. 193–199, 1990.

[14] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural Networks* **2**, pp. 459–473, 1989.

[15] Y. Zhang and J. Weng, "Complementary candid incremental principal component analysis," Tech. Rep. MSU-CSE-01-24, Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, August 2001.

[16] W.-S. Hwang and J. Weng, "Hierarchical discriminant regression," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **22**, pp. 1277–1293, November 2000.