A Hierarchical Bayesian Model of Invariant Pattern Recognition in the Visual Cortex

Dileep George
Department of Electrical Engineering
Stanford University and
Redwood Neuroscience Institute
Menlo Park, CA 94305
E-mail: dil@stanford.edu

Jeff Hawkins
Redwood Neuroscience Instiute
Menlo Park, CA 94305
E-mail: jhawkins@rni.org

***Abstract*— We describe a hierarchical model of invariant visual pattern recognition in the visual cortex. In this model, the knowledge of how patterns change when objects move is learned and encapsulated in terms of high probability sequences at each level of the hierarchy. Configuration of object parts is captured by the patterns of coincident high probability sequences. This knowledge is then encoded in a highly efficient Bayesian Network structure.The learning algorithm uses a temporal stability criterion to discover object concepts and movement patterns. We show that the architecture and algorithms are biologically plausible. The large scale architecture of the system matches the large scale organization of the cortex and the micro-circuits derived from the local computations match the anatomical data on cortical circuits. The system exhibits invariance across a wide variety of transformations and is robust in the presence of noise. Moreover, the model also offers alternative explanations for various known cortical phenomena.**

I. Introduction

Recognizing objects despite different scalings, rotations and translations is something humans perform without conscious effort, but this still is a hard problem for computer vision algorithms.

We believe that the geometric invariances that humans so effectively handle are intimately linked to the motion in this world. When we move in this world while still looking at an object, the patterns that fall on our retina change continuously while the underlying cause for those patterns - the object itself - remains the same. Rigid objects have the property that they produce the same change of patterns for the same pattern of motion. Rigid objects in this world can be thought of as the underlying causes for persistent patterns on our retina. Thus, learning persistent patterns on the retina would correspond to learning objects in the visual world. Associating these patterns with their causes corresponds to invariant pattern recognition.

In this model we use many concepts which are familiar and accepted in neuroscience and computer vision. It well known that the visual cortex is organized in a hierarchy and several models of invariant pattern recognition [6][16] make use of this. Temporal slowness has been shown to be a plausible criterion for learning invariances [19] and our idea of most likely sequences can be related to this. We derive our architecture and algorithms based on the idea that the goal of the cortex is to make predictions [7]. Predictive models [14] can explain the role of feedback connections in the cortex. However there are no predictive models available in the literature that do invariant pattern recognition as well. The framework of ideas that we use here was described in [7] and we consider that as the starting point of the work we describe here.

The rest of this paper is organized as follows. In section 2, we describe our system architecture and in section 3 the learning algorithm. In section 4 we describe how the system performs invariant pattern recognition. In section 5 we connect the architecture and algorithms to biology. In section 6 we describe the simulation setup and performance results. Our model provides alternative explanations for some cortical phenomena and these are explained in section 7. We conclude the paper in section 8 with a discussion on related work.

II. Architecture and Assumptions

The system we describe here is organized in a hierarchy and our learning and recognition algorithms exploit this hierarchical structure. Each level in our system hierarchy has several *modules*. These modules model cortical regions. A module can have several children and one parent. Thus the modules are arranged in a tree structure. The bottom most level is called level 1 and the level number increases as you go up in the hierarchy. Inputs go directly to the modules at level 1. The level 1 modules have small *receptive fields* compared to the size of the total image, i.e., these modules receive their inputs from a small patch of the visual field. Several such level 1 modules tile the visual field, possibly with overlap. A module at level 2 is connected to several adjoining level 1 modules below. Thus a level 2 module covers more of the visual field compared to a level 1 module. However, a level 2 module gets it information only through a level 1 module. This pattern is repeated in the hierarchy. Thus the receptive field sizes increase as one goes up the hierarchy. The module at the root of the tree covers the entire visual field, by pooling inputs from its child modules. The set of level 1 modules can be considered analogous to V1, the set of level 2 modules analogous to V2 and so on.

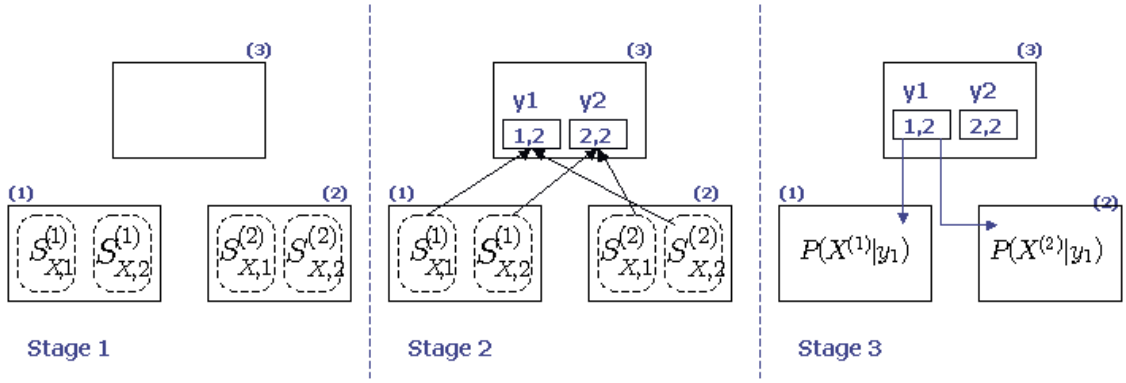

Fig. 1. Learning Stages: Learning starts at the bottom of the hierarchy and proceeds to the top. The modules at the very top of the hierarchy receive their inputs from a small section of the visual field. During Stage 1, these modules observe their inputs in time and learn the *most likely sequences* of a particular length of inputs. Once stage 1 learning is finished,these modules start passing up the index of the sequence whenever they observe one of the most likely sequences at its inputs. A higher level module gets its inputs from several lower level modules. During Stage 2, the higher level module learns frequent coincidences of sequence indices. These become the alphabets or concepts for the higher level. Note that this alphabet abstracts what patterns occur together in space and time. During the third stage of the learning procedure, the higher level concepts are fed down to the lower regions so that they learn the occurrences of the lower level patterns in the context of the higher level concepts. Repeating this in a hierarchy we obtain a graphical model as shown in figure 1

III. Learning Algorithm

We describe our learning algorithm taking a two level hierarchical arrangement as shown in figure 1 as the example. The inputs to the system are given to the modules at the bottom most level. Let the random variable prefix X indicate all the the inputs to level 1 modules. Let $\{X_n^{(1)}\}$ and $\{X_n^{(2)}\}$ denote the sequence of inputs to modules 1 and 2 in figure 1.

Learning in this model occurs in three stages. During the first stage of learning, a module learns the most likely sequences of its inputs. Let $B_\delta^{(l)} = \left\{S_{X,1}^{(m)}, S_{X,2}^{(m)}, \cdots, S_{X,N}^{(m)}\right\}$ be the set of sequences of length l with their fraction of occurrences greater than δ. A module learns this set empirically by observing its sequence of inputs. Once a module has learned $B_\delta^{(l)}$, any high probability sequence seen by this module can be uniquely represented by its index k into the set $B_\delta^{(l)}$. At the end of learning stage 1, a module has learned $B_\delta^{(l)}$ and it produces at its output the index of the high probability sequences that it observes on its input.

A module enters the second stage of the learning process once its child modules have finished the first stage of learning and is communicating with this module in terms of the indices of the high probability sequences of those modules. Lets consider module number 3 at the second level in figure 1. The input to this module consists of the concatenation of the outputs from its child modules 1 and 2. A particular concatenation represents a simultaneous occurrence of a combination of high probability sequences in the child modules. Depending on the spatio-temporal statistics of the inputs seen by the lower level modules, some of these coincidences will occur frequently and some will not. During the second stage of learning, a parent module learns the most frequent coincidences (according to an ϵ criterion) of sequences in the levels below it. We denote the most frequent patterns at this level 2 module by Y and the number of such patterns by M. These patterns become the alphabet for this module.

The third stage, called contextual embedding, involves feedback from the level 2 module to its child modules to embed the the lower level patterns in the context of the higher level patterns. This stage is initiated once the level 2 module has formed its alphabet Y as we described above. Assume that at a particular point in time the higher level pattern $Y = y_k$ is active. (This pattern was made active by the simultaneous occurrence of a combination of sequences in lower levels). This information, i.e., the index of the high level concept, is fed back to the level 1 modules. This information is used by the level 1 modules to obtain a conditional probability distribution (CPD) matrix of its patterns given the patterns at a higher level. During the learning process, this CPD matrix is updated by incrementing the count for all level 1 patterns that were part of the sequence which caused the high level pattern y_k. At the end to the learning process, the rows of this matrix are normalized to obtain the conditional probability distribution $P(X^{(1)}|Y)$ for module 1 at level 1. This process is identical for all the modules at level 1.

The learning process defined above can be repeated in a hierarchy. This is done by considering the frequent spatial patterns seen by a module at any level to be the alphabet of that region and then repeating stage 1, 2 and 3 of the learning algorithms in a manner identical to the description above. In our example, the learning can be continued between levels 2 and 3 by considering the frequent spatial patterns Y of the level 2 module as its alphabet and then learning the high probability sequences on this alphabet to continue to stages 2 and 3 of the algorithm.

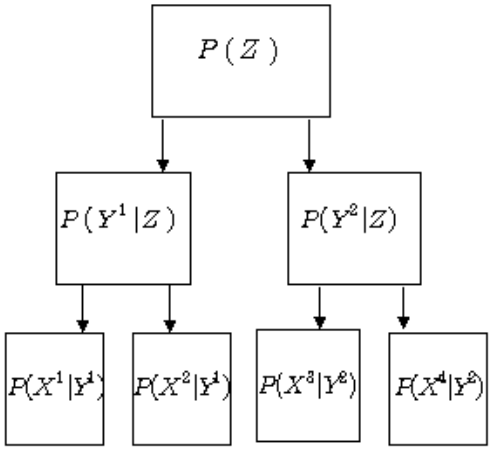

Fig. 2. Structure of a typical Bayesian Network obtained at the end of the learning procedure. The random variables at the bottom level nodes correspond to quantizations of input patterns. The random variables at intermediate levels represent object parts which move together persistently. The random variables at the top node correspond to objects. During training the definitions of the intermediate object-parts and then the top-level objects are obtained using algorithms described in figure 1. The probability tables are also filled according to Stage 3 of figure 1.

IV. Recognition as Inference In a Bayesian Network

Once all the modules in the hierarchy have learned according to the algorithms described in section 3, we get a tree structured Bayesian Network [13], an example of which is shown in figure 2. The modules correspond to the nodes in a probabilistic graphical model and each node stores a conditional probability distribution. Every module can be thought of as having a set of states. The CPDs at each node encode the probability distribution of the states of that module given the state of its parent module.

If we assume that the learning algorithm has produced meaningful states at each module of the hierarchy with the states at the top of the hierarchy corresponding to object categories, then the recognition problem can be defined as follows. Given any image I, find the most likely set of states at each module, the combination of which best explains the given image in a probabilistic sense. Specifically, if $\mathbf{W}$ is the set of all random variables corresponding to node states, then the most probable explanation for the image I is a set of instantiations $\mathbf{w}^\star$ of the random variables such that

$$P(\mathbf{w}^\star|I) = \max_{\mathbf{w}} P(w|I) \tag{1}$$

If Z is the random variable representing the states at the top of the hierarchy, then the category label that best explains any given image is the index of $z^\star$, where $z^\star$ is the assignment to Z which maximized the above equation. Its a well known result that given an acyclic Bayesian Network as the one we have here, inference can be performed using local message passing.

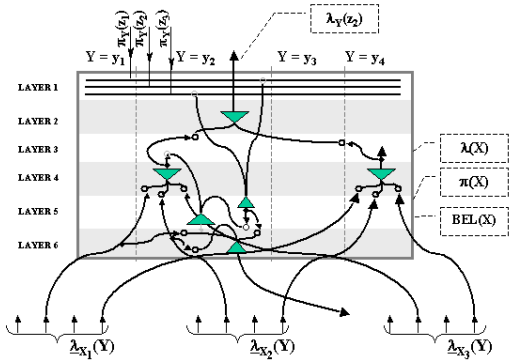

Fig. 3. Belief Propagation and Cortical Anatomy: The belief propagation equations that we used for inference in our model has an anatomical mapping which matches anatomical data [18] to a large extent. Shown here is the cortical circuit resulting from such a mapping. This mapping enabled us to replicate some of the physiological experiments in our system.

We use Pearl's Bayesian Belief Propagation algorithm [13] to obtain the most likely explanation given an image.

V. Connection to Biology

It is well known that cortical system is organized in a hierarchy and by virtue of the connections, some regions are hierarchically above some other [3]. Moreover, it is well known that the receptive field size increases as you go up in the hierarchy. It is generally accepted that neurons in the higher level of the visual cortex represent more complex features with neurons/columns in IT representing objects or object parts. The lateral connections within layer 2-3 of the cortex and the connections between layers 1,2 and 3 through the thalamus could provide adequate mechanisms for learning of sequences [7]. Thus the large scale organization of our system is in agreement with the structure of the visual cortex.

We also found a fine mapping of these algorithms to the cortical anatomy by mapping the Bayesian Belief Propagation (BBP) [13] equations to a neural instantiation. A cortical region can be thought of as encoding a set of concepts in relation to the concepts encoded in regions hierarchically above it. The set of concepts encoded by a region can be thought of as a random variable. A cortical column represents a particular value of this random variable. At every time instant, the activity of a set of cells in a column represents the probability that a particular hypothesis is active. The feed forward and feed back connections to a cortical region carry the Belief Propagation messages. Observed information anywhere in the cortex is propagated to other regions through these messages and can alter the probability values associated with the hypotheses maintained by other regions. Figure 3 shows the detailed cortical micro-circuitry derived from BBP equations. The anatomical details of this circuit match the known anatomical data [18] to a great extent. The BBP equations that we used for deriving this micro-circuit is given as part of the appendix.

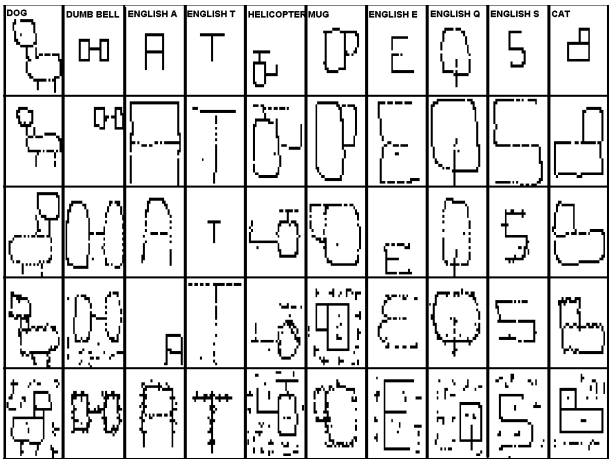

Fig. 4. **Recognition:** Shown here are examples of test images that the system could recognize correctly along with their labels. The system shows very robust scale, translation and distortion invariance works well with very noisy inputs. Note that some patterns (table lamp, dog) are recognized irrespective of their orientation. The invariances developed in the system are the ones to which the system is exposed to during the training phase. Our system has the feature that small eye-movements during the recognition stage improves performance. With eye-movements we have a recognition accuracy of 97 percent for viewer drawn images.

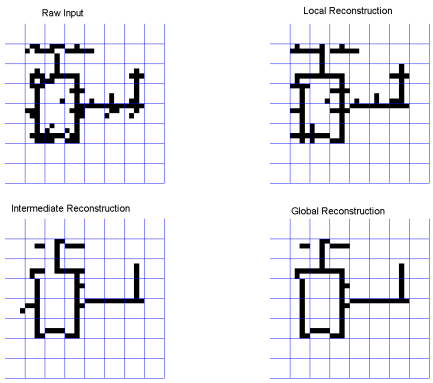

Fig. 5. Prediction/Filling-in: This experiment demonstrates the predictive capabilities of the system. The raw input (top left) is very noisy and an immediate reconstruction using the information in a 4x4 window has all the features wrong (top right). The intermediate reconstruction (bottom left) is obtained by operating the belief propagation till the second level in the hierarchy and then passing the beliefs down to the lower level again. Thus the intermediate level reconstruction the statistics of patterns in an 8x8 neighborhood. The global reconstruction (bottom right) is obtained by doing the belief propagation globally. This reconstruction is consistent with the recognition of the input as a helicopter.

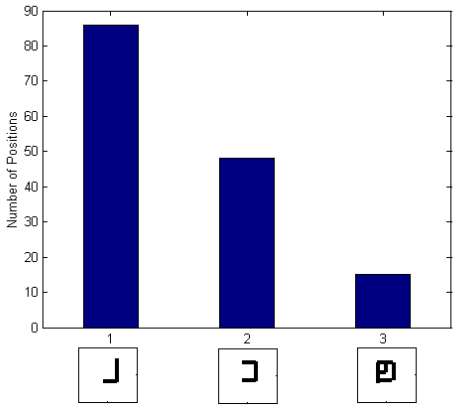

Fig. 6. In this experiment we showed the system snapshots of 3 novel images at 10 randomly chosen positions. What is plotted is the number of positions to which the system generalized for each of these novel images (shown along the X axis). It can be seen that the generalization performance goes down with increasing complexity of the novel pattern.

VI. Simulation Setup and Results

We simulated the above algorithms for a data set of line drawing movies. These movies were created by simulating straight-line motions of line drawings of objects belonging to 91 classes. There were 10 exemplar images of different scales for each category. Each image was of size 32 pixels by 32 pixels. The movies were created by picking a random category index and then moving the picture belonging to that category in straight lines. Once a particular direction of motion was picked, the object moved in that direction for a minimum of 10 time steps before changing directions. An object that was picked remained in the field of view for at least 30 time steps before a new object category was picked at random. This way of simulating a movie gives us an infinitely long input sequence to verify various performance aspects of the algorithms described above. We describe the results of these investigations in the following subsections.

All these simulations are based on a hierarchical arrangement of modules as described in section 2. The system consisted of 3 levels. The lowest level, level 1, consisted of modules receiving inputs from a 4x4 patch of images. Sixty four level 1 modules tiled an input image. Learning started at L1 and proceeded to the higher levels. A level 2 module received its inputs from 4 adjoining level 1 modules. There were a total of 16 level 2 modules. A single level 3 module received all the information from these level 2 modules.

A. Recognition, Prediction and Generalization

The full network was trained up according to the algorithm described in section 3. Recognition is performed according to the inference procedure outlined in section 4. An input image to be recognized is converted to uncertain evidence using a hamming distance metric on each module (at the level of 4x4 pixels) as described in section 4. Recognition is defined as obtaining the most likely explanation (MPE) of the evidence given the conditional probabilities that we learned on the graph. We used Pearl's Bayesian Belief Propagation algorithm for inference [13].

The system exhibited robust recognition performance invariant to large scale changes, deformations, translations and noise. Figure 4 shows examples of correctly recognized images. Note that some categories are recognized irrespective

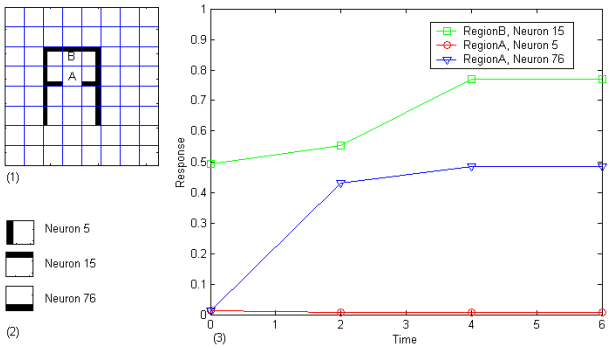

Fig. 7. Neurons Responding to Illusory Contours/Contour Continuation: Such neurons were observed in V1 [9]. Here we show the results of an experiment which demonstrates analogous results. Illusory contours are the result of the higher levels imposing its knowledge of higher level structures on to the lower levels. To test this we deleted a small portion of a familiar pattern and gave that as the input to the system. This pattern (an incomplete *a*) is shown in the figure. We then recorded the activities of neurons in regions marked A and B as a function of time. The image is shown to the system at $t = 0$. Neuron 15 in of region B shows a robust response at $t = 0$ because this region receives a perfect input that is tuned to neuron 15. Whereas, neuron 76 of region A does not show any response at this time. At time $t = 2$ the information has propagated one level up and has propagated back down. This forces region A to change its current belief about its state, thus increasing the activity of neuron 76. At $t = 4$, the global feedback information reaches all level 1 regions and for region A, this increases the belief in neuron 76. Note that the pattern corresponding to neuron 76 correctly fills the missing portion of the input pattern. Neuron 15 is an example of a neuron in region A whose activity was not affected by the feedback information. At $t = 4$, all regions have received feedback from everywhere and hence the responses do not change after this point.

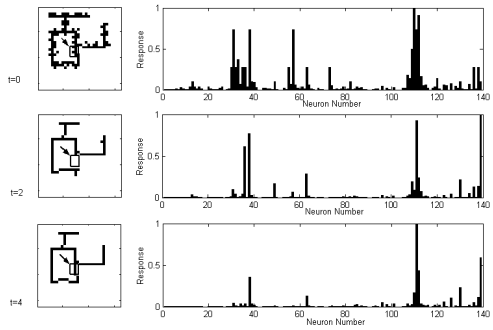

Fig. 8. Shape perception reduces activity in lower levels [10]: Our model offers an alternative explanation to this phenomenon compared to the subtraction theory [10]. Reduction in activity occurs because incorporating global information narrows the hypotheses space maintained by a lower level region. In this experiment, we showed our system a highly noisy picture of a *helicopter* and recorded the activity of the cells which represent the current belief in a rectangular Level1(V1) region(pointed by the arrow). At $t = 0$, the input is highly ambiguous as shown and hence the belief of the region is highly spread out. At $t = 1$ the level 2 regions integrate the information from multiple level 1 regions and feed back information to level 1 regions. At $t = 2$, the level 1 region uses this information to update its belief. Figure shows that this reduces the spread of the belief as compared to $t = 0$. The corresponding picture of the helicopter is the reconstruction at this stage if you take the best guesses from all level 1 regions. At $t = 4$, the level 1 regions get feedback which incorporates the global information. This further narrows the posterior distribution. Note also that the reconstruction at this point is the correct one.

of a flip about the vertical axis. This is because for those categories, we included sequences which had a flip as a part of the training sequences. This shows the capability of the algorithm to learn the transformations it is exposed to during the training. If the input is ambiguous, the cortex can gather further information from the input by making small eye movements. Many psycho-physical studies show that recognition performance is greatly hampered when eye movements are not allowed [11]. Between the eye-movements, the correct hypothesis would remain stable while the competing incorrect hypotheses will vary in a random manner. Our system exhibits this property and we used it to improve the signal to noise ratio during recognition.

A critical test for whether a system generalizes correctly would be to test whether it can correct noisy or missing inputs using its knowledge about the statistics of patterns. We tested this for our system and the results are shown in figure 5.

We also tested that our system generalizes well when trained on novel patterns. Generalization occurs in our system due to the hierarchy. Objects are made of the same lower level components. After having seen many images, the lower levels of the system have seen everything that is there (*sufficient statistic*) in the visual world at the spatial and temporal scales of those modules. Thus if a new pattern is to be learned, most of the lower level connections do not need to be changed at all for learning that pattern. Figure 6 shows the generalization performance of the system in learning new patterns.

VII. Alternative Explanations for Biological Phenomena

Some physiological experiments [9] found that neurons in V1 of the visual cortex respond to illusory contours in a Kanizsa figure. This means that the neuron is responding to a contour that does not exist in its receptive field. Another way of interpreting this is that the activity of the neuron represents the probability that a contour should be present in its input, given its own input and the contextual information from above. We found such neurons in our model using the anatomical mapping we described in section 5. See figure 7 for the results of our experiment.

Functional MRI studies [10] report that the perception of an object in the Infero Temporal cortex reduces the activity in lower levels of the hierarchy. We could observe this in our model and we offer a Bayesian explanation for this phenomenon as opposed to the current subtraction hypothesis [10]. See figure 8 for details.

VIII. Discussion

Invariant pattern recognition has been an area of active research for a long time. Earlier efforts used only

the spatial information in images to achieve invariant representations[6][16][15]. However performance of these systems was limited and generalization questionable. We believe that continuity of time is the cue that brain uses to solve the invariance problem [5], [17]. Some recent models have used temporal slowness as a criterion to learn representations [8], [19], [2]. However those systems lacked a Bayesian inference-prediction framework [9] and did not have any particular role for feedback.

Our model captures multiple temporal and spatial scales at the same time. This goes beyond the use of Hierarchical Hidden Markov Models (HHMMs)[4] to capture structure at multiple scales either in space or in time. Several other models [1], [12] attempt to solve the invariance problem by explicitly applying different scalings, rotations and translations in a very efficient manner. However, as our test cases in section 4 indicate, none of the novel patterns we receive are pure scalings or translations of stored patterns.

In our current system, sequence information is used only during the training stage to form concepts at intermediate levels. Future work will include methods for preserving this sequence information so that the system can predict forward in time. The current model deals only with the ventral visual pathway of the cortex. Dealing with the dorsal pathway will require integrating motor information with visual information. This is also part of future work.

APPENDIX: BAYESIAN BELIEF PROPAGATION EQUATIONS

The following equations, adapted from [13] were used for the derivation of the circuit shown in figure 3.

$$\lambda(y_k) = \prod_j \lambda_{X_j}(y_k) \quad (2)$$

$$\pi(y_k) = \sum_z P(y_k|z)\pi_Y(z) \quad (3)$$

$$BEL(y_k) = \alpha\lambda(y_k)\pi(y_k) \quad (4)$$

$$\lambda_Y(z_m) = \sum_y \lambda(y)P(y|z_m) \quad (5)$$

$$\pi_{X_j}(y_k) = \alpha\pi(y_k)\prod_{i\neq j}\lambda_{X_i}(y_k) \quad (6)$$

These equations are specified with respect to a module/region that encodes the random variable Y. Equations 2 to 4 represent how the internal values $\lambda(Y)$, $\pi(Y)$ and $BEL(Y)$ are calculated from incoming messages and locally stored probability tables. Equations 5 and 6 describe how to derive the messages that are to be set as feed forward and feed back outputs of this region. See [13] for more details on Belief Propagation.

REFERENCES

[1] David W. Arathorn. *Map-Seeking Circuits in Visual Cognition: A Computational Mechanism for Biological and Machine Vision*. Stanford Univ Pr, Stanford, CA 94305, Sept 2002.

[2] Suzanna Becker. Implicit learning in 3D object recognition: The importance of temporal context. *Neural Computation*, 11(2):347–374, February 1999.

[3] D. C. Van Essen, C. H. Anderson, and D. J. Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science*, 255(5043):419–423, Jan 24 1992. LR: 20041117; JID: 0404511; RF: 38; ppublish.

[4] Shai Fine, Yoram Singer, and Naftali Tishby. The Hierarchical Hidden Markov Model: Analysis and Applicaitons. *J Opt Soc Am A*, 20(7):1237–1252, 2003.

[5] Peter Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991.

[6] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[7] Jeff Hawkins and Sandra Blakeslee. *On Intelligence*. Times Books, Henry Holt and Company, New York, NY 10011, Sept 2004. In Press.

[8] Aapo Hyvrinen, Jarmo Hurri, and Jaakko Vyrynen. Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *J Opt Soc Am A*, 20(7):1237–1252, 2003.

[9] Tai Sing Lee and David Mumford. Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis*, 20(7):1434–1448, Jul 2003.

[10] S. O. Murray, D. Kersten, B. A. Olshausen, P. Schrater, and D. L. Woods. Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 99(23):15164–15169, Nov 12 2002. LR: 20041117; DEP: 20021104; JID: 7505876; 2002/11/04 [aheadofprint]; ppublish.

[11] T. A. Nazir and J. K. O'Regan. Some results on translation invariance in the human visual system. *Spatial vision*, 5(2):81–100, 1990. LR: 20041117; JID: 8602662; ppublish.

[12] Bruno A. Olshausen, Charles H. Anderson, and David C. Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, 13(11):4700–4719, November 1993.

[13] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman Publishers, San Francisco, California, 1988.

[14] R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, Jan 1999. LR: 20041117; JID: 9809671; CIN: Nat Neurosci. 1999 Jan;2(1):9-10. PMID: 10195172; ppublish.

[15] Rajesh P. N. Rao and Dana H. Ballard. Development of localized oriented receptive fields by learning a translation-invariant code for natural images. *Network: Computation in Neural Systems*, 9(2):219–234, 1998.

[16] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, November 1999.

[17] Simon M. Stringer and Edmund T. Rolls. Invariant object recognition in the visual system with novel views of 3D objects. *Neural Computation*, 14(11):2585–2596, November 2002.

[18] Alex M. Thomson and A. Peter Bannister. Interlaminar connections in the neocortex. *Cerebral cortex (New York, N.Y. : 1991)*, 13(1):5–14, 2003.

[19] Laurenz Wiskott and Terrence J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.