Invariant Pattern Recognition using Bayesian Inference on Hierarchical Sequences

Dileep George
Electrical Engineering, Stanford University
and Redwood Neuroscience Institute
Menlo Park, CA 94025
dil@stanford.edu

Jeff Hawkins
Redwood Neuroscience Institute
Menlo Park, CA 94025
jhawkins@rni.org

Abstract

Real world objects have persistent structure. However, as we move about in the world the spatio-temporal patterns coming from our sensory organs vary continuously. How the brain creates invariant representations from the always-changing input patterns is a major unanswered question. We propose that the neocortex solves the invariance problem by using a hierarchical structure. Each region in the hierarchy learns and recalls sequences of inputs. Temporal sequences at each level of the hierarchy become the spatial inputs to the next higher regions. Thus the entire memory system stores sequences in sequences. The hierarchical model is highly efficient in that object representations at any level in the hierarchy can be shared among multiple higher order objects, therefore, transformations learned for one set of objects will automatically apply to others.

Assuming a hierarchy of sequences, and assuming that each region in the hierarchy behaves equivalently, we derive the optimal Bayes inference rules for any level in the cortical hierarchy and we show how feedfoward and feedback can be understood within this probabilistic framework. We discuss how the hierarchical nested structure of sequences can be learned. We show that static group formation and probability density formation are special cases of remembering sequences. Thus, although normal vision is a temporal process we are able to recognize flashed static images as well. We use the most basic form of one of these special cases to train an object recognition system that exhibits robust invariant recognition.

1 Introduction

Look at any object in front of you. As you move your head, your eyes, or move towards that object while still looking at it, the images that fall on your retina vary significantly from one instant to another. However your percept of the object remains stable despite this variation. This is known as the invariance property. Your cortex does not want your perception of an object to vary with every small eye movement or neck tremor. We consider this invariance property as a technique evolved by the cortex to produce stable percepts of this world. How does the cortex achieve this invariance property?

Think of the different retinal images formed by an object. Although the retinal images are different, the underlying cause of all those images are the same - the object itself. An object is composed of several parts. And those parts are tied to the object in a particular way. When the object moves, it produces a particular motion pattern of the parts. The parts themselves causally influence sub-parts. For example a contour which moves to the left causes a line-segment that is part of it to move in a particular way. A particular sequence of movement of a line segment can be caused by a contour or a corner. A particular sequence of movement of a corner could be due to a table or a chair. The same lower level sequences are reused as part of different high level contexts. Thus the world seems to be naturally organized into a hierarchy of sequences. We believe that the cortex is capturing this causal hierarchical structure of the world using its own hierarchical cortical structure to solve the invariance problem.

Suppose that a region of cortex which can see only a small patch of any image learns all possible ways a line segment can move when it is part of a corner. Now, whenever one of those sequences of movement of that line seqment occurs, the region would be able to say that although the inputs are changing they all belong to the same corner. It seems plausible that by learning the sequences in the context of their causal influences, the invariance problem can be tackled. By doing that in a hierarchy, the same lower level representations can be shared among multiple higher level objects. Therefore, invariances learned for one set of objects will automatically apply to others.

The known anatomy of the visual cortex seems to be conducive to this idea. Visual cortex is organized in a hierarchy and the receptive field size of neurons increase as you go up the hierarchy. Each region in the cortex receives input from below as well as feedback from above. The feedback pathways can provide the contexts of higher level sequences. There are also recurrent connections within a region and between regions via thalamus and such connections could store sequences of different durations.

These ideas are discussed and elaborated in [6] and we consider that as the starting point for this work. The rest of this paper is organized as follows. Section 2 is a mathematical description of how Bayesian inference can be done on hierarchical sequences. In this section we show that the large scale and small scale anatomical structure of the visual cortex is consistent with the idea of Bayesian inference on hierarchical sequences. In section 3, we discuss how such hierarchical structures can be learned. In section 4 we describe an invariant pattern recogniton system that we built based on a subset of principles described in sections 2 and 3. We conclude the paper in section 5 with a discussion on related and future work.

2 Inference and Prediction using Hierarchical Sequences

The goal of this section is to illustrate how Bayesian inference and prediction can occur in a hierarchical sequences framework and how it relates to the known anatomical structure of the visual cortex. We assume that images in this world are generated by a hierarchy of causes. A particular cause at one level influences a sequence of causes to unfold in time at a lower level. For clarity and for notational convenience, we consider a three-level hierarchy. Let the random variables X_i, Y_i and Z_i denote the highest level, intermediate level and lowest level of causes respectively, where i indexes different regions in space active at the same time. We restrict our analysis to cases with only one highest level cause active at any time.

We assume that a particular highest level cause x_k causes a set of sequences $S_{Y_1}^{(k)}$ of Y_1's and $S_{Y_2}^{(k)}$ of Y_2's more likely to simultaneously occur in the child regions Y_1 and Y_2 of X_1. In other words, the higher level cause x_k is identified as the co-occurence of a sequence

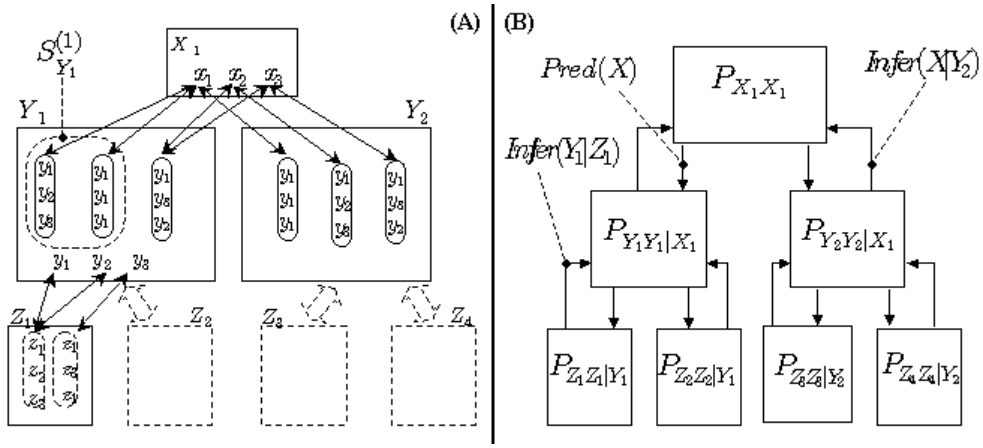

Figure 1: (A) A particular instantiation of hierarchical sequences. The high level cause x_1 of region X causes either sequence $y_1y_2y_3$ or sequence $y_1y_1y_1$ in region Y_1 along with sequence $y_1y_1y_1$ in region Y_2. Elements of these sequences, say for example y_2, act as causes for sequences at lower levels. A particular sequence at any level ($y_1y_3y_2$ in Y_1) can be part of multiple higher level causes (x_2 and x_3 in X_1). (B). Bayesian inference-prediction architecture of the visual cortex based on the derivations in section 2.

in the set $S^{(1)}_{Y_1}$ and a sequence in the set $S^{(1)}_{Y_2}$ in adjacent intermediate level regions. The high level causes vary at a slower rate compared to the lower level causes. For example the higher level cause x_k on an average would stay active for a substantial duration of a sequence in $S^{(k)}_{Y_1}$. In a similar fashion the intermediate level causes Y_i's influence their corresponding lowest level Z variables and vary at a slower rate compared to the Z sequences. A particular instantiation of these ideas is illustrated in figure 1(A).

We assume that the cortical hierarchy matches the causal sequences hierarchy of image generation. This means that there are cortical regions corresponding to the random variables X_i, Y_i and Z_i. For the rest of the discussion we use these labels also to denote their corresponding cortical regions. To simplify the analysis, we assume markovity of sequences at each level. Thus, learning the structure of sequences of region Y_1 would mean learning the probability transition matrix $P_{Y_1Y_1|X_1=x_{1,k}}$ for all k. The highest level propagates itself forward according to $P_{X_1X_1}$. In order to obviate complicated time indexing, we assume that the slower time variation of the high level sequences are captured within their probability transition matrices. Whenever we condition a sequence of causes in a lower level on a particular cause at the higher level, we implicity assume that the higher level cause has not changed for the duration of the lower level sequence.

Lets say that at time t, the region X_1 wants to make a prediction about the probability distribution of $X_1(t+1)$ assuming that it knows $X_1(t), Y_1(t), Y_2(t)$ and $Z_1(t), ..., Z_4(t)$. This can be done as

$$Pred(X_1) = P_{X_1(t+1)|X_1(t),Y_1(t),Y_2(t),Z_1(t),...,Z_4(t)} = P_{X_1(t+1)|X_1(t)} \tag{1}$$

Thus the region X_1 needs only its learned and locally stored matrix $P_{X_1X_1}$ to make predictions. Similarly, region Y_i can make a prediction about the probability distribution of $Y_i(t+1)$ according to

$$\begin{aligned} Pred(Y_i) &= P_{Y_i(t+1)|Y_i(t),X_1(t),Z_1(t),...,Z_4(t)} && (2) \\ &= \sum_j P_{Y_i(t+1)|X_1(t+1)=j,Y_i(t)} P_{X_1(t+1)=j|X_1(t)} && (3) \end{aligned}$$

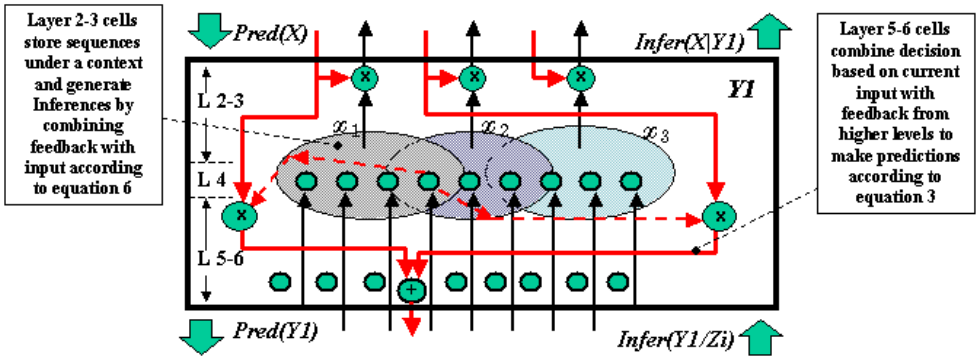

Figure 2: The laminar structure of cortical regions is conducive to the Bayesian inference-prediction architecture based on hierarchical sequences.

Note that the second term on the right hand side of the above equation, $P_{X_1(t+1)=j|X_1(t)}$, is same as the predictive probability distribution calculated by region X_1 in equation 1. Thus for region Y_i to make a prediction about its next state, it has to combine information computed by its parent region with its own locally stored $P_{Y_i(t+1)|X_1(t+1)=j,Y_i(t)}$. Thus the $Pred(X_1)$ information computed by the region X_1 has to be fed down to regions Y_i for those regions to make predictions.

Now lets consider the case when after having observed a sequence of Z_1's and Z_2's, region Y_1 decides to update its estimate of its current state. The optimal estimate of the current state $Y_1(t+1)$ is obtained according to the MAP rule.

$$\hat{Y}_1(t+1) = \arg\max_{Y_1(t+1)} P_{Y_1(t+1)|Y_1(t),Z_1^{t_0:t+1},Z_2^{t_0:t+1},X_1(t)} \quad (4)$$

$$= \arg\max_{Y_1(t+1)} P_{Z_1^{t_0+1:t+1},Z_2^{t_0+1:t+1}|Y_1(t+1),Z_1(t_0),Z_2(t_0)} P_{Y_1(t+1)|Y_1(t),X_1(t)} \quad (5)$$

$$= \arg\max_{Y_1(t+1)} \left[P_{Z_1^{t_0+1:t+1}|Y_1(t+1),Z_1(t_0)} (P_{Y_1(t+1)|Y_1(t),X_1(t)})^{1/2}\right] \times \quad (6)$$

$$\left[P_{Z_2^{t_0+1:t+1}|Y_1(t+1),Z_2(t_0)} (P_{Y_1(t+1)|Y_1(t),X_1(t)})^{1/2}\right] \quad (7)$$

where $Z_i^{t_0:t+1}$ is a sequence of Z_i's extending from time t_0 to time $t+1$. In the above equation, the terms within square brackets can be computed in the regions Z_1 and Z_2 using local information, given that they have the $Pred(Y_1)$ information fed down to them. If these regions send up a set of winning arguments $Y_1(t+1)$ (lets denote it $Infer(Y|Z_i)$), then the region Y_1 can finish the argmax computation exactly and decide the most likely Y_1 state based on that information.

The analysis above shows that the idea of hierarchical sequences and the idea of a hierarchical cortex with feedforward and feedback connection are consistent with each other in a Bayesian inference-prediction framework. The feedforward pathways are used to carry the inferences made based on current as well as locally stored past observations. The feedback pathways carry expectations/predictions to lower levels (figure 1(B)) . Higher level regions have converging inputs from multiple lower level regions. Thus feedback information from a higher level of the cortex can be used as context to interpret/disambiguate an observed pattern at a lower level. Recognition occurs when the highest level converges on a cause by disambiguating several parallel and competing hypotheses at all levels. This derivation also suggests roles for different cells in the known laminar architecture of cortical regions

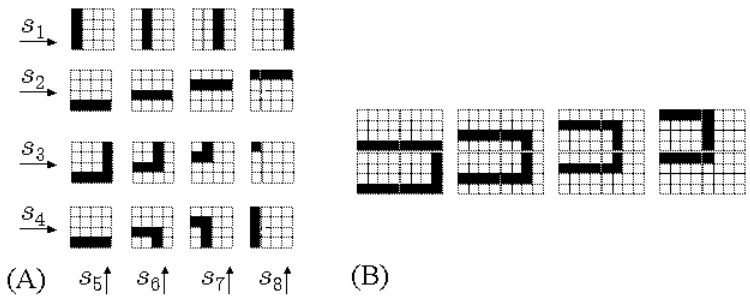

Figure 3: (A). Examples of most likely and most unlikely sequences of length 4 observed in a line-drawing movie by a region of size 4x4. The movie was generated by simulated straight-line motions of images drawn using vertical and horizontal lines. The sequences s_1, s_2, s_3 and s_4 (read left to right) occured much more frequently compared to the sequences s_5, s_6, s_7 and s_8. (read top to bottom). (B)The higher level object/cause shown, moving left and up, is learned by a Y (Level 2) region receiving inputs from 4 Z (Level 1) regions. In this case the object corresponds to simultaneous occurance of s_2, s_4, s_3 and s_2 in the top-left, top-right, bottom-right and bottom-left regions respectively

as shown in figure 2. First order markovity was assumed so that we do not have to carry too many terms in the conditional probability calculations. We believe that similar conclusions as above could be drawn if this assumption is relaxed.

3 Learning Hierarchical Sequences

How can a region of cortex learn sequences within sequences? Consider a region Y_1 receiving the context information of a high level cause $X_1 = x_k$. If this region now learns to associate with this context x_k all sequences of Y_1 that occur at its input while x_k is active, then that region is essentially learning sequences within sequences. After learning, whenever a sequence of Y_1s occur at the input, this region can produce the corresponding X_1 at its output. For example, if the sequences were markov then learning would correspond to learning the matrices $P_{Y_1 Y_1 | X_1 = x_k}$ for every k. In this way, a region of cortex can learn to collapse a sequence at its input to one or more higher level causes based on its learning.

The high level causes themselves have to be learned from the low level inputs. This could be done as follows. Lower level regions learn the most frequent sequences of their inputs. After learning, whenever a part of one of those sequences occur, those regions pass up a pattern corresponding to the sequence. A higher level region with converging inputs from several lower level regions then looks at sequences occuring simultaneously in the low level regions. Patterns of sequences which consistently occur in multiple low level regions become the objects at the next higher level. This process can be repeated between levels of the hierarchy to obtain causes at the highest level. For example if region Y_1 observes that the sequence $s^j_{Z_1}$ of region Z_1 and the sequence $s^k_{Z_2}$ of region Z_2 occur at the same time very often, then their combination becomes an object or cause at region Y_1. Examples of learned sequences and higher level causes for line drawing movies are shown in figure 3.

Note that if under the context $X_1 = x_k$, a Y region stored only the frequency of occurences of its inputs Y_1, then this corresponds to learning the conditional probability distribution $P_{Y_1 | X_1 = x_k}$. This is a special case of sequence learning where sequences are of length 1. In the markov case this would correspond to learning the steady state distribution of the markov chain $P_{Y_1 Y_1 | X_1 = x_k}$. Now, if under the context $X_1 = x_k$ we just group all the

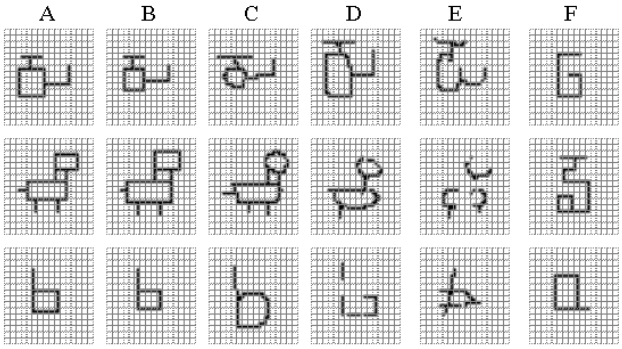

Figure 4: Examples of images used to train and test the invariant pattern recognition system. The rows correspond to 3 out of the 91 categories for which the system was trained. Columns A and B correspond to the two different images of a category that were used for training. Columns C and D are examples of test patterns for which the system recognized the image category correctly. Note that the system shows a high degree of shift, scale and distortion invariance. Some of these test images were drawn by our lab mates using a mouse. Column E gives examples of patterns for which the system made an error in recogniton and column F gives the category the column E pattern was incorrectly classified as. The complete set of training and test patterns and the MATLAB code for the system can be downloaded from http://www.rni.org/nips2004/

inputs Y_1 then it becomes a special case of learning the probability distribution. In this case the probability distribution is uniform over all Y_1s having non-zero probability under the contex x_k.

4 Simulation of a Line Drawing Recognition System

Using a subset of the principles outlined above, we simulated a hierarchical system for line drawing recognition and measured various aspects of its performance. Instead of storing sequence information, we considered a sequences as groups (sets), thus dropping timing information within a sequence. As noted in the previous section, this is can be considered as a special case of learning sequences. We do not make use of feedback in this implementation. This can also be considered as a special case where all feedback probability densities are uniform. Using the full set of principles outlined in sections 2 and 3 can only improve the performance of the system.

The system consisted of 3 levels - L1, L2 and L3. The lowest level, L1, consisted of regions receiving inputs from a 4x4 patch of images which were of size 32 pixels by 32 pixels. These 4x4 regions regions tiled an input image with 2 pixels overlap between adjacent regions. This overlap between regions ensured that spatial continuity constraints are maintained. Learning started at L1 and proceeded to the higher levels. The L1 regions learned by obtaining the most likely sequences caused by simulated motion of of black and white straight-line drawings (figures 3 and 4). For example, vertical lines and all shifts of vertical lines within an L1 region became the *vertical line group* and left-bottom corners and all shifts of them formed the *left-bottom corner group* . With this, an L1 region presented with a vertical line at its input would produce the output *vertical line group* irrespective of the position of the vertical line within that region. (In our implementation, with 13 groups in L1, it set one of out 13 bits to 1). For novel patterns appearing at the input, the output was set as the group of the closest (euclidean distance) familiar pattern.

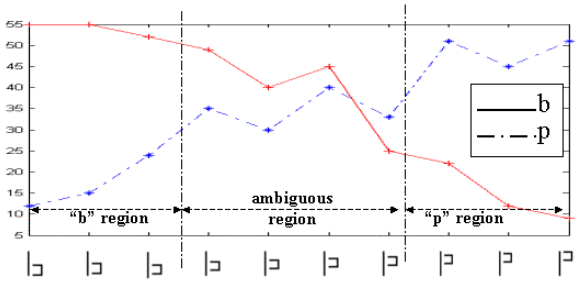

Figure 5: Variation of perceptual strength with pattern distortion : For this test we gradually distorted a test pattern of one category (category *b*) into a pattern of another category (category *p*). None of the test patterns were identical to the ones used for training the *b* and *p* categories. Plotted along the Y axis is the score obtained at L3 for categories *b* and *p* when patterns along the X axis were shown to the system. In the region marked as *b region* the pattern was identified as belonging to category *b* and similarly for the *p region*. In the ambiguous region the system classified the pattern as neither *b* nor *p* but (erroneously) idenitied it as various other categories.

An L2 region received its inputs from 16 L1 regions. In our implementation this pattern is of length 208. The groups at L2 were formed in a semi-supervised manner, with learning context coming from L3. We showed the network moving images of objects of a particular category all the while setting a constant context from L3 to all L2 regions. Thus the L2 regions learned to associate all the inputs from L1 region that occured under a particular category context with that category. During the recognition phase, an L2 region set at its output the category memberships of the pattern it received at its input. If the membership was null, it output an all zero pattern. Thus, during the recognition phase, each L2 region sent up its multiple hypotheses regarding the possible L3 causes. A single L3 region pooled all such hypotheses from 16 L2 regions below it. The L3 region would make a decision regarding the categry of object by counting the votes from all L2 regions.

We observed in our introdcution that the perception of an object should remain stable despite eye movements as long as the object remains within the field of view and is attended to. If the input is ambiguous, the brain can gather further information from the input by making small eye movements. Between these eye movements, the correct hypothesis would remain stable while the competing incorrect hypotheses will vary in a random manner. We made use of this idea to improve the signal to noise ratio for detecting novel patterns.

The system was trained on simulated motions of 91 objects. Two examples of every object were shown to the system during training (figure 4 A, B). Accuracy of detection on the training set was 100% without any eye movement. For test cases, we limited the maximum number of eye movements to 12. The system showed a high degree of invariance to position, scale and distortion on novel patterns as displayed in figures 4 and 5.

5 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[5, 12, 11]. 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 [4, 13]. Some recent models have used temporal slowness as a criterion to learn representations [7, 14, 2]. However those systems lacked a Bayesian inference-prediction framework [8] 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)[3] to capture structure at multiple scales either in space or in time. Moreover, algorithm stuctures like HHMMs and Markov Random Fields [9] have remained as abstract computer vision models because they haven't made any connections with known cortical structure. Several other models [1, 10] 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.

We demonstrated invariant pattern recognition using only a subset of the principles outlined in sections 2 and 3. We believe that, using the full strength of the outlined theory, we will be able to demonstrate other well known cortical phenomena [8]. Although we used supervised learning in our simulation of the pattern recognition system, this is not a necessary component of the theory. We believe that it is possible to learn high level causes in an unsupervised fashion by learning sequences of sequences as demonstrated in figure 3. Future work will include application of these ideas to natural videos.

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] 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.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] Kevin Murhpy, Antonio Torralba, and William T. Freeman. Using the Forest to See the Trees: A Graphical Model Relating Features Objects and Scenes. *Advances in Neural Information Processing Systems 16, Vancouver, BC*, 16, 2004.

[10] 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.

[11] 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.

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

[13] 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.

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