Submitted to NIPS96, Section: **Cognitive Science**. Preference: Oral presentation.

Neural Models for Part-Whole Hierarchies

Maximilian Riesenhuber **Peter Dayan**
Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139
{max,dayan}@ai.mit.edu

Abstract

We present a connectionist method for representing images that explicitly addresses their hierarchical nature. It blends data from neuroscience about whole-object viewpoint sensitive cells in inferotemporal cortex[7] and attentional basis-field modulation in V4[3] with ideas about hierarchical descriptions based on microfeatures.[4,10] The resulting model makes critical use of pathways for both analysis and synthesis.[5] We illustrate the model with a simple example of representing information about faces.

1 Hierarchical Models

Substantial recent effort has been devoted to analysis-by-synthesis models for visual processing.[15,5] The synthetic or *generative* models form the map:

$$\text{`object'} \rightarrow \text{`image'} \tag{1}$$

where 'object' implies the identity of the object (such as a face) and its relevant instantiation parameters (such as whether it is smiling or frowning) and 'image' stands in for the pattern of activity over the input units that is caused by observing the object. The Helmholtz machine[5] (HM) employs unsupervised learning to construct an explicit generative model of this form, using a top-down layered processing structure that is a form of factorial belief network.[9] The activities of the units in the top layer, or possibly all hidden layers, represent the object and the choices of its instantiation parameters.

The task for analysis or *recognition* is to take an 'image' and produce (a probability distribution over) the object that might have generated it, and also the instantiation parameters of that object. The HM uses a bottom-up belief net over the same units to represent the recognition model, and trains top-down and bottom-up models to be mutual inverses.

We thank Larry Abbott, Geoff Hinton, Bruno Olshausen, Tomaso Poggio, Alex Pouget, Emilio Salinas and Pawan Sinha for discussions and comments.

Images of objects constitute an important paradigm case of a representational hierarchy. This is not the sort of *processing* hierarchy seen in multi-layer structures such as the HM, but rather a 'part-whole' hierarchy (*eg* a mouth is part of a face). The representation and manipulation of part-whole hierarchical information in fixed hardware is a heavy millstone around connectionist necks. For example, anterior inferotemporal cortex (IT) appears to construct representations of visually presented objects. Both mouths and faces are objects themselves, and so require fully elaborated representations, presumably at the level of anterior IT, using different (or possibly partially overlapping) sets of cells. If the cells representing mouths are at the same level of the processing system as those representing faces, how can information naturally available at the level of the description of the mouths influence the recognition of whole faces; conversely, how can knowledge about a face generate a pattern of activity over the mouth units representing the mouth that it contains? Having the mouth units just feed into the face units is not enough, since this places implausible restrictions on the number of levels in the hierarchy, and is in any case incapable of representing true recursive structure (consider representing in a similar manner a room which contains a painting which itself depicts another room). Models such as the HM ignore this issue by ignoring the relationship between parts and wholes.

We develop a scheme for representing and reasoning about hierarchical image knowledge which is based on a different synthetic model:

$$\text{`object'} \otimes \text{`attentional eye-position'} \rightarrow \text{`image'} \qquad (2)$$

(shown in cartoon form in Fig. 1) in which generative knowledge about an object is really (probabilistic) knowledge about what it looks like on the retina (the 'image') under various transformations (the 'attentional eye-position'). In the simplest form of the model, in which we ignore its proper statistical basis, the image is exactly the activity over the input units that would be observed were the object to be placed in a position on the retina corresponding to the attentional eye-position. In the future, we plan to generalize this to include the parameters of arbitrary transformations, *eg* rotation. Just as for the non-hierarchical case, the recognition counterpart to the generative map in mapping (2) has to take an image and return (probability distributions over) the object and its instantiation parameters (and also the transformation parameters corresponding to the attentional eye-position).

If a system could acquire the map in (2) and its recognition inverse, it would have solved the part-whole hierarchical problem – *eg* it would be able to work out what the image of the lower middle part of a face is, and recognise it as a mouth. There is a natural formulation of this pair of recognition and generative models in terms of areas of the cortex responsible for processing visual input:

a) Images consist of patterns of activity over the lateral geniculate nucleus.

b) The recognition model is bottom-up, leading to distributed patterns of activity of units in anterior IT which directly represent the object. In general these activities would also represent some of the instantiation parameters of the object.

c) The viewpoint dependent 2d image sensitive cells studied by Logothetis *et al*[7] are crucial to the scheme. To the data that there are cells that fire reliably when a monkey is presented with a particular 2d image of a face, we add the capacity to regenerate the activity caused by the appearance of that face, based on the activity of these cells, and also the capacity to generate the activity caused by the

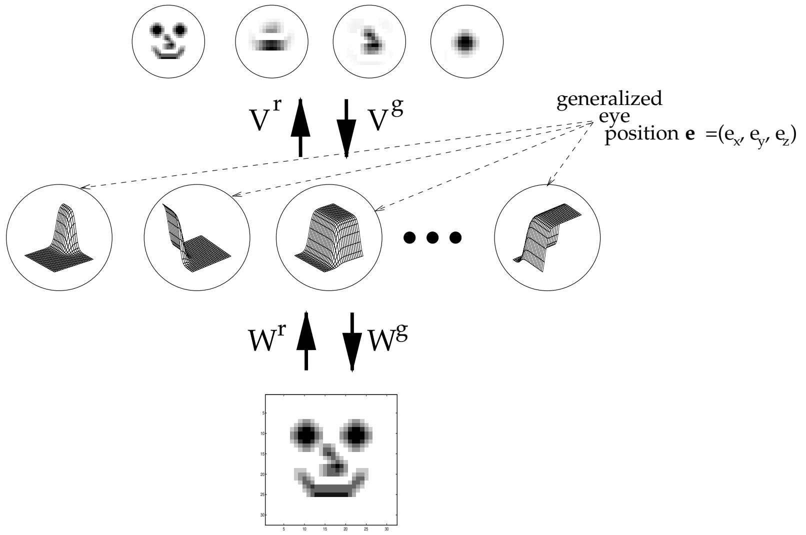

Figure 1: Cartoon of the model. In the top-down, generative, direction, the model generates images of faces, eyes, mouths or noses based on an attentional eye position and a selection of a single top-layer unit; the bottom-up, recognition, direction is the inverse of this map. The response of the neurons in the middle layer is modulated sigmoidally (as illustrated by the graphs shown inside the circles representing the neurons in the middle layer) by the attentional eye position. See section 2 for more details.

appearance of sub-parts of the face.

d) The generative model is top-down, leading, in principle, from these same units all the way to V1. It may not be necessary to generate an image all the way on V1 to know that faces contain mouths – it may require creating activity only on immediately preceding layers.

e) Both the recognition and the generative pathways require particular modulation for the effective eye position. The requirement for this is clear in the generative model – one can ask what a face at some scale would look like if it were evident only from the left corner of one's eye. This leads to a top-down computational task very like the bottom-up one addressed using a multiplicatively controlled synaptic matrix in the shifter model of Olshausen *et al.*[8] Two groups, Pouget & Sejnowski[12] and Salinas & Abbott[14] interpreted in terms of basis fields the eye-position modulated cells in parietal cortex that Andersen *et al*[1] find, and showed that these basis fields can be used to achieve the same task as the shifter model but with neuronal rather than synaptic multiplicative modulation. Connor *et al*[3] found such neuronal modulation in many cells in area V4 in the object processing ventral stream, only now for attentionally defined eye-position rather than the actual eye-position. Eye-position modulation almost certainly occurs at many levels in the system, possibly including V1.[11]

In the model, analysis progresses in a similar way to the HM, except that the cells are modulated by an attentional eye-position just as during synthesis. This modulating attentional eye-position must be able to become detached from the spatial eye-position – Connor *et al* collected evidence for part of this hypothesis; although the coordinate system(s) of the modulation is not entirely clear from their data. Under this view, recognition requires finding the correct attentional eye-position – this itself can either be learned directly or controlled in a top-down manner.[8] Note

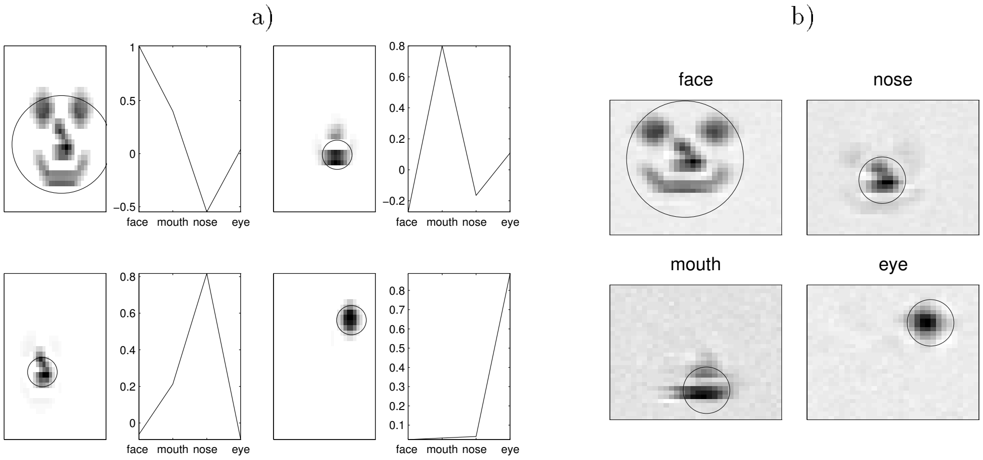

Figure 2: **a)** Recognition: the left column of each pair shows the stimuli; the right shows the resulting activations in the top layer (ordered as face, mouth, nose and eye). The stimuli are faces at random positions in the retina. Recognition is performed by setting the attentional eye position in the image and setting the attentional scale, which creates a window of attention around the attended to position, shown by a circle of corresponding size and position. **b)** Generation: each panel shows the output of the generative pathway for a randomly chosen attentional eye position on activating each of the top layer units in turn. The focus of attention is marked by a circle whose size reflects the attentional scale. The name of the object whose neuronal representation in the top layer was activated is shown above each panel.

that this is an instance of exactly the same translation invariance problem that Olshausen *et al*[8] addressed with their shifter mechanism, except that receptive field shifting is again replaced by receptive field modulation.

f) The bottom-up and top-down mappings are learned taking the eye-position modulation into account. In the experiments below, we used a version of the wake-sleep algorithm[5] for its conceptual and computational simplicity. This requires learning the bottom-up model from generated imagery (during sleep) and learning the top-down model from assigned explanations (during observation of real input during wake). In the current version, for simplicity, the eye position is set correctly during recognition.

2 Results

We have developed a simple model that illustrates the feasibility of the scheme presented above in the context of recognizing and generating cartoon drawings of a face and its parts. Recognition involves taking an image of a face or a part thereof (the mouth, nose or one of the eyes) at an arbitrary position on the retina, and setting the appropriate top level unit to 1 (and the remaining units to zero). Generation involves imaging either a whole face or of one of its parts (selected by the active unit in the top layer) at an arbitrary position on the retina. In general, the top level would implement a distributed code for the identity and instantiation parameters of the objects.

The model (Fig. 1) consists of three layers. The lowest layer is a 32×32 'retina'. In the recognition direction, the retina feeds into a layer of 500 hidden units. These project to the top layer, which has four neurons. In the generative direction, the connectivity is reversed. The network is fully connected in both directions. The activity of each neuron based on input from the preceding (for recognition) or the following layer (for generation) is a linear function (weight matrices W^r, V^r in the recognition and V^g, W^g in the generative direction). The attentional eye position influences activity through multiplicative modulation of the neuronal responses in the hidden layer. The linear response $r_i = (W^r \mathbf{p})_i$ or $r_i = (V^g \mathbf{o})_i$ of each neuron i in the middle layer based on the bottom-up or top-down connections is multiplied by $\xi_i = \phi_i^x(e_x)\phi_i^y(e_y)\phi_i^s(e_s)$, where $\phi_i^{\{x,y,s\}}$ are the tuning curves in each dimension of the attentional eye position $\mathbf{e} = (e^x, e^y, e^s)$, coding the x- and y- coordinates and the scale of the focus of attention, respectively. Thus, for the activity m_i of hidden neuron i we have $m_i = (W^r \mathbf{p})_i \cdot \xi_i$ in the recognition pathway and $m_i = (V^g \mathbf{o})_i \cdot \xi_i$ in the generative pathway. The tuning curves of the ξ_i are chosen to be sigmoid with random centers c_i and random directions $d_i \in \{-1, 1\}$, *eg* $\phi_i^s = \sigma(4 * d_i^s * (e^s - c_i^s))$. In the recognition direction, the attentional eye position also has an influence on the activity in the input layer by defining a 'window of attention',[6] which we implemented using a Gaussian window centered at the attentional eye position with its size given by the attentional scale. This is to allow the system to learn models of parts based on experience with images of whole faces.

To train the model, we employ a variant of the unsupervised wake-sleep algorithm.[5] In this algorithm, the generative pathway is trained during a wake-phase, in which stimuli in the input layer (the retina, in our case) cause activation of the neurons in the network through the recognition pathway, providing an error signal to train the generative pathway using the delta rule. Conversely, in the sleep-phase, random activation of a top layer unit (in conjunction with a randomly chosen attentional eye-position) leads, via the generative connections, to the generation of activation in the middle layer and consequently an image in the input layer that is then used to adapt the recognition weights, again using the delta rule. Although the delta rule in wake-sleep is fine for the recognition direction, it leads to a poor generative model – in our simple case, generation is much more difficult than recognition. As an interim solution, we therefore train the generative weights using back-propagation, which uses the activity in the top layer created by the recognition pathway as the input and the retinal activation pattern as the target signal. Hence, learning is still unsupervised (except that appropriate attentional eye-positions are always set during recognition).

Figure 2a shows several examples of the performance of the recognition pathway for the different stimuli after 300,000 iterations. The network is able to recognize the stimuli accurately at different positions in the visual field. Figure 2b shows several examples of the output of the generative model, illustrating its capacity to produce images of faces or their parts at arbitrary locations. By imaging a whole face and then focussing the attention on *eg* an area around its center, which activates the 'nose' unit through the recognition pathway, the relationship that, *eg*, a nose is part of a face can be established in a straightforward way.

3 Discussion

Representing hierarchical structure is a key problem for connectionism. Visual images offer a canonical example for which it seems possible to elucidate some of the underlying neural mechanisms. The theory is based on 2d view object selective cells in anterior IT, and attentional eye-position modulation of the firing cells in V4. These work in the context of analysis by synthesis or recognition and generative models such that the part-whole hierarchy of an object such as a face (which contains eyes, which contain pupils, *etc*) can be traversed in the generative direction by choosing to view the object through a different effective eye-position, and in the recognition direction by allowing the real and the attentional eye-positions to be decoupled to activate the requisite 2d view selective cells.

The scheme is related to Pollack's Recursive Auto-Associative Memory (RAAM) system.[10] RAAM provides a way of representing tree-structured information – for instance to learn an object whose structure is $\{\{A, B\}, \{C, D\}\}$, a standard three-layer auto-associative net would be taught AB, leading to a pattern of hidden unit activations α; then it would learn CD leading to β; and finally $\alpha\beta$ leading to γ, which would itself be the representation of the whole object. The compression operation ($AB \rightarrow \alpha$) and its expansion inverse are required as explicit methods for manipulating tree structure. It is not clear how recognition might naturally work for Pollack's system – for instance one would seem to have to RAAMify the various parts (eyes, nose, mouth, *etc*) of an observed image in order to compare the net representation with that for a stored face.

We are adopting a similar strategy to RAAM for representing hierarchical information, except that a) RAAM normally constructs its own codes for intermediate levels of the trees that it is fed, whereas here, images of faces are as real and as available as those of their associated mouths; b) by contrast with RAAM's discrete symbols at the leaves and discrete levels in the hierachy, the images and the levels are continuous; and c) the notion of 'eye-position' stands in for the compression and expansion operations that are central to RAAM. The computational methods are also different.

Various aspects of our scheme require comment: the relationship between eye-position and recognition and generative imagery, variants of the scheme for objects that are too big or too geometrically challenging to 'fit' in one go into a single image (*eg* the room example above), and to hierarchical objects other than images. We are also working on a more probabilistically correct version, taking advantage of the statistical soundness of the HM.

Eye position information is ubiquitious in visual processing areas,[11] including the LGN and V1,[16] as well as the parietal cortex[1] and V4.[3] Further, it can be revealed as having a dramatic effect on perception, as in Ramachandran *et al*'s[13] study on intermittent exotropes. This is a form of squint in which the two eyes are normally aligned, but in which the exotropic eye can deviate (voluntarily or involuntarily) by as much as 60°. The study showed that even if an image is 'burnt' on the retina in this eye as an afterimage, and so is fixed in retinal coordinates, at least one component of the percept *moves* as the eye moves. This argues that information about eye position dramatically effects visual processing in a manner that is consistent with the model presented here of shifts based on modulation. This is also required by Bridgeman *et al*'s[2] theory of perceptual stability across fixations, that essentially builds up an impression of a scene in exactly the form of mapping (2).

Knowledge of images that are too large to fit naturally in a single view at a canonical location and scale, or that theoretically cannot fit in a view (like 360° information about a room) can be handled in a straightforward extension of the scheme. All this requires is generalizing further the notion of eye-position. One can explore one's generative model of a room in the same way that one can explore one's generative model of a face.

We have described our scheme from the perspective of images. This is convenient because of the substantial information available about visual processing. However, images are not the only examples of hierarchical structure – this is also very relevant to words, music and also inferential mechanisms. We believe that our mechanisms are also more general – proving this will require the equivalent of the attentional eye-position that lies at the heart of the method.

References

[1] Andersen, R, Essick, GK & Siegel, RM (1985). Encoding of spatial location by posterior parietal neurons. *Science,* **230**, 456-458.

[2] Bridgeman, B, van der Hejiden, AHC & Velichkovsky, BM (1994). A theory of visual stability across saccadic eye movements. *Behavioral and Brain Sciences,* **17**, 247-292.

[3] Connor, CE, Gallant, JL, Preddie, DC & Van Essen, DC (1996). Responses in area V4 depend on the spatial relationship between stimulus and attention. *Journal of Neurophysiology,* **75**, 1306-1308.

[4] Hinton, GE (1981). Implementing semantic networks in parallel hardware. In GE Hinton & JA Anderson, editors, *Parallel Models of Associative Memory.* Hillsdale, NJ: Erlbaum, 161-188.

[5] Hinton, GE, Dayan, P, Frey, BJ & Neal, RM (1995). The wake-sleep algorithm for unsupervised neural networks. *Science,* **268**, 1158-1160.

[6] Koch, C & Ullmann, S (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology,* **4**, 219-227.

[7] Logothetis, NK, Pauls, J, & Poggio, T (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology,* **5**, 552-563.

[8] Olshausen, BA, Anderson, CH & Van Essen, DC (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience,* **13**, 4700-4719.

[9] Pearl, J (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann.

[10] Pollack, JB (1990). Recursive distributed representations. *Artificial Intelligence,* **46**, 77-105.

[11] Pouget, A, Fisher, SA & Sejnowski, TJ (1993). Egocentric spatial representation in early vision. *Journal of Cognitive Neuroscience,* **5**, 150-161.

[12] Pouget, A & Sejnowski, TJ (1995). Spatial representations in the parietal cortex may use basis functions. In G Tesauro, DS Touretzky & TK Leen, editors, *Advances in Neural Information Processing Systems 7*, 157-164.

[13] Ramachandran, VS, Cobb, S & Levi, L (1994). The neural locus of binocular rivalry and monocular diplopia in intermittent exotropes. *Neuroreport,* **5**, 1141-1144.

[14] Salinas, E & Abbott LF (1996). Transfer of coded information from sensory to motor networks. *Journal of Neuroscience,* **15**, 6461-6474.

[15] Sung, K & Poggio, T (1995). *Example based learning for view-based human face detection.* AI Memo 1521, CBCL paper 112, Cambridge, MA: MIT.

[16] Trotter, Y, Celebrini, S, Stricanne, B, Thorpe, S & Imbert, M (1992). Modulation of neural stereoscopic processing in primate area V1 by the viewing distance. *Science,* **257**, 1279-1281.