Max-Planck-Institut
für biologische Kybernetik
Arbeitsgruppe Bülthoff

Spemannstraße 38 • 72076 Tübingen • Germany

Technical Report No. 36 August 1996

Presentation order affects human object recognition learning

Guy Wallis

Abstract

The view based approach to object recognition relies upon the co-activation of 2-D pictorial elements or features. This approach is limited to generalising recognition across transformations of objects in which considerable physical similarity is present in the stored 2-D images to which the object is being compared. It is, therefore, unclear how completely novel views of objects might correctly be assigned to known views of an object so as to allow correct recognition from any viewpoint.
The answer to this problem may lie in the fact that in the real world we are presented with a further cue as to how we should associate these images, namely that we tend to view objects over extended periods of time. In this paper, neural network and human psychophysics data on face recognition are presented which support the notion that recognition learning can be affected by the order in which images appear, as well as their spatial similarity.

Guy Wallis was supported with a research fellowship from the Max-Planck Gesellschaft

Introduction

To successfully interact with the everyday objects that surround us we must be able to recognise these objects under widely differing conditions, such as novel viewpoints or changes in retinal size and location. Only if we can do this correctly can we determine the behavioural significance of these objects and decide whether the sphere in front of us should, for example, be kicked or eaten. Similar, although often finer, discriminations are required in face recognition. One might be presented with the task of deciding which side of the aisle is reserved for the groom's family at your cousin's wedding - a decision of familiar verses unfamiliar categorisation. On the other hand, the faces may be familiar and the task becomes one of distinguishing family members, such as your aunt and your sister. Such decisions have clear social significance and are crucial in deciding how to interact with other people. The question of how we manage this remains open.

Theories for how we represent objects and ultimately solve object recognition abound. Examples include the building and matching of mental 3D models (Marr, 1982; Marr & Nishihara, 1978), decomposition into locally interrelated basic geometric primitives (Biederman, 1987; Biederman & Gerhardstein, 1993), and template matching algorithms (Ullman, 1989; Yuille, 1991). However, the proposal that I shall be considering here suggests that recognition is supported by a series of picture elements or features associated into 2D views of an object. Support for this theory comes both from psychophysical (Tarr & Bülthoff, 1995; Tarr & Pinker, 1989; Bülthoff & Edelman, 1992) and neurophysiological (Logothetis & Pauls, 1994) sources.

These models propose that generalisation of recognition to novel views could be achieved by a large number of broadly tuned feature sensitive units each tolerant to small deformations of their preferred features. This would then be sufficient to perform recognition over small transformations (Poggio & Edelman, 1990) - at least given some form of supervised training regime. Ultimately, however, there has to be a limit to the amount of generalisation one can afford to make from a set of feature sensitive cells before they lose their power to discriminate. Even assuming some scheme for the pre-normalisation of object size and translation (a big assumption!), one would still require separate feature detectors for large view changes. In the absence of a supervised training signal, it is not clear how a series of different views of an object - which may share very few, if any, of the features supporting the recognition - might be associated together.

To describe a potential solution to this problem it is worth reflecting on what clues our environment gives us about how to associate the stream of images that we normally perceive. The fact that our natural environment is full of higher-order spatial correlations has received considerable attention (Field, 1987), whereas the existence of statistical regularity in the temporal domain (Dong & Atick, 1995), has not. Temporal regularity emerges from the simple fact that we often study objects for extended periods, resulting in correlations in the appearance of the retinal image from one moment to the next. This regularity may provide us with a simple heuristic for deciding how to associate novel images of objects with stored object representations. Since objects are often seen over extended periods, any unrecognised view coming straight after a recognised one is most probably of the same object. This heuristic will work as long as accidental associations from one object to another are random and associations from one view of an object to another are experienced regularly. There is every reason to suppose that this is actually what will happen under normal viewing conditions, and that by approaching an object, watching it move, or rotating it in our hand, for example, we will receive a consistent associative signal capable of bringing all of the views of the object together.

In the last few years both neurophysiologists (Perrett & Oram, 1993; Rolls, 1992) and neural network theorists (Edelman & Weinshall, 1991; Földiák, 1991; Wallis & Rolls, 1996) have been exploring this theme on the basis of neurophysiological recordings in the infero-temporal lobe (IT) of primates. Recording in this area, Miyashita and his colleagues (Miyashita, 1988, Miyashita & Chang, 1988) studied the effect of repeatedly showing a sequence of randomly selected fractal images. They discovered that cells in IT would learn to respond to one stimulus in the series very strongly, but also to images appearing in close succession, purely as a function of temporal and not spatial disparity between stimuli.

There is good evidence from single cell recordings (Rolls, 1992; Desimone, 1991; Tanaka et al., 1991) and anatomical studies (Ungerleider & Mishkin, 1982; Plaut & Farah, 1990) that the neurons in IT play an important role in object recognition. It thus seems plausible that what Miyashita *et. al.* have observed is the functioning of a system for associating views of objects simply on the basis of their appearance in time. It is this hypothesis that serves to motivate the work described in this paper. Evidence to support the hypothesis is presented as a significant effect of temporal order in establishing the perceived similarity and identity of the views of faces.

Psychophysical Experiments

Introduction

If object recognition is affected by the temporal order in which images of objects appear, it seems reasonable to test for an effect psychophysically. This first section sets out to describe just such an effect in face recognition.

The field of face recognition has been very extensively studied and several researchers have reported that we make recognition errors when the viewing position is changed from view to test - especially if the faces are unfamiliar (Krouse, 1981; Logie et al., 1987; Patterson & Baddeley, 1977; Wogalter & Laughery, 1987; Troje & Bülthoff, 1996). The experiment described in this section exploits this fact by testing recognition performance on a set of interleaved faces presented in smooth sequences. These sequences consisted of five views of a face, presented in even steps from left profile to right. The time-based association hypothesis described above, predicts that the visual system would associate these images together as being views of the same face.

If, however, the subject's task is to identify individuals, any associations made across different faces would be erroneous. This should then become apparent by the increased number of discrimination errors for these faces in comparison with faces not seen in sequences. Figure 1 puts this hypothesis in a more graphical light by displaying three possible sequences each containing five different faces seen from five evenly spaced viewing angles. The temporal association hypothesis would be supported if confusion rates for faces within sequences S1, S2 and S3 are higher than for faces selected between the sequences.

Methods

After a brief familiarisation phase using a separate set of faces, subjects viewed three sequences of faces each containing five poses of five different faces - see figure 1 - with the distribution of faces varying randomly between subjects. The faces were displayed on a black background, on an SGI Indigo workstation. Each image subtended approximately $10° \times 6°$ at a viewing distance of 50cm. In any one sequence the pose of the face was altered smoothly from left profile to right profile and back, with each of the five sequence member faces appearing in one of the five poses. Each sequence was seen five times such that each face was shown from each one of the five viewpoints. Each view within a sequence was seen for 300ms with no delay between images. The delay between sequences was set at 2500ms. Before viewing the sequences, subjects were instructed to 'Attend closely to the faces as they turn'.

After viewing the five permutations of each sequence, subjects were tested in a standard same/different 2AFC paradigm. One face was presented for 150ms, then a colour mask was presented for 150ms, and then finally a second face was presented for 150ms. The subjects' task was to respond by pressing a key to indicate 'same' or 'different' to each of the pairs shown. Each trial fell under one of three possible conditions.

- A The same face was shown from different viewpoints.
- B Two different faces from the same sequence were shown.
- C Two different faces from different sequences were shown.

To balance the number of 'same' and 'different' trials, condition A contained 30 trials whilst the B and C conditions contained 15 trials each. Trials from the three conditions were interleaved and repeated three times, making a total of 180 trials per trial block. The entire block - including both the training and testing phases - was then repeated twice more, yielding a total of 540 trials per subject.

Results

Twelve naive subjects participated in the experiments. The data of 2 subjects were excluded from the analysis because their recognition rates did not exceed chance. The overall performance is shown averaged over all three blocks in figure 2 and broken down into individual blocks in figure 3.

A 2-way ANOVA was used to analyse percent correct with test condition and trial block number as independent variables. There was a significant effect of test condition ($F(2,18) = 14.978, p < 0.01$). Tukey's Honestly Significant Difference Test indicated a significant difference between all three condition means with condition A significantly greater than condition C which was significantly greater than condition B ($p < 0.05$). The fact that performance on same trials (condition A) was better than for different trials (conditions B and C) has been described in the face recognition literature before (Patterson & Baddeley, 1977).

Of particular interest here, however, was the significant effect of sequence on the different trials. Subjects confused different faces from the same sequence (condition B) more often than they confused different faces from different sequences (condition C) - see figure 2. The results also appear to show that the effect increases across trial blocks, because the performance in condition B decreases over the three blocks - see figure 3. This effect was not, however, significant at the 5% level.

Sensitivity and response bias was also computed over all three trial blocks and all ten subjects used

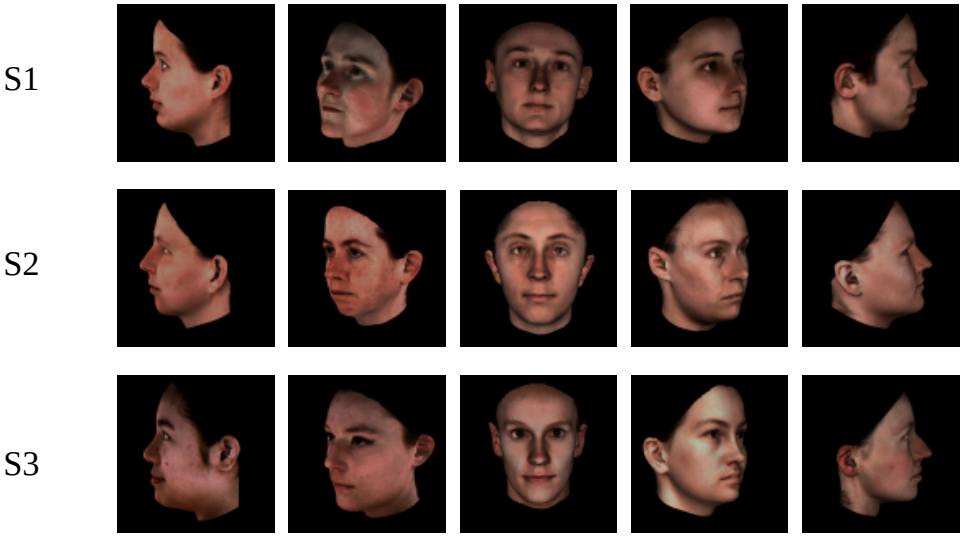

Figure 1: Example of the faces used and the sequences (S1,S2,S3) presented.

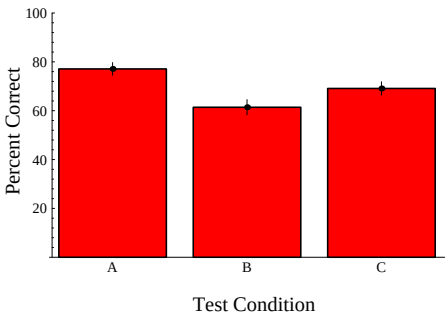

Figure 2: Average recognition performance under the three test conditions.

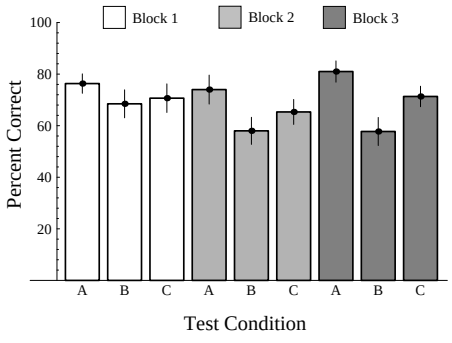

Figure 3: Subject performance broken down into consecutive trial blocks.

in the analysis. Hit rates were established from condition A and correct rejection rates from the average of conditions B and C. Sensitivity was fairly high ($d' = 1.424$) and no strong biasing effects were measured ($c = -0.027, \beta = 0.96$).

Neural Network Simulations

Introduction

As described above, the psychophysical experiments conducted here were inspired by both neurophysiological findings and theoretical results using neural network models. In this section a short experiment is presented in which the same faces used in the previous experiment are presented to a simple neural network. This network utilises a learning rule similar to that proposed by Hebb (1949), but which is designed to establish neurons selective for images appearing in sequences as well as simply on the basis of physical appearance.

The learning rule in question was first used in the context of invariant object recognition by Földiák (1991), who demonstrated its use for associating sequences of parallel lines[1]. More recently, I have extended these ideas by tying them more closely to neurophysiological data (Wallis & Rolls, 1996), and exploring the theoretical basis of how the learning rule works (Wallis, 1996a,b).

By presenting the same faces used in the previous experiment to a simple network utilising this learning rule, the hope is to confirm that the type of effects seen here in the human data can indeed be replicated by virtue of this unsupervised learning rule.

The version of the learning rule used in this work is equivalent to Földiák's (1991) and can be summarized as follows:

$$\Delta w_{ij}{}^{(t)} = \alpha \overline{y}_i{}^{(t)} x_j \quad (1)$$

$$\sum_j w_{ij}{}^2 = 1 \; for \; each \; i^{th} \; neuron \quad (2)$$

$$\overline{y}_i{}^{(t)} = (1-\eta) y_i{}^{(t)} + \eta \overline{y}_i{}^{(t-1)} \quad (3)$$

$$y_i = \Phi \left[\sum_j x_j w_{ij} \right] \quad (4)$$

where x_j is the j^{th} input to the neuron, y_i is the output of the i^{th} neuron, w_{ij} is the j^{th} weight on the i^{th} neuron, η governs the relative influence of the trace and the new input, and $\overline{y}_i{}^{(t)}$ represents the value of the i^{th} cell's recent activity at time t. The function Φ implements lateral inhibition within a local region of neurons and transforms input activation into a firing rate by passing it through a sigmoidal activation function.

Equation 1 has the familiar form of Hebb learning except that the standard instantaneous neural activity term y_i has been replaced with the term $\overline{y_i}^{(t)}$. The value is related to y_i but is now time dependent indicated by the (t) superscript and is also an average, indicated by the line above the y. What $\overline{y_i}^{(t)}$ actually represents is the running average, that is, the recent average activity of the neuron. This average is calculated by the recursive formula for $\overline{y_i}^{(t)}$ given in equation 3. This serves to make learning in a neuron dependent on previous neural activity as well as current activity. This allows neurons to generalise to novel inputs given strong recent activation.

Network architecture

A two layer network was constructed - see figure 4. The first layer acts as a local feature extraction layer and consists of a 40x40 grid of neurons arranged in 100 4x4 inhibitory pools. Each pool fully samples a corresponding 4x4 patch of the 40x40 input image[2]. Competition within these pools is of the 'winner take *most*' type, otherwise referred to as leaky learning (Hertz et al., 1990). In the context of this network, this implies establishing which neuron within each pool is firing most strongly and electing it the winner. All other neurons within the same pool then have their firing rate reduced to one third of their initial rate so as to implement local inhibition. All learning in this layer is simple Hebbian.

Above the input layer is a second layer consisting of a single inhibitory pool of 15 neurons (one per face) each of which fully samples the first layer. Neurons in this layer are trained with the Hebb-like rule described in equations 1 to 4 above. All neurons in both layers also have a separate, non-linear activation function which transforms the cell's calculated weighted input into an output firing rate. This was achieved by scaling the outputs within each inhibitory pool to 1 and then passing the result through a sigmoidal activation function. The action of the inhibition and non-linear activation function are represented by the function Φ in equation 4. The rescaling was intended to keep the amount of learning taking place for each stimulus roughly constant.

Methods

The same 75 face images (5 views each of 15 faces) used in the previous psychophysical exper-

[1] The rule was originally proposed by Klopf (1972; 1988) and first implemented in its current form by Sutton & Barto (1981) in Pavlovian conditioning.

[2] Subsampled versions of the faces shown to the subjects earlier.

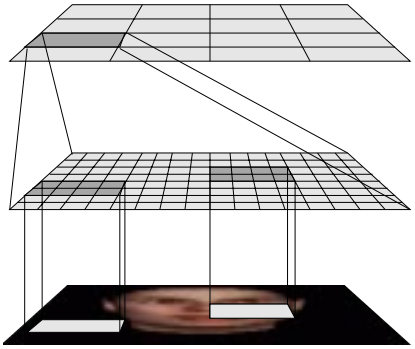

Figure 4: The hierarchical network architecture used in the simulations.

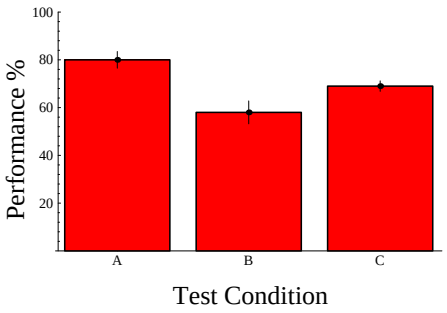

Figure 5: Results obtained in the network simulations for the same test conditions used in the human psychophysics experiments.

iment were prepared for presentation to the network by reducing their resolution to 40x40 pixels from the 200x200 pixels seen by the subjects. This was done to reduce the number of free parameters in the network and hence training time. The new image size was believed to be sufficient because the network was still able to identify the faces to 95% accuracy if each of the fifteen output neurons was trained on the five views of one particular face. Earlier pilot studies for the previous experiment suggested that this was already better than the peak in human performance of 90% under the same training conditions.

The use of a time based learning rule requires some concept of time to be built into the simulations - since the value of $\overline{y_i}^{(t)}$ in equations 1 to 4 will change over time. The basic unit of time was taken as the time for the presentation of a single view of a single face, namely 150ms. The value of η in equation 3 was set to 0.6 such that the effect of a single image on learning would decrease by a factor of around 10 after five subsequent presentations - a period of 750ms. This was chosen partly as a reasonable period for significant image association and also to ensure that the inter-sequence interval of 2500ms was sufficient to erase any residual effects of the previous image on learning. The overall intention of this was to restrict any association due to temporal order to within sequences rather than across them - as was the intention of the 2500ms inter-sequence delay used in the human training. In other words, the neurons effectively forget all previous activity during the long delay period.

Training then proceeded exactly as in the human case with the network exposed to a total of 90 sequences - equivalent to the full training received by a subject after all three training blocks. The enire process was repeated a total of 10 times using different combinations of faces to yield 10 different results from the network.

Results

The network was assumed to have responded 'same' if the winning neuron in the output layer was the same for the target and probe faces and 'different' if the winning neurons differed between target and probe image. Figure 5 shows how the network performed on the same-different recognition task originally posed to the subjects. The results found in the psychophysical experiment have indeed been reproduced - namely, good performance for same trials and a strong distinction between performance on discriminating faces trained within sequences compared with faces from separate sequences.

Discussion

The underlying hypothesis of this paper is that object recognition learning can be affected by the order in which images of objects appear as well as by their physical appearance. This hypothesis was confirmed in a psychophysical experiment with human observers. Faces were more easily confused if the subject had previously seen them presented in interleaved smooth sequences than if they were seen separately. This finding is, to my knowledge, the first evidence for a such a psychophysical effect.

I have also used a simple Hebb-like learning rule in a small neural network simulation. The network itself is obviously not intended to reproduce the sophistication of face recognition in humans. Its winner-take-all output is far from the representation described in IT cortex (Rolls & Tovee, 1995). However, despite its simplicity the network was shown to be capable of reproducing the psychophysical results described here, supporting the

idea that the type of learning rule used may underlie time-based association learning[3].

The ability of a time-based association mechanism to correctly associate arbitrary views of objects without an explicit external training signal means that it could overcome many of the weaknesses of using supervised training schemes or associating views simply on the basis of physical appearance. Its discovery in neurophysiological and now human psychophysical experiments may well represent a significant new step in establishing the 2-D multiple view approach to object recognition within a unified model of object representation and recognition learning.

Acknowledgments

I am deeply indebted to Alice O'Toole and Jeff Liter whose ideas and experience of conducting psychophysical experiments gave these experiments their direction and impetus. I am also grateful to Niko Troje for assembling the Max-Planck-Institute's most excellent collection of full 3-D head models, from which the face images used here were generated.

References

Biederman, I. (1987). Recognition–by–components: A theory of human image understanding. *Psychological Review*, 94(2):115–147.

Biederman, I. and Gerhardstein, P. (1993). Recognizing depth–rotated objects: Evidence and conditions for 3d viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, 20(1):80.

Bülthoff, H. and Edelman, S. (1992). Psychophysical support for a two–dimensional view interpolation theory of object recognition. In *Proceedings of the National Academy of Science, USA*, volume 92, pages 60–64.

Desimone, R. (1991). Face–selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, 3:1–8.

Dong, D. and Atick, J. (1995). Statistics of natural time-varying images. *Network*, 6(3).

Edelman, S. and Weinshall, D. (1991). A self–organising multiple–view representation of 3d objects. *Biological Cybernetics*, 64:209–219.

Field, D. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America, A*, 4:2379–2394.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

Hebb, D. (1949). *The Organisation of Behaviour*. New York: Wiley.

Hertz, J., Krogh, A., and Palmer, R. (1990). *Introduction to the theory of neural computation*. Santa Fe Institute: Addison Wesley.

Klopf, A. (1972). Brain function and adaptive systems – a heterostatic theory. Technical Report AFCRL–72–0164, Air Force Cambridge Research Laboratories, L.G. Hanscom Field, Bedford, MA.

Klopf, A. (1988). A neuronal model of classical conditioning. *Psychobiology*, 16:85–125.

Krouse, F. (1981). Effects of pose, pose change, and delay on face recognition performance. *Journal of Applied Psychology*, 66:651–654.

Logie, R., Baddeley, A., and Woodhead, M. (1987). Face recognition, pose and ecological validity. *Applied Cognitive Psychology*, 1:53–69.

Logothetis, N. and Pauls, J. (1994). Viewer-centered object representations in the primate. *Cerebral Cortex*, 3:270–288.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman and Co.

Marr, D. and Nishihara, H. (1978). Representation and recognition of the spatial organization of three dimensional structure. *The Proceedings of the Royal Society, London [B]*, 200:269–294.

Miyashita, Y. (1988). Neuronal correlate of visual associative long–term memory in the primate temporal cortex. *Nature*, 335:817–820.

Miyashita, Y. and Chang, H. (1988). Neuronal correlate of pictorial short–term memory in the primate temporal cortex. *Nature*, 331:68–70.

Patterson, K. and Baddeley, A. (1977). When face recognition fails. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 3:406–417.

Perrett, D. and Oram, M. (1993). Neurophysiology of shape processing. *Image and Vision Computing*, 11(6):317–333.

Plaut, D. and Farah, M. (1990). Visual object representation: Interpreting neurophysiological data within a computational framework. *Journal of Cognitive Neuroscience*, 2(4):320–343.

[3]The rule has also been shown to successfully establish distributed codes closer to those described by Tovee & Rolls in larger, more detailed simulations (Wallis & Rolls, 1996).

Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266.

Rolls, E. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical areas. *Philosophical Transactions of the Royal Society, London [B]*, 335:11–21.

Rolls, E. and Tovee, M. (1995). Sparseness of the neural representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology*, 73:713–726.

Sutton, R. and Barto, A. (1981). Towards a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, 88:135–170.

Tanaka, K., Saito, H., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, 66:170–189.

Tarr, M. and Bülthoff, H. H. (1995). Is human object recognition better described by geon-structural-descriptions or by multiple-views? *Journal of Experimental Psychology: Human Perception and Performance*, 21:1494–1505.

Tarr, M. and Pinker, S. (1989). Mental rotation and orientation–dependence in shape recognition. *Cognitive Psycholgy*, 21:233–282.

Troje, N. and Bülthoff, H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research*, 36:1761–1771.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32:193–254.

Ungerleider, L. and Mishkin, M. (1982). Two cortical visual systems. In Ingle, D., Goodale, M., and Mansfield, R., editors, *Analysis of Visual Behaviour*, pages 549–586. Cambridge, Massachusetts, USA: MIT press.

Wallis, G. (1996a). Optimal unsupervised learning in invariant object recognition. *Neural Computation,*
submitted for review. www: ftp://ftp.mpik-tueb.mpg.de/pub/guy/nc.ps.Z.

Wallis, G. (1996b). Using spatio-temporal corrolations to learn invariant object recognition. *To appear*
in Neural Networks. www: ftp://ftp.mpik-tueb.mpg.de/pub/guy/nn.ps.Z.

Wallis, G. and Rolls, E. (1996). A model of invariant object recognition in the visual system. *Progress in Neurobiology, submitted for review.* www: ftp://ftp.mpik-tueb.mpg.de/pub/guy/pnb.ps.Z.

Wogalter, M. and Laughery, K. (1987). Face recognition: effects of study to test maintenance and change of photographic mode and pose. *Applied Cognitive Psychology*, 1:241–253.

Yuille, A. (1991). Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–71.