Max–Planck–Institut
für biologische Kybernetik
Spemannstraße 38 • 72076 Tübingen • Germany

Technical Report No. 078

Effects of Temporal Association on Recognition Memory

Guy Wallis[1] & Heinrich H. Bülthoff[2]

March 2000

Guy Wallis is now at:
Department of Human Movement Studies
University of Queensland, St. Lucia 4072, QLD Australia
Email: gwallis@hms.uq.edu.au
Tel: +61 7 3365 6817 Fax: +61 7 3365 6877

[1] AG Bülthoff, E–mail: guy.wallis@tuebingen.mpg.de
[2] AG Bülthoff, E–mail: heinrich.buelthoff@tuebingen.mpg.de

Effects of Temporal Association on Recognition Memory

Guy Wallis & Heinrich H. Bülthoff

Abstract. The influence of temporal association on the representation and recognition of objects was investigated. Subjects were presented sequences of novel faces in which the identity of the face changed as the head rotated. The subjects showed a tendency to treat the views as if they were of the same person. The results counter the proposal that object views are recognised simply on the basis of objective, structural components. Instead, they suggest that we are continuously associating views of objects to support later recognition, and that we do so not only on the basis of their physical similarity, but also their correlated appearance in time.

1 Introduction

The human visual system can rapidly and accurately identify objects inspite of the nearly endless variety of images that any one object can produce on the retina, as viewing distance, viewing angle, or lighting conditions change. Although recognition of an object undergoing small view-point changes can be achieved on the basis of physical similarity to a stored view (Poggio & Edelman, 1990), it is unclear how larger changes can be achieved in this way, because the object's appearance will have changed considerably (Koenderink & van Doorn, 1979). Theorists have proposed the solution that views of an object are associated not only on the basis of physical but also temporal similarity. Temporal similarity provides information about object identity because different views of an object are usually seen in close succession (Földiák, 1991; Edelman & Weinshall, 1991; Wallis & Bülthoff, 1999; Stryker, 1991). This paper presents the results from three recognition experiments that support this proposal. In these experiments, different views of two different people's faces were erroneously perceived by observers as belonging to the same person, if previously seen in a temporally smooth sequence.

2 Stimuli

As a first test of the temporal association hypothesis, a task was devised in which subjects had to discriminate a set of briefly presented faces. As part of the preparation for the experiment, the heads of twelve female volunteers were scanned using a Cyberware 3D® laser scanner. The scanner samples texture and shape information on a regular cylindrical grid with a resolution of 0.8 degrees horizontally and 0.615 mm vertically. Using 3D head models greatly facilitated the morphing process in comparison with 2D, image based morphing techniques. This way we produced a set of twelve 3D models which were then divided equally into three groups (Fig. 1). The group membership of each head was randomized for each observer, to mitigate the effect of similarities between particular heads. Within each group, six morph heads were generated by taking the mean shape coordinates of each head pair, coloured with their mean facial pigment (Vetter, 1998). For each face pair a sequence of views was then generated containing veridical views of, say, face α in left and right profiles, face β in frontal view and then the facial morph between α and β in the $\pm 45^\circ$ views (Fig. 2).

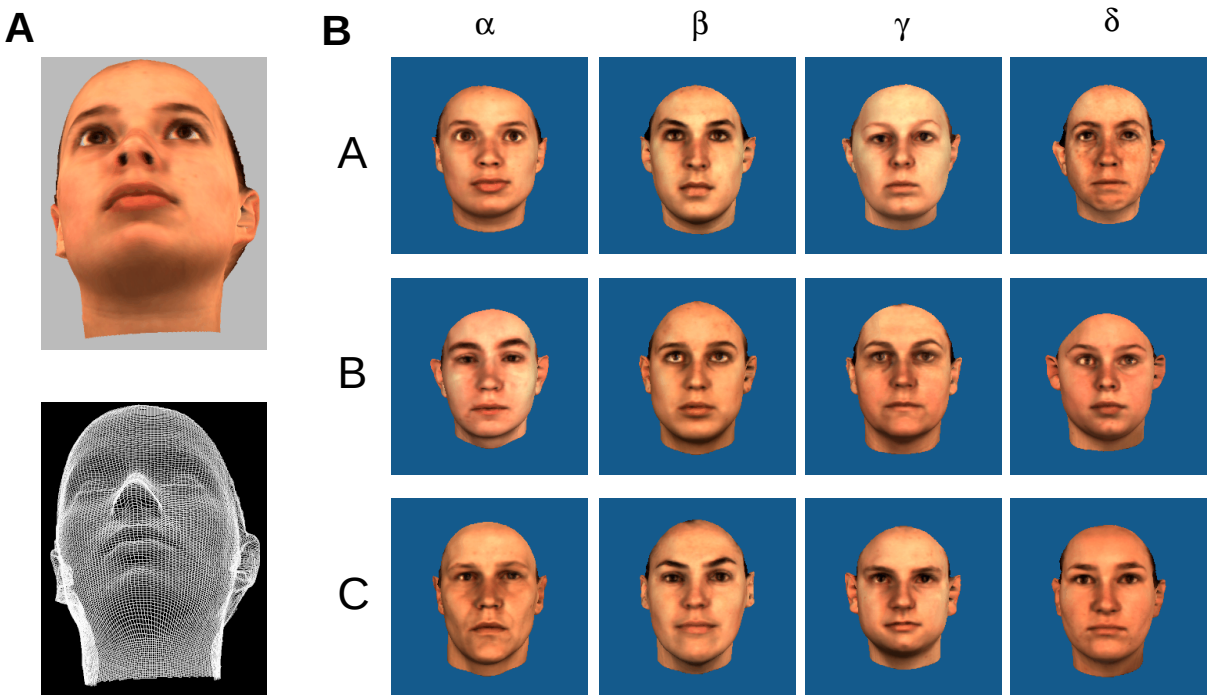

Figure 1: **(A)** Twelve, three-dimensional head models were generated by scanning the heads of 12 female volunteers. These scans, which contained both textural and shape information, were then cropped to remove extraneous cues such as hair. **(B)** The heads were split into three groups (A, B and C) of four (α, β, γ and δ). The figure shows the grouping used for one of the ten observers.

3 Procedure

Ten naive observers took part in the experiment, which was divided up into four blocks, split evenly over two days. Each block contained a training phase followed by a testing phase. During training, observers were shown all 36 of the prepared sequences twice. Images in each sequence appeared for 300ms before being replaced immediately by the next view, giving the impression of a head moving from left to right profile. By also presenting the same sequence of views in reverse order, the head was made to rotate back and forth two times. Testing took the form of a standard delayed match to sample task, in which observers were shown one view, an image mask, a second view, and then the mask again. Their task was then to indicate whether the two faces were different views of the same head or not (Fig. 3). The testing phase consisted of 384 test trials in which the total number of match and non-match trials was the same. The number of between group non-matches and within group non-matches was also balanced with 96 of each. Test images always depicted a face either directly from the front or in profile, i.e. no morphed images were tested.

4 Results

If object views are associated on the basis of their appearance in time, then exposure to image sequences should cause the images within these sequences to be represented as views of a single object. Although single cell recording results in inferior temporal lobe cortex suggest that any sequence of images can lead to image association (Miyashita, 1988), we chose to present smooth morphing sequences. We did so in the expectation that this would facilitate the association process by allowing generalization to be built up over relatively small steps, as would be the case for real world objects. Similar considerations also motivated our choice of faces as an object category, in which the exemplars have similar form, especially after removal of extraneous cues (hair).

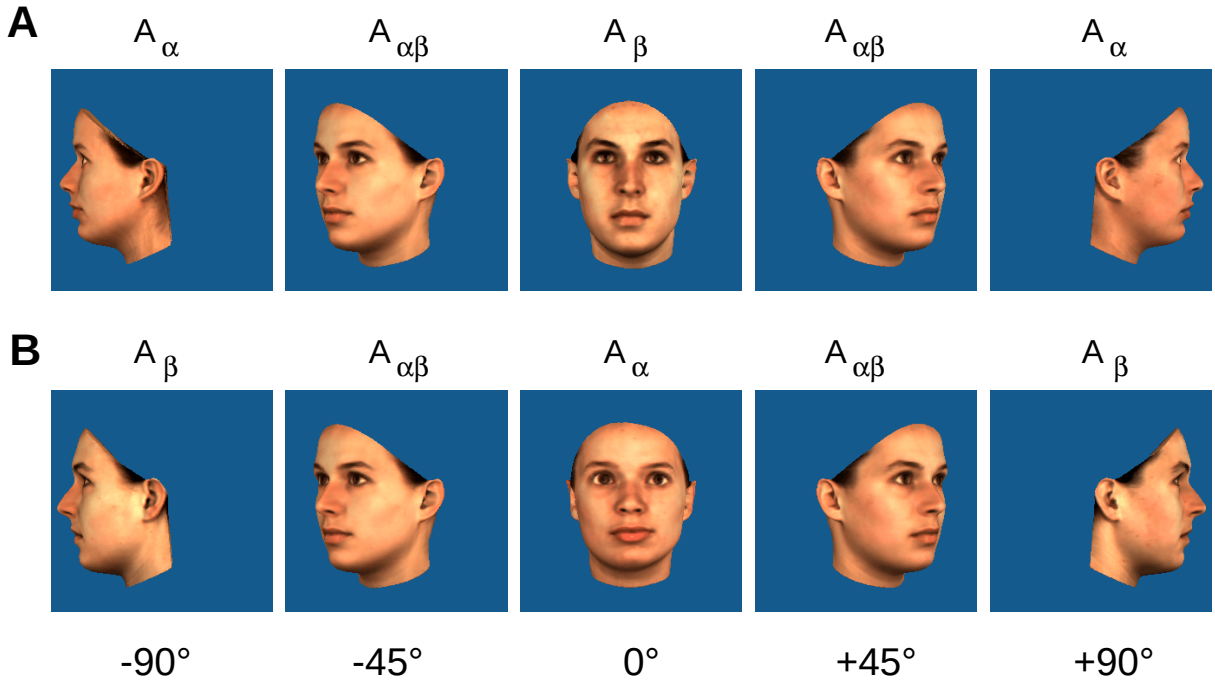

Figure 2: **(A)** The heads were used to render 2D facial images in the frontal (0°) and both profile views (±90°). A new set of head models were then generated by morphing both the shape and textural information of pairs of heads selected from a single training group. All possible pairwise within-group morphing combinations were calculated, resulting in 6 new heads per group. These new heads were then rendered to 2D facial images in left and right ±45° views. The images were organized into sequences of views in which a head appeared in left profile; a morph of this head with a second head seen from −45°; the second head in frontal view; the same morph seen from +45°; and finally the first head again, in right profile. **(B)** The complementary sequence was also prepared, in which the second head was seen in profile and the first head from the front, resulting in a total of 12 such sequences per training group.

4.1 Experiment 1

Generalisation from frontal to profile view is a difficult task for observers (Patterson & Baddeley, 1977; Bruce, 1982; Troje & Bülthoff, 1996). Hence, by exposing observers to sequences comprising two different faces, we expected to elicit greater confusion of the identity of these faces relative to that of faces not associated in this way. We tested this hypothesis in the second phase of each block, by comparing discrimination performance for pairs of faces from within a group (WG), to those from between groups (BG). Since the WG faces had been seen together in a sequence and the BG faces had not, we expected to see lower discrimination performance for WG faces than for BG faces.

The results were analysed using a within subject ANOVA, with block and stimulus group as factors. Subjects showed significantly poorer discrimination performance in the WG condition than in the BG condition, $F(1,8) = 11.768$, $MS_e = 0.0035$, $P < 0.01$. In general, task performance appeared to increase slightly over the four blocks, due to increasing observer familiarity both with the task and the stimuli (Fig. 4A). In the WG condition, however, no such trend was apparent, and the difference in performance between the WG and BG conditions showed a commensurate increase. Overall performance on the second day of testing was higher (Fig. 4B), and statistical significance of the difference between the WG and BG conditions also rose ($P < 0.001$). The results hence confirm that exposure to sequences containing pairs of faces caused a reduction in discrimination performance for these faces, as the temporal association hypothesis would predict.

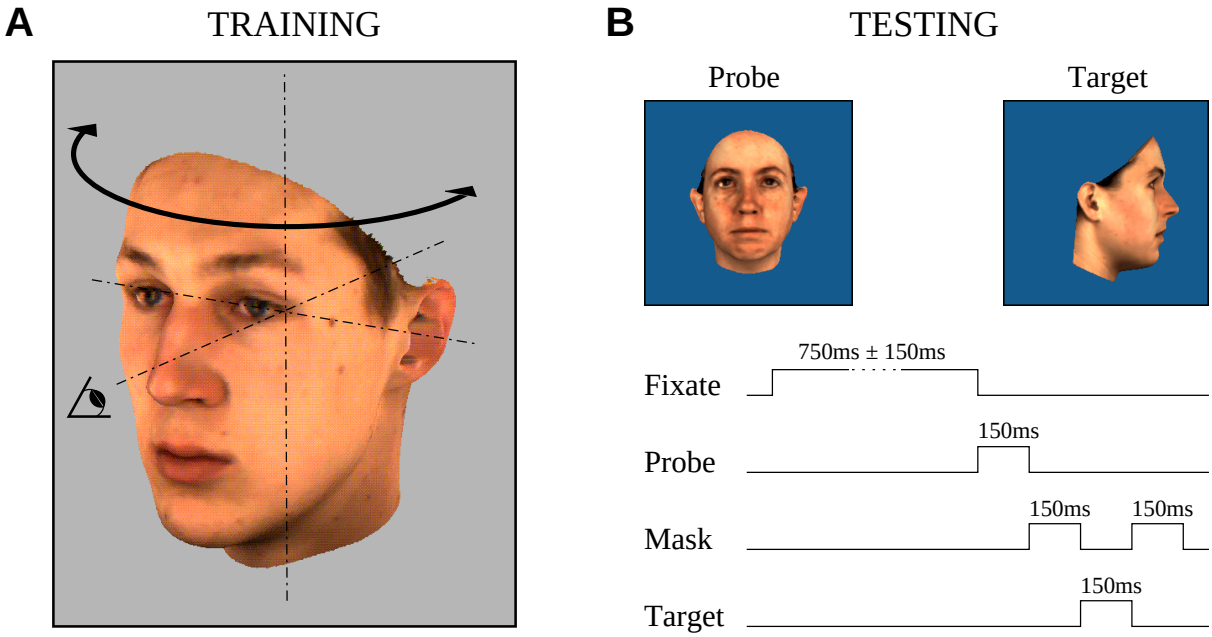

Figure 3: **(A)** During training, subjects were exposed to the morph sequences such that for each sequence, a single head appeared to rotate twice from left profile to right and back. Examples of the training sequences can be viewed at the following web site http://www.kyb.tuebingen.mpg.de/people/personal/guy/morph.html **(B)** After training, individual faces were tested in a delayed match to sample task in which observers were asked to indicate whether the two faces were different views of the same head or not. Test images always depicted a face either directly from the front or in profile, i.e. no morphed images were tested.

4.2 Experiment 2

One question raised by the first experiment is whether the use of morph faces affected recognition, in other words whether seeing intermediate views of the faces was decisive in confusing identity of the WG faces, rather than their being seen in smooth temporal order. To test this we presented the same sequences to a new set of ten observers. The only difference to the conditions of the previous experiment was that the five views of each sequence were presented simultaneously. The five views appeared along the circumference of a circle centred at the point of fixation for a period equal in length to the total viewing time of the sequences used in the first experiment (6000ms). Subjects were once again exposed to four blocks of alternating training and testing, but in this case no effect of group was observed $F(1,8) = 1.865$, $MS_e = 0.0056$, $P > 0.2$. From these results it is possible to conclude that neither the viewing of morphs nor the simultaneous appearance of the five views, was sufficient to associate the WG faces. The latter conclusion is particularly interesting in that it indicates that the refixation of the faces disrupted association, suggesting that the mechanism for learning invariance is spatio-temporally triggered rather than simply temporally. This would seem reasonable, because a change of fixation often indicates the fixation of a new object. Hence, associating across fixations would tend to lead to the erroneous association of different objects.

Another question raised by these results is whether the representation of the faces was altered, or merely associations between separate faces. Performance was good, which suggests that the subjects were not confused by training. The overall performance in the task was on average 74.7%. In a study with the same face database, subjects with no prior exposure to the faces performed on average at 65.4% correct (Troje & Bülthoff, 1996). This is lower even, than the worst performance of 72.6% recorded in the first block of this experiment,

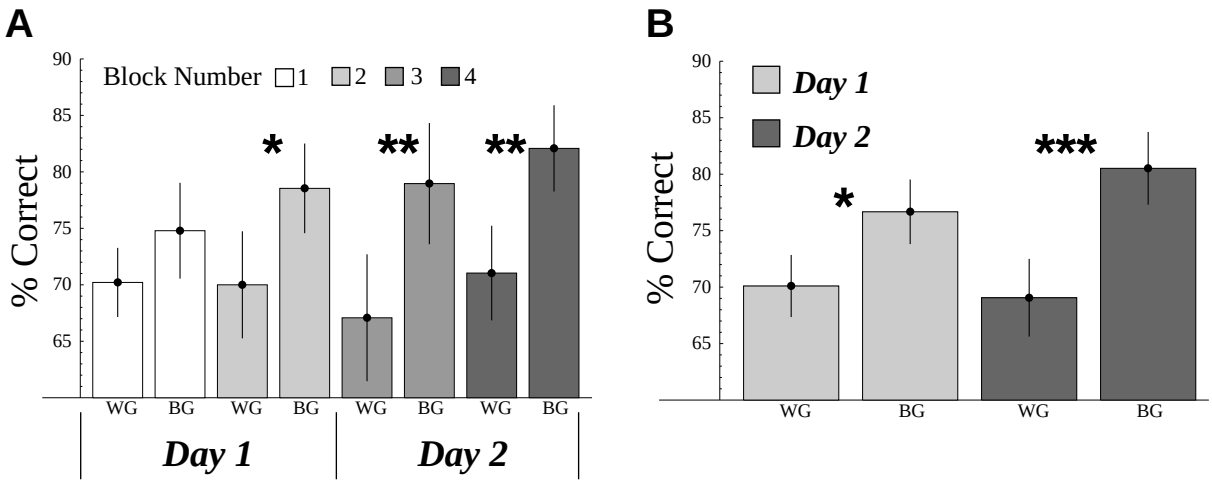

Figure 4: **(A)** Discrimination performance across training blocks, measured as percent correct in the mismatch trials - i.e. correct rejections. Figures show results for within group (WG) and between group (BG) comparisons. **(B)** The same results broken down across the two training days. Significance of the difference between conditions WG and BG are indicated as follows: $(P < 0.1) = *$, $(P < 0.01) = **$, $(P < 0.001) = ***$.

confirming that exposure to the morph sequences had not impaired overall performance in the task. However, to test whether the various views had truly been associated into the representation of a single face we conducted a third experiment, in which the same morph sequences were used. The assumption of this experiment was that if one sees the frontal view of face A turn to the profile view of face B, there will be an associated subjective impression of a change in identity, and of facial distortion during rotation. Conversely, if no such change is detected then presumably the profile and frontal views must appear to belong to the same face.

4.3 Experiment 3

To test this, six subjects were shown five morph sequences of the type used in the previous experiments. During testing, subjects saw these sequences again, along with five new sequences. The new sequences were not, however, morphs but instead depicted the true appearance of five from the ten faces seen during training. Subjects were then asked to indicate whether the heads appeared to change their form in any way during rotation. Despite the correct appearance of the five faces not seen during training, subjects showed a significant tendency to chose the faces which had been seen during training as non-deforming, (71.1%). Subjects also showed a strong tendency to perceive the true (non-deforming) heads as deforming (84.8%), presumably because the profile view which they saw, did not fit the one learned during training (Table 1A). To confirm that this effect was due to training, we tested the performance of a further six subjects not exposed to the training sequences. These untrained subjects showed precisely the opposite tendency to that shown by the trained subjects (Table 1B). The results demonstrate that naive subjects were able to predict the true profile appearance of the faces without ever having seen them before. Thus the training apparently caused the association of the frontal and profile views of the faces seen in the temporal sequence, to the exclusion of the true pairings.

5 Discussion

The question of how humans represent objects and the mechanisms behind recognition remain under debate. We have argued that objects are represented as collections of associated, two-dimensional views (Tarr & Pinker, 1989; Bülthoff & Edelman, 1992), consistent with the findings reported here. Other explainations including either object-centered

A **TRAINED**

STIMULUS	RESPONSE: DEFORMING	RESPONSE: NON DEFORMING
DEFORMING	0.2889	0.7111
NON DEFORMING	0.8484	0.1516

B **UNTRAINED**

STIMULUS	RESPONSE: DEFORMING	RESPONSE: NON DEFORMING
DEFORMING	0.6333	0.3667
NON DEFORMING	0.3444	0.6556

Table 1: Stimulus response matrices for subjects told to discriminate rotating heads containing views from more than one person (deforming), from real heads containing views from just one person (non-deforming). **(A)** Prior exposure to the deforming heads caused subjects to see these heads as non-deforming in preference to the real ones, ($d' = -1.589$). **(B)** Naive subjects, however, were able to select the real heads in preference to the deforming ones, ($d' = 0.744$).

models (Ullman, 1979; Marr, 1982; Hasselmo, Rolls, Baylis, & Nalwa, 1989), or structural descriptions that contain explicitly associated parts (Perrett, Mistlin, & Chitty, 1987; Biederman, 1987), may well have to be modified to take account of the data reported here.

Evidence from neurophysiology indicates that multiple views of objects are represented by collections of neurons, each selective to a combination of features within each view (Young & Yamane, 1992; Rolls, 1992; Abbott, Rolls, & Tovee, 1996; Logothetis, Pauls, & Poggio, 1995). From the accumulated evidence, neurophysiologists have proposed that the visual system builds up tolerance to changes in object appearance, over several processing stages. By the time processing has reached the temporal lobe, single cells are tuned to respond invariantly over any one of the many natural transformations (Perrett & Oram, 1993; Rolls, 1992). There is good reason to think that this invariance is achieved via a final layer of processing, in which the output of view dependent neurons feed forward to build view invariant neurons (Perrett & Oram, 1993). Further recording work in the same cortical areas has lead to the proposal that these associations are rapidly modifiable via experience(Rolls, Baylis, Hasselmo, & Nalwa, 1989), and are partially made on the basis of temporal contiguity of the visual input (Stryker, 1991; Miyashita, 1993; Wallis & Bülthoff, 1999), supporting the main premise of this paper.

A common misconception is that the stored views described above are equivalent to inflexible templates, selective for only one particular view. Such templates would not support recognition of objects from novel viewpoints (Pizlo, 1994). There is also concern that whole object views are encoded at the level of single neurons in the style of earlier theories of object representation, since criticised for their inefficiency and suceptability to cell damage (Sherrington, 1941; Konorski, 1967; Barlow, 1972). Both of these problems are countered by the use of a distributed, feature based recognition system (Hinton, McClelland, & Rumelhart, 1986). At the neural level, many hundreds or thousands of neurons, each selective for its specific feature, would act together to represent an object. New combinations of these features could then be recruited to uniquely represent a completely new object. The upshot of this is that although a face may be new, experience with a similar nose or configuration of mouth and eyes for example, would provide some level of generalization for a novel face across view change. Indeed, the numerous beneficial, emergent properties of a distributed representation have long been realized by neural network theorists (Hinton et al., 1986; Poggio & Edelman, 1990).

The idea that temporal information can be used in setting up spatial representations is not new, and was recognized by some of

the earliest researchers studying learning in cortical circuits (Pitts & McCulloch, 1947). The aim of our work has been to test the temporal association hypothesis in humans, and in so doing, to provide concrete evidence for a behavioral level equivalent to the theoretical predictions and neurophysiological data cited. Although the evidence presented here relates directly to face recognition, we would argue that the mechanism extends beyond face recognition to all types of object representation and recognition, consistent with recent evidence that it also affects the perception of object rigidity (Sinha & Poggio, 1996). The inescapable consequence of these findings is that the machinery underlying visual perception is contingent upon the temporal as well as physical appearance of our world.

Acknowledgments

We are grateful to Jeff Liter for help in designing the experiment, to Alexa Ruppertsberg for preparing the morph sequences, to Thomas Vetter and Volker Blanz whose program we used for morphing the 3D heads, and to Niko Troje for scanning and preparing the head models. We are also grateful to Francis Crick, Anya Hurlbert, Fiona Newell, and Alice O'Toole for comments on earlier drafts of the paper.

References

Abbott, L., Rolls, E., & Tovee, M. (1996). Representational capacity of face coding in monkeys. *Cerebral Cortex*, **6**, 498–505.

Anderson, J., & Rosenfeld, E. (1988). *Neurocomputing: Foundations of Research*. Cambridge: MIT Press.

Barlow, H. (1972). Single units and sensation: A neuron doctrine for perceptual psychology?. *Perception*, **1**, 371–394.

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

Bruce, V. (1982). Changing faces: Visual and non-visual coding proceses in face recognition. *British Journal of Psychology*, **73**, 105–116.

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

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

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

Hasselmo, M., Rolls, E., Baylis, G., & Nalwa, V. (1989). Object–centred encoding by face–selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, **75**, 417–429.

Hinton, G., McClelland, J., & Rumelhart, D. (1986). Distributed representations. In D. Rumelhart & J. McClelland (Eds.), *Parallel Distributed Processing*, Vol. 1: Foundations, chap. 3. Cambridge, Massachusetts: MIT Press.

Koenderink, J., & van Doorn, A. (1979). The internal respresentation of solid shape with respect to vision. *Biological Cybernetics*, **32**, 211–216.

Konorski, J. (1967). *Integrative Activity of the Brain: An Interdisciplinary Approach*. Chicago: University of Chicago Press.

Logothetis, N., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, **5**, 552–563.

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

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

Miyashita, Y. (1993). Inferior temporal cortex: Where visual perception meets memory. *Annual Review of Neuroscience*, **16**, 245–263.

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

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

Perrett, D., Mistlin, A., & Chitty, A. (1987). Visual cells responsive to faces. *Trends in Neurosciences*, **10**, 358–364.

Pitts, W., & McCulloch, W. (1947). How we know universals: the perception of auditory and visual forms. *Bulletin of Mathematical Biophysics*, **9**, 127–147. Reprinted in (Anderson & Rosenfeld, 1988).

Pizlo, Z. (1994). A theory of shape constancy based on perspective invariants. *Vision Research*, **34**(12), 1637–1658.

Poggio, T., & 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., Baylis, G., Hasselmo, M., & Nalwa, V. (1989). The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, **76**, 153–164.

Sherrington, C. (1941). *Man On His Nature*. Cambridge: Cambridge University Press.

Sinha, P., & Poggio, T. (1996). Role of learning in three-dimensional form perception. *Nature*, **384**, 460–463.

Stryker, M. (1991). Temporal associations. *Nature*, **354**, 108–109.

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

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

Ullman, S. (1979). *The Interpretation of Visual Motion*. Cambridge, Massachusetts: MIT Press.

Vetter, T. (1998). Synthesis of novel views from a single face image.. *International Journal of Computer Vision*, **28**(2), 103–116.

Wallis, G., & Bülthoff, H. (1999). Learning to recognize objects. *Trends in Cognitive Sciences*, **3**, 22–31.

Young, M., & Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science*, **256**, 1327–1331.