7
Prior Knowledge and Learning in 3D Object Recognition

Markus Gschwind[1], Hans Brettel[2] and Ingo Rentschler[1]

1 Introduction

Biological 3D object recognition is restricted to the sensing of 2D projections, or images, and is further constrained by the lack of transparency. The most common assumption then is that image data are referenced to mental object representations. Such representations, or object models, must be contrasted with object recognition in so far as the latter involves the understanding of image data. This distinction is central to recognition-by-components (RBC; Biederman 1987), a theory of human image understanding based on the assumption that input images are parsed into regions that display nonaccidental properties of edges. These properties provide critical constraints on the identity of 3D primitives ("geons") the images come from, e.g., cylinders, blocks, wedges, and cones, and are (relatively) invariant with viewpoint and image degradation.

RBC can be implemented by building structural representations from geons linked through explicit categorical relations (Hummel and Biederman 1992). This theory predicts that object identification will be fast and accurate if geons are readily identified in characteristic arrangements. It also implies that viewpoint invariance in 3D object recognition is achieved for all views that activate the same geon structural description (GSD; Biederman and Gerhardstein 1993). However, viewpoint invariance is not found for stimuli based on irregular blob structures ("amoebae"; Edelman and Bülthoff 1992; Bülthoff and Edelman 1992) and wire-like objects ("paper-clips"; Bülthoff and Edelman 1992). It has been held that the latter result is incompatible with recognition theories involving 3D representations. This gave rise to the multiple-views hypothesis, according to which a set of views of an object is stored in memory and the object is recognized by normalizing the input view to the most nearly compatible among such stored views (Tarr and Pinker 1989; Bülthoff and Edelman 1992).

[1]Institute of Medical Psychology, University of Munich, Goethestrasse 31, D-80336 München, Germany
[2]CNRS UMR 5141, École Nationale Supérieure des Télécommunications, Paris, France

Given these different perspectives on human object recognition, it is helpful to consider the development of object recognition by computer. Early approaches to this problem used the concept of generalized cones applied to the domain of line drawings of objects and scenes composed of polyhedral or curved parts. The understanding of such "engineering" drawings was demonstrated by producing a line drawing of the arrangement of parts as it would appear from any desired viewpoint. Yet it was clear that the interpretation of "naturalistic" images was another matter altogether (see Ballard and Brown 1982, chapter 9). To solve the latter type of problem, part-based recognition schemes are now employed in a more flexible way. For instance, the analysis of parts may be initiated by segmenting input images into regions that are recognized as parts of objects in the database. If no recognition occurs, the parameters of the initial segmentation are varied. Clearly, such approaches do not succeed in one stroke. These processes typically involve closed-loop systems where the current interpretation state is used to drive the lower level image processing functions. For these reasons, "world knowledge" and learning play key roles in second-generation image understanding and object recognition by computer (see Caelli and Bischof 1997), and the chapter by M. Jüttner, this volume).

The latter development prompted this study of the roles of prior knowledge and learning in the recognition by human observers of "structure-only" 3D objects composed of identical parts in varying spatial arrangement. As the left-right categorization of mirror-image forms is a typical feature of visual expertise (Johnson and Mervis 1997; Tanaka and Taylor 1991; Rentschler and Jüttner 2007), the test stimuli included handed objects.

2 Separating Representation and Recognition

Valid conclusions as to the nature of object representations cannot be drawn unless their dependence on stimulus information (Liu 1996; Liu et al. 1999) and task demands (Tjan and Legge 1998) is taken into account. The latter two studies made this point using an ideal observer model based on statistical pattern recognition. Thereby patterns are classified using sets of extracted features and an underlying statistical model for the generation of these patterns (see Haykin 1999).

Tjan and Legge (1998) showed that viewpoint dependence of recognition is low for structurally regular objects, but dependence increases as regularity decreases. They were further able to demonstrate a correspondence between the predicted view-point complexity (VX) of a recognition task and published human data on viewpoint dependence. For instance, they found low VX values for simple geometric objects (single geons) and mechanical compositions (distinct multiple-geon objects) consistent with the observations by Biederman and Gerhardstein (1993). By contrast, wire-like and amoebae objects showed high

VX consistent with the findings by Edelman and Bülthoff (1992). Tjan and Legge concluded that confusion about the nature of object representations can be attributed at least partly to a failure to distinguish betweeen visual processing and the type of recognition task including the physical characteristics of test objects.

The findings of Tjan and Legge would seem to be consistent with reports from object recognition by computer (see Dickinson 1993). On the one hand, 2D indexing primitives, i.e., image structures that are matched to stored object models, are useful for small object databases. The reason for this limitation is increasing search complexity and reliance on verification with decreasing complexity of primitives. On the other hand, the reliable recovery of 3D indexing primitives from input images is a very difficult problem. Nevertheless, due to a concomitant decrease in search complexity for matching, 3D indexing primitives may be more successful than 2D indexing primitives for large databases.

Against the conclusions from ideal observer models, it might be held that these models rely on traditional pattern recognition, where classification is achieved by partitioning feature space into regions associated with different pattern classes. However, there are many recognition problems that cannot be solved this way. For instance, the efficiency of object recognition systems may be judged using the criterion of "stability and sensitivity" (Marr and Nishihara 1978, p. 272). Accordingly, descriptions must reflect the similarity of objects thus enabling generalization. At the same time subtle differences need to be preserved to allow discrimination. Stable information representing global aspects of object shape must be decoupled, therefore, from information representing finer details. This can be achieved by relying on prominent pattern components for similarity judgments, whereas full pattern representations are used for discrimination (Rentschler et al. 1996).

More generally, traditional pattern recognition works well for simple isolated patterns but is inadequate for complex patterns and objects embedded in scenes. Image interpretation by computer therefore relies on the extraction of features of image parts and features of part relations that are linked together to form structural descriptions. Sets of hierarchically organized rules ("graphs") are then generated for classification to the extent needed for solving a given recognition problem. Classification performance can be improved further by feeding back results from rule evaluation to earlier stages of the rule generation system. Such methodologies of syntactic pattern recognition (see Caelli and Bischof 1997) have been adapted to the analysis of human image understanding (Rentschler and Jüttner 2007; see also the chapter by M. Jüttner, this volume) and object recognition (Osman et al. 2000). That approach would seem to be particularly appropriate for implementing cognitive functions as it integrates bottom-up and top-down processing characteristics. However, the various degrees of freedom of implementing such systems warrant further experimental research into the roles of prior knowledge and learning in human 3D object recognition.

We therefore sought to distinguish representations and recognition using a psychophysical paradigm of category learning involving priming. Priming is a technique from memory research using the beneficial influence of pre-exposure to a stimulus in the absence of explicit instructions to remember the stimulus (Biederman and Cooper 1991; Cooper et al. 1992). When used in combination with an invariant procedure of recognition involving fixed stimulus sets, any effect of priming must be attributed to object memory, i.e., representation.

3 Learning 3D Structure from Images

Our recognition paradigm used two sets of 3D objects consisting of one bilaterally symmetric object and one pair of handed (left and right) objects each (Fig. 1). Following priming (Fig. 2), participants were trained to classify a set of 22 learning views (Fig. 3). Upon reaching 90% correct, participants classified 83 test views (64 *novel* views, 19 *learned* views). Classification performance was measured in terms of signal detection accuracy (d prime; see Rentschler et al. 2004) and response time.

In the first experiment (Gschwind et al. 2004), we used objects built from spheres termed *spheres.* Resulting views were poor in ordered feature elements

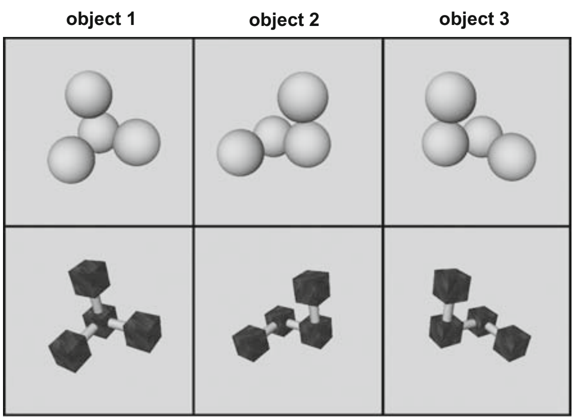

Fig. 1. Test sets of one bilaterally symmetric 3D object and one pair of handed (mirror symmetric) 3D objects. Each object was composed of four spheres (top) or cubes connected by rods (bottom). Three object parts formed an isosceles rectangular triangle, while the fourth one was placed perpendicularly above the centre of one of the base parts. Objects were generated both as physical models and virtual models. Physical models were constructed of polystyrene balls (6 cm diameter) or wooden cubes (3 cm sides) and rods (3 cm long, 1.2 cm diameter). Virtual models were generated and displayed as perspective 2D projections by the Open Inventor™ (Silicon Graphics, Inc.) 3D developer's toolkit. A lighting model of mixed directed and diffuse illumination and a lack of cast shadows was used

FIG. 2. For *vision* priming (*top*), participants watched one after the other computergraphic projections of the 3D objects successively rotating around the three principal axes. Two cycles of exposure of 90 s and 10 s per axis were used. For *motor* priming (*bottom*), the blindfolded subjects manipulated the physical models without restriction. No instructions other than the invitation to familiarize themselves with the objects were given. Priming lasted for 5 min and was followed by category learning

and connectivity of parts. This raised a question regarding the extent to which priming effects depended on stimulus information. We sought to answer this question in the second experiment using a set of modified stimuli termed *cubes*. The latter set had the same macrogeometric structure as *spheres* but textured cubes and rods as parts (see Fig. 1). The conditions of generating learning and test views, priming, as well as category learning and generalization were identical for both experiments.

Figure 4 shows the effects of priming in terms of classification performance in the first unit of category learning. With *spheres*, priming did not significantly affect the accuracy for object 1, perhaps because subjects were already at ceiling. Yet *motor* priming significantly improved classification of the handed objects 2 and 3 (Fig. 4, top left). For *cubes* (Fig. 4, top right), both *motor* and *vision* priming were equally effective in inducing classification, with the induction effect being most pronounced for non-handed object 1. Response times tended to be increased by *vision* and *motor* priming for the classification of *spheres* (Fig. 4, bottom left), although significance was only reached with *motor* priming for non-handed object

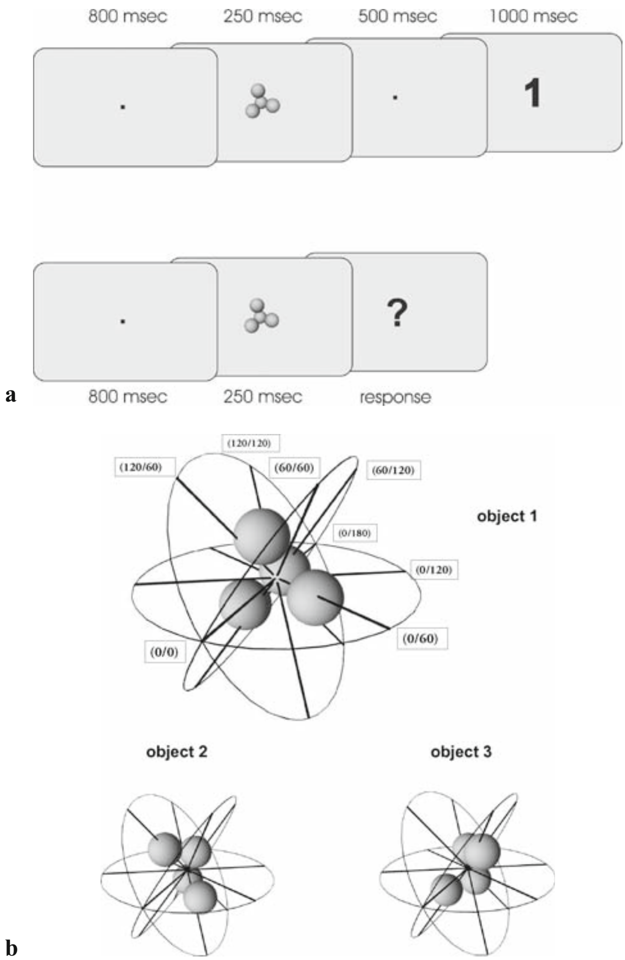

FIG. 3. **a** Supervised category learning was partitioned into a series of learning units, each consisting of a learning phase and a test phase. For learning, subjects saw in random sequence each of the learning views followed by the corresponding object label. For testing, they saw the learning views again but had to indicate the object labels by pressing a key on the computer keyboard. **b** Learning sets of 22 views (6 different views for object 1, 8 for each of the objects 2 and 3) obtained by sampling the viewing sphere in steps of 60°. In addition, a random rotation angle around the (virtual) camera axis was employed. Test sets of 83 views (21 different views for object 1 and 31 different views for each of the objects 2 and 3) were obtained by sampling the viewing sphere in steps of 30°. 19 of the test views were already used during category learning (5 for object 1 and 7 each for objects 2 and 3). Sixty-four test views were from novel viewpoints (16 for object 1, and 24 each for objects 2 and 3)

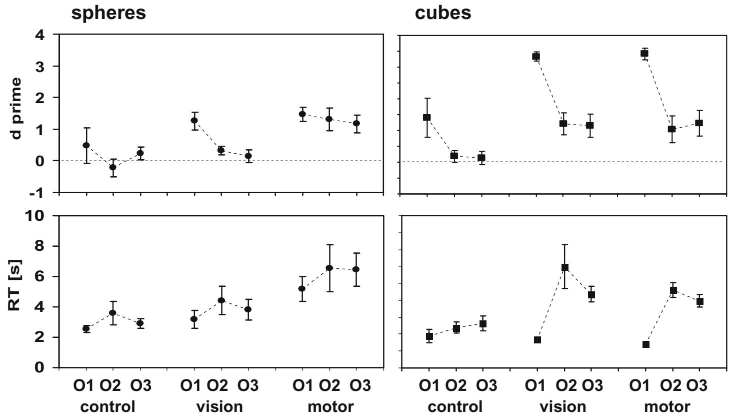

FIG. 4. Effects of priming on object recognition. Priming conditions were *control*, *vision*, and *motor* both for *spheres* (*filled circles*, left) and *cubes* stimuli (*filled squares*, right). Classification accuracies (d prime, top) and response times (RT, bottom) obtained from the first learning unit. 10 subjects entered category learning for each condition but only 7 *control* subjects reached criterion. Error bars: ± 1 S.E. (N = 7 * 6 control/object1, N = 7 * 8 control/object23; N = 10 * 6 object1, N = 10 * 8 object23 for *vision* and *motor*)

1. For *cubes*, an increase in response times was obtained by *vision* and *motor* priming for the handed objects 2 and 3 only (Fig. 4, bottom right).

Category learning continued until observers reached a criterion of 90% correct. For *spheres*, the number of learning units to criterion was not significantly dependent on experimental conditions (N = 25.7 ± 6.3 *control*, N = 33.1 ± 6.9 *vision*, N = 16.2 ± 4.3 *motor*). For *cubes*, both types of priming strongly enhanced category learning (N = 25.4 ± 5.8 *control*, N = 8.6 ± 4.0 *vision*, N = 3.8 ± 1.0 *motor*).

4 Generalization to Novel Viewpoints

The experiments continued with measuring generalization to novel viewpoints and re-classification of learned views (Fig. 5). With *spheres,* the accuracies for non-handed object 1 were found to be relatively high and virtually unaffected by priming (Fig. 5, top left). The accuracies for handed objects 2 and 3 were poor under the conditions of *control* and *vision. Motor* priming, however, strongly improved accuracies to yield values equal to those. Except for the performance involving the non-handed object under the conditions of *control* and *vision* priming, accuracies for *spheres* were significantly better for the learned views than for the novel views. *Motor* priming caused longer response times for both types of object but there was no significant difference in response times between novel and learned objects across priming conditions. With *cubes* (Fig. 5, right), maximum accuracies were obtained for both object types and there was no

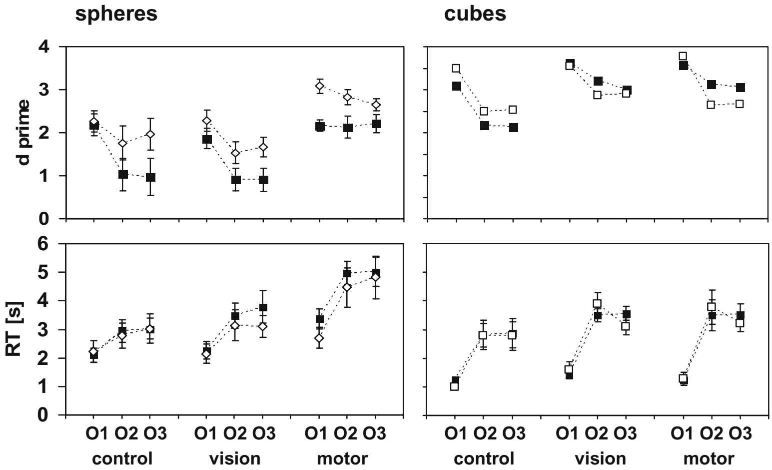

FIG. 5. Generalization to novel viewpoints for *spheres* (left) and *cubes* (right). Accuracies (d prime) cumulated over views, classification decisions, and observers at the top, corresponding response times at the bottom. Closed symbols denote generalization performance from novel viewpoints, open symbols from learned viewpoints. Each view was presented three times. Error bars: ± 1 S.E. (novel viewpoints: N = 7 ∗ 3 ∗ 16 control/object1, N = 7 ∗ 3 ∗ 24 control/objects23, N = 10 ∗ 3 ∗ 16 vision/motor/object1, N = 10 ∗ 3 ∗ 24 vision/motor/objects23; learned viewpoints: N = 7 ∗ 3 ∗ 5 control/object1, N = 7 ∗ 3 ∗ 7 control/objects23, N = 10 ∗ 3 ∗ 5 vision/motor/object1, N = 10 ∗ 3 ∗ 7 vision/motor/objects23). No error bars are given for the classification accuracies for *cubes* stimuli because of deviation from normal distribution

significant effect of priming conditions nor a significant difference between learned and novel views. That is, with *cubes* there occurred no differences in recognition performance between object types, priming conditions, or learned and novel views.

5 Inverse Problems and Spatial Transformations

Non-discrimination of handed objects is predicted by structural recognition models using non-directed part relations (e.g., Hummel 2001) and by view-based recognition models (e.g., Riesenhuber and Poggio 1999). Moreover, with *spheres* we found performance to be view-dependent consistent with the predictions of view-based recognition models. This would seem to support the "rotation-for-handedness" hypothesis (Tarr and Pinker 1989; Tarr 1995; Gauthier et al. 2002), according to which humans rely for recognition on reflection-invariant mechanisms in the brain and use mental rotation (Shepard and Metzler 1971) to disambiguate handedness.

The latter hypothesis, however, meets two difficulties when applied to the results of our experiments. First, images are generally ambiguous with regard to the 3D structures they are derived from. The solution for such inverse problems critically depends upon the operation of constraints, i.e., prior object knowledge (Pizlo 2001). This is why in previous studies on mental rotation subjects were given "a great deal of feedback about the 3D structure of each object" (Tarr 1995, p. 61). Our observers had no prior object knowledge under the *control* condition and were found to be completely unable, both for *spheres* and *cubes*, to disambiguate handedness early in practice (Fig. 4, top). Second, we used 2D views generated by conjointly varying the three Euler angles of rotation (see Korn and Korn 1968, Section 14.10). To reliably recover these angles from a given 2D view is impossible, and the rotation of the image plane was an additional source of uncertainty for the matching process. On these grounds, we reject the rotation-for-handedness hypothesis according to which our observers could have achieved disambiguation of handedness by employing continuous rotations around specific axes in 3D. Instead, for both non-handed and handed objects, they must have iteratively used combinations of spatial transformations.

Consistent with the latter conclusion, subjects with *motor* priming needed for the recognition of both non-handed and handed *spheres* prolonged response times, typically moved their hands during classification, and spontaneously reported having mentally rotated the candidate models for classification. The retardation of response times would seem to reflect, therefore, the times needed for generating internalized candidate models, transforming them during the matching process, and executing additional transformations to align mirror symmetric counterparts. This implies that, for *spheres* with *motor* priming, our recognition paradigm could not be separated into one of recognizing the non-handed object and one of discriminating handedness. Indeed, the improvement of category learning through motor priming was most pronounced for non-handed object 1. The signal detection analysis of data from the generalization phase demonstrated that this resulted at least partly from a reduction of the misclassification of views of the handed objects as views of the non-handed object.

We then turn to the question of how *motor* priming facilitated the classification of stimuli built from spheres. Clearly, such type of priming drew the attention of subjects to the third stimulus dimension. This enabled them to explicitly generate relational 3D representations (Thoma et al. 2004). Participants may have solved the inverse problem for *spheres* by encoding temporal sequences of exploratory finger and hand movements along the *physical* object models. As object palpation directly evokes mental imagery (Critchley 1953, chap. IV), it seems that some sort of kinetic object traces were stored in multimodal representations (e.g., Zangaladze et al. 1999). Subjects may then have inferred the connectivity of sphere parts, i.e., 3D structure, from linking object parts exposed in 2D views to such internalized representations. Conversely, we suggest that the type of prior knowledge provided by *vision* priming did not allow the solution of the inverse problem for *spheres*. Indeed, during *vision* priming subjects noted ambiguous rotation-in-depth of the *spheres* objects. These effects were caused by uncertainties of correspondence between object views displayed during motion.

6 Role of Image Understanding in Invariant Recognition

From the equivalence of *vision* and *motor* priming for classifying *cubes* (Fig. 5), we conclude that the clear connectivity between parts and the related ordering of feature elements helped the solution of the inverse problem right from the visual stimulus. Moreover, the parallel contours of cube parts facilitated matching thus supporting the verification of candidate 3D object models. Therefore, the classification of *cubes* would seem to be an instance of fast and accurate recognition that is viewpoint invariant as predicted by RBC. Indeed, for *cubes* we found recognition performance to be view-invariant. Furthermore, the classification of handed objects built from *cubes* entailed prolonged response times, thus indicating the need of aligning internalized object models to an external reference system.

In case of objects built from *spheres*, the extraction of part relations from 2D views was difficult. The parts as such left the axes of connectivity between them completely unspecified. The image understanding of the observers therefore benefited greatly from structural cues obtained from motor memory, thus presumably using 2D representations augmented by 3D information from motor memory (see Liu et al. 1995). The matching of such reduced object models to input data, however, entailed an increase in search complexity, i.e., the amount of spatial transformations and matching needed for categorization. As a result, the response times for classifying both types of objects built from spheres, non-handed and handed, were prolonged.

These findings emphasize the role of image understanding in object recognition. The two sets of objects had identical structural characteristics relevant for classification, and their respective members were readily decomposed into identical parts. Object recognition relied, therefore, entirely on the ability to recover part relations from 2D views.

7 Conclusions

We have shown that early in practice, humans were virtually blind to structural differences of 3D objects composed of identical sphere-shaped parts. Category learning improved recognition but more for non-handed objects than for handed objects. Prior knowledge from passively inspecting 2D views of depth-rotating objects did not affect recognition, whereas active haptic exploration of physical 3D models enabled equally accurate but view-dependent recognition of both non-handed and handed objects. Using objects with the same macrogeometrical features but clear connectivity of cube-shaped parts yielded very different results. Recognition was fast and accurate early in practice for the non-handed object.

Yet, with both types of prior knowledge, category learning enabled equally accurate and view-independent recognition for both non-handed and handed objects.

These results demonstrate, on the one hand, that there is no absolute difference between stimuli that allow distinct structural descriptions for 3D object recognition and stimuli that do not (e.g., Biederman and Gerhardstein 1993). Prior knowledge and learning play an important role in determining the extent to which image regions and their relations can be referenced to mental object representations. On the other hand, the structure-based recognition of 3D objects is not accommodated by the multiple-views theory of recognition (e.g., Bülthoff and Edelman 1992). These observations would seem to be consistent with the conclusions by Christou and Bülthoff (2000), according to whom the nature of object representations depends on whether there is enough stimulus information for the recognition task at hand.

We therefore propose that observers build 3D representations for object recognition as long as sufficient stimulus information and prior knowledge are available. Yet internalized 3D models may be too similar to allow their disambiguation concerning class membership, a situation typically encountered in classification at the subordinate level. Alternatively, observers may fail early in practice to extract from input images view-invariant geometric primitives in distinct relations. Category learning might then enable them to derive such structural descriptions. Otherwise, they would resort to the use of object representations in image format and corresponding matching behavior, thus increasing classification performance for learned views at the expense of decreased performance in generalization to novel views.

Acknowledgment. This chapter benefited greatly from the reviews and comments of Irving Biederman, Terry Caelli, Martin Jüttner, and Zili Liu.

References

Ballard DH, Brown CM (1982) Computer vision. Prentice Hall, Englewood Cliffs NJ

Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev 94:115–147

Biederman I, Cooper EE (1991) Priming contour-deleted images: evidence for intermediate representations in visual object recognition. Cognit Psychol 23:393–419

Biederman I, Gerhardstein PC (1993) Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. J Exp Psychol 19:1162–1182

Bülthoff H, Edelman S (1992) Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proc Natl Acad Sci USA 89:60–64

Caelli T, Bischof WF (1997) Machine learning and image interpretation. Plenum Press, New York

Christou C, Bülthoff HH (2000) Perception, representation and recognition: a holistic view of recognition. Spat Vis 13:265–276

Cooper LA, Schacter DL, Ballesteros S, Moore C (1992) Priming and recognition of transformed three-dimensional objects: effects of size and reflection. J Exp Psychol Learn Mem Cogn 18:43–57

Critchley M (1953) The parietal lobes. Edward Arnold, London

Dickinson SJ (1993) Part-based modeling and qualitative recognition. In: Jain AK, Flynn PJ (Eds) Three-dimensional object recognition systems. Elsevier, Amsterdam, pp 201–228

Edelman S, Bülthoff HH (1992) Orientation dependence in the recognition of familiar and novel views of 3D objects. Vision Res 32:2385–4000

Gauthier I, Hayward WG, Tarr MJ, Anderson AW, Skudlarski P, Gore JC (2002) BOLD activity during mental rotation and viewpoint-dependent object recognition. Neuron 34:161–171

Gschwind M, Brettel H, Osman E, Rentschler I (2004) Structured but view-dependent representation for visual 3-D object classification. Perception 33(Suppl):73

Haykin S (1999) Neural networks. Prentice Hall, Upper Saddle River NJ

Hummel JE (2001) Complementary solutions to the binding problem in vision: implications for shape perception and object recognition. Vis Cogn 8:489–517

Hummel JE, Biederman I (1992) Dynamic binding in a neural network for shape recognition. Psychol Rev 99:480–517

Johnson KE, Mervis CB (1997) Effects of varying levels of expertise on the basic level of categorization. J Exp Psychol Gen 126:248–277

Korn GA, Korn TM (1968) Mathematical handbook for scientists and engineers. McGraw-Hill, New York, Section 14.10

Liu Z (1996) Viewpoint dependency in object representation and recognition. Spat Vis 9:491–521

Liu Z, Knill DC, Kersten D (1995) Object classification for human and ideal observers. Vision Res 35:549–568

Liu Z, Kersten D, Knill DC (1999) Dissociating stimulus information from internal representation – a case study in object recognition. Vision Res 39:603–612

Marr D, Nishihara HK (1978) Representation and recognition of the spatial organisation of three-dimensional shapes. Proc R Soc Lond B 200:269–294

Osman E, Pearce AR, Jüttner M, Rentschler I (2000) Reconstructing mental object representations: a machine vision approach to human visual recognition. Spat Vis 13:277–286

Pizlo Z (2001) Perception viewed as an inverse problem. Vision Res 41:3145–3161

Rentschler I, Jüttner M (2007) Mirror-image relations in category learning. Vis Cogn 15:211–237

Rentschler I, Barth E, Caelli T, Zetzsche C, Jüttner M (1996) Generalization of form in visual pattern classification. Spat Vis 10:59–85

Rentschler I, Jüttner M, Osman E, Müller A, Caelli T (2004) Development of configural 3D object recognition. Behav Brain Res 149:107–111

Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025

Shepard RN, Metzler J (1971) Mental rotation of three-dimensional objects. Science 171:701–703

Tanaka JW, Taylor M (1991) Object categories and expertise: is the basic level in the eye of the beholder? Cognit Psychol 23:457–482

Tarr M (1995) Rotating objects to recognize them: a case study of the role of viewpoint dependency in the recognition of three-dimensional objects. Psychonom Bull Rev 2:55–82

Tarr MJ, Pinker SM (1989) Mental rotation and orientation dependence in shape recognition. Cognit Psychol 21:233–282

Thoma V, Hummel JE, Davidoff J (2004) Evidence for holistic representations of ignored images and analytic representations of attended images. J Exp Psychol 30:257–267

Tjan BS, Legge GE (1998) The viewpoint complexity of an object-recognition task. Vision Res 38:2335–2350

Zangaladze A, Epstein CM, Grafton S, Sathian K (1999) Involvement of visual cortex in tactile discrimination of orientation. Nature 401:587–590