INFORMATION SOCIETY TECHNOLOGIES (IST) PROGRAMME

Cognitive Vision Systems – CogVis (IST-2000-29375)

DR 1.2. Psychophysical results from experiments on recognition & categorisation (MPIK)

Graf, M., Schwaninger, A., Wallraven, C., & Bülthoff, H.H.

A firm understanding of how the human visual system recognises and categorises objects is important in order to build a successful cognitive vision system. We have reviewed the relevant literature both on visual object recognition and categorisation (chapter 1). Based on this review and the technical annex of this project we have addressed several topics in a series of psychophysical experiments, focusing on structural aspects of recognition memory, object similarity in the context of categorisation, shape transformations in categorisation, the role of context in recognition and categorisation, and the interplay between object motion and shape for categorisation decisions (chapter 2). Based on our psychophysical results we present our view on recognition and categorisation, proposing an integrative framework that serves as a theoretical basis for a computational recognition system grounded in cognitive research (chapter 3).

1 Models of object recognition and categorisation

Usually we are able to visually recognise objects even when we see them from different points of view, in different sizes due to changes in distance, and in different positions in the environment. Even young children recognise objects so immediately and effortlessly that it seems to be a rather ordinary and simple task. However, changes in the spatial relation between observer and object lead to immense changes of the image that is projected onto the retina. Hence, to recognise objects regardless of orientation, size and position is not a trivial problem, and no computational system proposed so far can successfully recognise objects over a wide range of object categories and contexts. The question how we recognise objects despite spatial transformations is usually referred to as the *first basic problem* of object recognition. We are not only able to recognise identical objects, but we can effortlessly see an unknown dog as a *dog*, or as a *beagle*, even though we have never seen it before. But how do we see different instances as members of the same object class? This ability for class level recognition or categorisation is considered as the *second basic problem* of object recognition. Categorisation is a fundamental capacity; an organism without the ability to categorise would be continually confronted with an ever-changing array of seemingly meaningless and unrelated perceptual experiences.

Objects can be recognised or categorised on different levels of abstraction. For example, a specific object can be categorised as an *animal*, as a *dog*, as a *collie*, or as *my dog Snoopy*. One of these levels has perceptual priority, and is called the *basic level* of categorisation (Rosch, Mervis, Gray, Johnson & Boyes-Braem, 1976). The basic level is usually also the entry level of categorisation (Jolicoeur, Gluck & Kosslyn, 1984): Typically we recognise or name objects at the basic level, i.e. we see or name something as a *dog*, a *cat*, a *car*, a *table*, a *chair*, etc. The level above the basic level is called *superordinate level* (e.g., *vehicle*, *animal*), while the level below the basic level is the *subordinate level* (*limousine*, *van*, *hatchback* or *collie*, *dachshund*, *beagle*, etc.). Finally, objects can also be recognised at an exemplar level, i.e. an entity can be seen (or identified) as *my dog Snoopy*, or as *John's car*.

Up to the present day, recognition and visual categorisation usually are treated as separate areas in the literature, even though the two terms are used almost exchangeable: Basically, to recognise an object on basic and subordinate level means to categorise it, and thus this separation seems artificial. The CogVis project aims at both recognition and categorisation of objects, and therefore models from both areas have to be considered. As a large number of approaches were suggested in the last decades, it is important to provide a survey over different models of recognition and categorisation. In order to get hold of the diversity of approaches, both recognition and categorisation models are classified within a single framework, using a modified version of Ullman's (1989) classification scheme for recognition models. We distinguish between view-independent models (invariant property models, traditional feature models, structural description models) and view-dependent (image-based) models (alignment models on one hand and interpolation, pooling and threshold models on the other hand).[1,2] Within every model type, both recognition and categorisation models will be introduced.

1.1 Invariant property models

Invariant property models of object constancy and object recognition are based on the assumption that objects can be characterized by certain invariant properties, which are unaffected by transformations of the proximal stimulus on the retina that result from a change in the spatial relation between observer and object. The idea is that recognition can be explained on the basis of formless mathematical invariants that are common to all views of an object, usually defined relative to certain geometrical transformations (e.g., Todd, Chen & Norman, 1998; Van Gool, Moons, Pauwels & Wagemans, 1994; see already Cassirer, 1944; Gibson, 1950; Pitts & McCulloch, 1947). For example, the cross ratio is a frequently used invariant of the projective group (e.g., Cutting, 1986). Invariant property approaches may be mathematically appealing, but often the postulated invariants

[1] We use the term view- or image-based to indicate that the representation is in some sense close to an image, based that recognition or categorisation performance depends systematically on image transformations. Note that we neither want to say that representations are pixel-based, nor that they are holistic in the sense of rigid holistic templates.

[2] The distinction between different view- or image-based models is not always trivial. Different classifications of image-based models seem possible.

could not be experimentally confirmed (e.g., Niall, 1992; Niall & Macnamara, 1990). In principle, invariant property models predict that recognition performance is independent of the amount of transformation, as the invariants are by definition unaffected by transformations (for further discussion see Wagemans, Van Gool & Lamote, 1996; Larsen & Bundesen, 1998). However, this stands in contrast to a large number of empirical studies (see chapter 1.4), and raises doubts whether invariant property models can be used to explain the recognition of individual exemplars.

What about object *categorisation* at the basic level? Invariant property models were not designed as models of categorisation. However, it can be argued that invariant property models can be enhanced to account for categorisation by extending them to topological invariants. Actually, there is evidence that topological invariants play a role in visual perception (e.g., Chen, 1982, 1985, 2001). However, it is not clear how categorisation can be modelled by invariant properties. To use Shimon Ullman's words: "What simple invariances would distinguish, for example, a fox from a dog?" (1989, p. 201). Probably for these reasons, invariant property models – i.e. models based on mathematical (formless) invariants – were not proposed in the categorisation literature, and are typically not discussed.

1.2 Traditional feature models

Traditional feature models of recognition suggest that recognition is achieved through an extraction of – both visual and more abstract – features and a subsequent comparison with stored representations. Two subtypes of traditional feature models can be distinguished. In one subtype objects are represented as lists of features (e.g., Selfridge & Neisser, 1963), while in the other subtype objects are represented as points in a multidimensional space whose dimensions correspond to the feature dimensions (e.g., Shepard, 1957, 1987).

Many of the traditional recognition models can be regarded as feature models, and also nearly all present models of *categorisation*. These categorisation models likewise can be sub-classified into models in which objects (respectively categories) are represented as lists of (usually discrete and binary) features (e.g., Estes, 1986, 1994; Medin & Schaffer, 1978; Tversky, 1977), or as points or regions in a multidimensional feature space (e.g., Erickson & Kruschke, 1998; Kruschke, 1992; Lamberts, 1994; Nosofsky, 1986, 1988). In both feature list and feature space models, categorisation is based on the similarity between the perceived object and the category representation (which is conceptualised either as a prototype or as a set of exemplars). Depending on the subtype, similarity is either computed on a set-theoretical basis through the weighted number of common and distinctive features (Tversky, 1977), or determined geometrically as the distance in multidimensional space. Feature models that specify processes and thus predict the time course of categorisation were developed only recently (Ashby & Maddox, 1994, 1996; Cohen & Nosofsky, 2000; Lamberts, 1998; Nosofsky & Palmeri, 1997).

These models were quite successful with rather artificial stimulus material, but allow only a very limited description of the shape of objects, even though shape is of crucial importance for recognition and categorisation (e.g., Biederman & Ju, 1988;

Landau, Smith & Jones, 1988; Rosch et al., 1976). The features which are typically used in these categorisation models are of a rather abstract nature; for example a banana could be described by the attributes "yellow, long, sweet, peel, sections" (see e.g., the presentation in Barsalou & Hale, 1993). Also in experimental tests the problem of shape is largely avoided: If shapes are employed as stimuli (and not colours or other non-shape-stimuli), they are constructed such that they vary along few salient dimensions and consequently can be described in a relative simple way (e.g., Ashby & Maddox, 1994, 1996; Brooks, 1978; Medin & Schaffer, 1978; Nosofsky, 1986) – and hence look rather artificial and lack ecological validity.[3] Therefore the problem remains unsolved how the similarity of shapes can be conceptualised, and how shape similarity influences categorisation performance. Moreover, it was argued that feature-based approaches are inadequate as models of similarity and categorisation, because they are unable to represent relationships between features (see Hahn & Chater, 1997; Medin, 1989).

1.3 Structural description models

The basic idea of structural description models is that object recognition or categorisation is based on a structural representation, which is defined as a configuration of elementary object parts that are regarded as shape primitives (e.g., Marr & Nishihara, 1978; Sutherland, 1968). Structural description models aim at supplying abstract and propositional descriptions of objects, which are immune to irrelevant spatial information. Therefore, structural description models typically predict that recognition performance is invariant regarding spatial transformations. Biederman's recognition-by-components (RBC) or geon structural description (GSD) model can be regarded as the best developed example of the structural description model type (Biederman, 1987; Biederman & Gerhardstein, 1993, 1995; Hummel & Biederman, 1992). According to this model objects are represented as configurations of elementary three-dimensional primitive parts, called geons. These geons are derived from nonaccidental properties (NAPs) in the image, i.e. from properties which unlikely arise by chance, and are more or less invariant over a wide range of views. For example, the properties straight vs. curved, symmetrical vs. asymmetrical, parallel vs. nonparallel are regarded as nonaccidental properties (nonaccidental properties were originally proposed within an image-based approach by Lowe, 1985, 1987). According to the model, geons and their spatial configuration are combined to a structural representation, called geon structural description. The spatial relations between parts are described in a categorical way, using relations like *above*, *below*, etc. Like other structural description models, Biederman's model predicts invariance in relation to position and size and also in relation to orientation in depth, as long as no parts are occluded. The model, however, does predict that recognition performance depends on the orientation of the stimulus in the picture plane, because the relations between parts are defined in a viewer-centred frame (Hummel & Biederman, 1992).

[3] Experiments conducted by Edelman (Edelman, 1995; Cutzu & Edelman, 1996) are an exception, because rather natural stimuli were used. However, Edelman's Chorus model is not a usual feature model (see 2.4.2).

The question has to be raised whether objects can be decomposed into geons at all. It was argued that Biederman's RBC cannot be applied to a whole range of biological stimuli (Ullman, 1996, p. 30), or that biological shapes in general cannot be adequately described by structural description models (Kurbat, 1994; Leyton, 1992, p. 411-413). In accordance with these arguments, it is far from clear how the biological objects from the CogVis morph database can be described by the existing set of geons. This problem extends also to artefact categories like *shoe*, *hat* or *backpack*, which seem to exceed the scope of the geon model. Therefore it has to be doubted seriously that object parts are necessarily represented as geons, or as similar geometrical primitives (for further problems of RBC see Edelman, 1999; Tarr & Bülthoff, 1995, 1998; see also chapter 3). However, this does not mean that category representations do not have a part structure: Actually there is evidence that object parts have an important role in recognition and categorisation (e.g., Biederman, 1987; Biederman & Cooper, 1991; Tversky & Hemenway, 1984). It should be noted that it is not the notion of part structure in object representations by itself which is problematic, but the use of parts and relations as a basis to derive invariant recognition performance (see Graf & Schneider, 2001).

Biederman's RBC is already in some way a model of *categorisation*, because it was designed to account for entry level recognition (which usually corresponds to basic level categorisation), and it was claimed that it can be extended to subordinate level recognition (Biederman et al., 1999). In the last decade, categorisation models that are similar to structural description models were developed also in the traditional categorisation community. One example is Barsalou's (Barsalou, 1992; Barsalou & Hale, 1993) frame model, in which categories are conceptualised as frames or schemata. But this model is not elaborated to account for the categorisation of familiar objects on the basis of their shapes, because category representations are conceptualised in terms of relatively abstract properties, similar to traditional feature approaches.

1.4 Image-based (view-dependent) models

In the last decade more and more studies accumulated which demonstrated that recognition is not view-independent. Orientation effects were found for novel objects (e.g., Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992; Tarr & Pinker, 1989), and also for common, familiar objects (e.g., Hayward & Tarr, 1997; Lawson & Humphreys, 1996, 1998; Murray, 1997, 1999; Newell & Findlay, 1997; Palmer, Rosch & Chase, 1981). Orientation-dependent performance could be verified with naming tasks, sequential matching tasks and priming tasks (for reviews see Jolicoeur & Humphrey, 1998; Lawson, 1999), with a visual search task (Jolicoeur, 1992) and even with figure-ground tasks (Gibson & Peterson, 1994). Orientation-dependent recognition performance is not limited to individual objects, like faces (e.g., Hill, Schyns & Akamatsu, 1997), or to objects on the subordinate level of categorisation (e.g., Edelman & Bülthoff, 1992; Tarr, 1995), but also was demonstrated for basic level recognition (Hayward & Williams, 2000; Jolicoeur et al., 1998; Lawson & Humphreys, 1998; Murray, 1998; Palmer, Rosch & Chase, 1981).

Moreover, recognition performance is not only influenced by the orientation, but also by the *size* of the stimulus. Results are quite similar: Reaction times (RTs) and error

rates depend on the extent of transformation that is necessary to align memory and stimulus representation. RTs increase in a monotonic way with increasing change of (perceived) size (e.g., Bundesen & Larsen, 1975; Bundesen, Larsen & Farrell, 1981; Cave & Kosslyn, 1989; Jolicoeur, 1987; Larsen & Bundesen, 1978; Milliken & Jolicoeur, 1992; for a review see Ashbridge & Perrett, 1998). Several studies even showed a systematic relation between the amount of *translation* and recognition performance: Increasing displacement between two sequentially presented stimuli led to a deterioration of performance, both for novel objects (Dill & Edelman, 2001; Dill & Fahle, 1998; Foster & Kahn, 1985) and familiar objects (Cave et al., 1994). Overall, view-independent models are difficult to reconcile with these findings which indicate that recognition performance depends systematically on different spatial transformations.

A number of image-based models were developed in the recognition literature in order to account for the systematic dependency on spatial transformations. Image-based recognition models were originally designed to explain the recognition of individual or identical objects. Meanwhile, extensions to category level recognition were proposed for all of the major image-based approaches to recognition. [4]

1.4.1 Alignment models

The basic idea of the alignment approach is that the differences between the visual image and the stored image-like object representation, which are caused by differences in position, size, and orientation, are compensated by an alignment of memory representation and stimulus representation, so that both can be compared in a straightforward way in a subsequent matching. If the alignment process is assumed to be time-consuming and error-prone, orientation-dependency of recognition can be accounted for easily (e.g., Jolicoeur, 1985; 1990; Tarr & Pinker, 1989). There is psychophysical evidence (Bundesen, Larsen & Farrell, 1981; Kourtzi & Shiffrar, 2001) that analogue compensation processes may be involved in object recognition, i.e. that compensation processes are continuous or incremental and traverse intermediate stages of the transformational path. This is confirmed by neurophysiological evidence which indicates that the orientation of objects is continuously mapped in the visual cortex (Wang, Tanifuji & Tanaka, 1998).

Several computational alignment models were developed, of which Ullman's (1989, 1996) 3D alignment model and Lowe's (1985, 1987) SCERPO model are probably the best known examples. Alignment models were originally designed for the identification of individual or identical objects. However, in the last years alignment models were extended to account for *categorisation*. In one attempt, entry level classification is explained by specific linear transformations and a subsequent tolerant matching (Yolles, in Ullman, 1996, p. 172-178). In this model the alignment is limited to

[4] Even though image-based models play a prominent role in the recognition literature, up to now no detailed image-based models were suggested in the *categorisation* literature. Barsalou's (1999) framework of perceptual symbol systems is in some way close to an image-based model, because it suggests that spatial transformations may play a role in categorisation. However, Barsalou's framework is not a model of perception and categorisation, as it does not explain how the fit between one representation and another is computed.

linear transformations. Shape differences between members of the same category are not accounted for by the alignment itself, but by a tolerant matching which occurs as a second step after normalization.

As an alternative to Ullman's (1989) model that relies on 3D object representations, Ullman and Basri (1991) suggested an alignment model on the basis of 2D views. In this model, an internal object model is constructed by a linear combination of a small number of stored 2D images. Thus, the alignment is not achieved by a spatial compensation process, but by linear combination of images. The intuition behind the linear combination approach can be explained in simple terms: Suppose that two views of the same three-dimensional object are stored, taken from somewhat different viewing directions. An intermediate view can then be described as a weighted sum of the views that are already stored.

Also for the linear combination approach an extension to class level recognition was proposed (Basri, 1996). Similar to Yolles' model, the alignment is limited to affine transformations in such a way that view and shape variations are decoupled: While view variations are compensated by the alignment process, shape variations within different members of the same category are not compensated for by the alignment process itself, but are accounted for by a tolerant similarity measure.

1.4.2 Interpolation, pooling and threshold models

In the interpolation model, recognition is achieved by localization in a multidimensional representational space, which is spanned by stored views (Poggio, 1990; Poggio & Edelman, 1990; Poggio & Girosi, 1990). The interpolation model is based on the theory of approximation of multivariate functions and can be implemented with radial basis functions (RBFs). In this scheme, the whole viewing space of an object is approximated by the learned views through a series of RBFs, which spread out the views in a high-dimensional feature space. Object recognition then means to examine whether a new point can be approximated by the existing tuned basis functions. Thus, recognition does not occur by transformation or reconstruction of an internal image, but by interpolation or approximation in a high-dimensional representational space.

Also *categorisation* models were developed within the interpolation approach (e.g., Edelman, 1998, 1999; Riesenhuber & Poggio, 2000). A well-developed extension of an image-based model to entry level categorisation is the Chorus model (e.g., Edelman, 1998, 1999; Edelman & Duvdevani-Bar, 1997; Edelman, 1995). In order to extend the model to categorisation, the RBF-framework is interpreted as representing not just single views, but prototypical shapes. Objects are represented by their similarity to several of the stored prototypes (chorus of prototypes). Thus, the similarity between shapes is represented in the Chorus model, but not the geometry of the shapes per se. Edelman distinguishes a multidimensional distal shape space and a proximal representational space. To represent shape differences between a perceived object and the prototypes in the proximal space, a common parameterisation of the distal shape space is necessary – at least within object categories. In this common parameterisation nonrigid transformations of shape (morphings) are defined (Edelman, 1998; Edelman &

Duvdevani-Bar, 1997). Categorisation is not achieved in the Chorus model by transformations of pictorial representations, but by assigning a location in the proximal space to the stimulus, according to the similarity to a number of prototypical templates.

Edelman's model is an important step towards an integrative model of recognition and categorisation, but has some limitations (see commentaries in BBS to Edelman, 1998; see also Riesenhuber & Poggio, 2000). For example, the assumption of a multidimensional feature space raises a number of difficulties (see e.g., Marko, 1973; Tversky, 1977), and the assumption of a linear representational space was criticized (Bonmassar & Schwartz, 1998; Gregson, 1998; van Leeuwen, 1998). Moreover, the holistic nature of representations in Chorus is problematic (e.g., Hummel, 2000; but see Edelman & Intrator, 2000, 2001).

At the end of the 90s, pooling and threshold models of recognition were developed (Perrett & Oram, 1998; Perrett, Oram & Ashbridge, 1998; Riesenhuber & Poggio, 1999, 2002; Wallis & Bülthoff, 1999). Recognition is explained on the basis of the behaviour of cells in IT cortex which are selectively tuned to specific image features (fragments or whole shapes) in a view-dependent (and size-dependent) way. A hierarchical pooling of the outputs of view-specific cells provides generalization over viewing conditions (Perrett & Oram, 1998). A similar proposal was made by Riesenhuber and Poggio (1999, 2002), reminiscent of the Pandemonium model (Selfridge and Neisser, 1963; see Tarr, 1999). The threshold model (Perrett, Oram & Ashbridge, 1998) also accounts for the systematic relation between recognition latencies and the amount of rotation (and size-scaling): The speed of object recognition depends on the rate of accumulation of activity from neurons selective for the object, evoked by a particular viewing circumstance. For a familiar object, more tuned cells will be activated in the views most frequently presented, so that a given level of evidence (threshold) can be achieved fast. When the object is seen in an unusual view, fewer cells will respond, and activity among the population of cells selective for the object's appearance will accumulate more slowly. Consequently, these threshold models explain orientation-dependency without the need to postulate transformation or interpolation processes.

How can pooling/threshold models account for basic level *categorisation*? In the model of Perrett et al. (1998), recognition depends only on how well the input image falls within the tolerance of neural representations of familiar objects. The speed of classification of an unfamiliar exemplar of a familiar object class (e.g., recognizing a new car model as a car) should depend only on the novel item's similarity to familiar exemplars. Also Riesenhuber and Poggio (2000) extended their model to class level recognition, introducing one layer of units (RBFs) which cover the stimulus space and are tuned in unsupervised training, and above another task-specific layer of units which are tuned by supervised training.

Interpolation, pooling and threshold models are interesting, because they integrate neurophysiological and psychological modelling, but they bring up several problems: First, the models are difficult to reconcile with psychophysical evidence for position-dependency in object recognition (Cave et al., 1994; Dill & Edelman, 2001; Dill & Fahle, 1998; Foster & Kahn, 1985). Second, it seems difficult to reconcile these models with evidence for analogue compensation processes in recognition (e.g., Bundesen et al., 1981; Kourtzi & Shiffrar, 2001). Third, these approaches rely on the notion of shape similarity,

without being able to account for the similarity of shapes. Fourth, these models are purely bottom-up models. Present models are not compatible with evidence for massive top-down processing in the cortex (e.g., Ullman, 1995). And fifth, the question how object structure can be dealt within these approaches needs further elaboration (see Edelman & Intrator, 2001; see also Perrett & Oram, 1998).

1.5 Recent developments

In the last years, several new developments can be traced in the literature on object recognition. We think that two of these developments are of special importance in the present context, as they provide possible extensions of the image-based approach to recognition and categorisation.

Recently, attempts were made to integrate object parts or fragments into image-based accounts of recognition and categorisation. Edelman responded to the criticism against the holistic nature of his Chorus model by postulating a Chorus of Fragments model which encodes object fragments and implicit spatial relations between them (Edelman & Intrator, 2000, 2001). He proposed part detectors (what + where neurons) which are coarsely tuned to the shapes of specific object fragments and a range of locations. Within this framework, configurations of parts can be represented, and illustrated by a pegboard whose spatial structure supports the arrangement of parts, which are suspended by pegs (see also Perrett & Oram, 1998).
Another image-based model which integrates the part-structure of object representations was proposed by Basri, Costa, Geiger and Jacobs (1998). The authors showed that similarities of part structure can be captured with an elastic matching approach. The underlying idea is that corresponding parts can be aligned when they are represented in an elastic representation. Moreover, Ullman et al. (2002) proposed a model which extracts image features of intermediate complexity which may be involved in object categorisation. Thus, in general, there is a tendency in the literature to integrate structured representations in image-based models of recognition and categorisation.

A second new development is related to the issue of shape transformations in basic level categorisation. One of the central questions in object categorisation is how the shape variability of different category members can be accounted for. Recently Graf (2001, 2002) proposed that the shape variability within categories up to the basic level of categorisation can be described by continuous transformations of object shape (topological transformations), which can be illustrated by locally deforming a rubber sheet on which the shape of an object was printed (see Figure 1). Topological transformations seem not only suited to account for shape variations within biological categories (Thompson, 1917/1942), but also for many artefact categories (Graf, 2002).

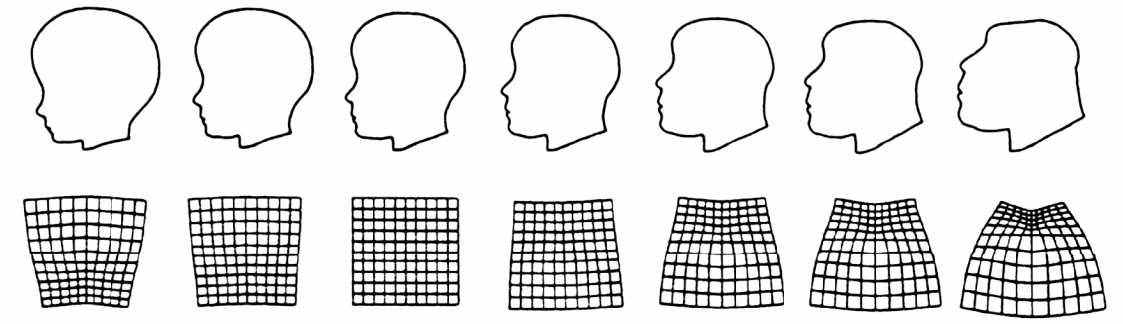

Figure 1. A series of different heads can be created by a topological transformation. The transformation is illustrated by the deformation of the corresponding coordinate system (Shaw & Pittenger, 1977).

A speeded categorisation experiment with line drawings demonstrated that categorisation performance depended systematically on the amount of shape transformation. Moreover, transformation times were sequentially additive, which suggests that analogue deforming transformation processes are involved in basic level categorisation, i.e. transformations which pass through intermediate positions in the transformational path. In order to account for these findings, the alignment approach was extended to nonlinear shape transformations (Graf, 2001, 2002). The basic idea is that basic level categorisation can be conceptualised simply by allowing for topological (warping or morphing) transformations which compensate for shape differences within category members. The model can be conceptualised on the basis of 2.5-D representations. This approach allows for an integrative model of recognition and categorisation up to the basic level, simply by using different transformations in recognition and categorisation. However, it still has to be investigated whether these results generalize to perceived similarity, to the categorisation of more realistic grey-level stimuli, and to objects that underwent both shape changes and rotations.

In summary, these two new developments provide interesting extensions which can overcome limitations of the existing image-based models of recognition, and therefore are the starting point for our own research within the CogVis project.

2 Psychophysical experiments

As already described in chapter 1.4, recognition performance depends systematically on spatial transformations, like rotations and size-scalings. These findings cannot be accounted for by models which predict that recognition performance is invariant regarding spatial transformations (view-invariant models). Consequently, image-based models were developed in order to account for view-dependency in object recognition. However, present image-based models still have a number of limitations (see introduction). One main issue is related to the question what role object parts play in object recognition and categorisation, and how structured representations can be integrated into an image-based model. A second major issue refers to the question how object categorisation, especially basic level categorisation, can be accounted for within an image-based approach. Furthermore, more research is required on the questions what role top-down processing, and dynamic aspects – like object motion – play in categorisation. Consequently, the following research questions arise:

- Do parts and structured representations play an important role in object recognition and categorisation? How can structural representations be integrated in an image-based model?
- Given that similarity is an essential determinant of categorisation: How are systematic changes of object shape reflected in perceived similarity? To put it more general: How can basic level object categorisation be accounted for within an image-based model?
- What is the role of context and top-down information in object categorisation?
- What is the interplay between object motion and shape for categorisation decisions?

We have addressed these topics in a number of psychophysical experiments. The next section contains an overview of the studies and main results, followed by separate research reports for every study (chapters 2.1 – 2.5).

The main approaches to explain human recognition and categorisation differ quite remarkably with regard to their assumptions on the structural aspects of recognition memory. Invariant property models, traditional feature models and structural description theories assume that *local featural or part-based information* plays a pivotal role in object recognition and categorisation, and assume the recognition and categorisation performance is essentially *viewpoint-invariant.* In contrast, many *view-based* models are *holistic*, i.e. they propose that objects are encoded and represented as unparsed perceptual wholes in which parts and spatial relations are not explicitly represented. In order to explore structural aspects of recognition memory we have developed a new psychophysical method for investigating the role of featural and spatial-relational information in recognition and categorisation (*study 1*). In the first experiments we have applied this method to old-new recognition of faces because this stimulus class has often been considered to be processed in an exclusively holistic way (e.g., Biederman & Kalocsai, 1997; Tanaka, & Farah, 1993, Farah, Tanaka, & Drain, 1995). Our results provided clear evidence that spatial-relational as well as featural information is important even for face recognition. This suggests that part information is even more important for

object recognition and categorisation. By comparing the recognition of unfamiliar and familiar faces we have addressed the question how memory representations develop over the time course of learning. We showed that there is no qualitative shift in terms of using featural vs. spatial-relational information, even though familiar faces were better recognised than unfamiliar ones.

A further central issue is how categorisation can be conceptualised within an image-based model. This question is essential, as we typically recognise objects at the basic level of categorisation – and thus categorise objects when we see them. As similarity has a central status in virtually all models of categorisation, we investigated object similarity in the context of categorisation (*study 2*). Using line-drawings, we found that object similarity is related in a systematic way to continuous transformations of object shape (topological transformations), which were produced by morphing between two images of the same basic level category. Thus, topological transformations seem well suited to account for the shape similarity of members of familiar categories on the basic level of categorisation. These findings are in accordance with the proposal that similarity judgements involve the alignment of corresponding parts (e.g., Markman, 2001) – as the morphing procedure relies on an alignment of corresponding features or parts.

The role of object shape in categorisation was further explored with speeded categorisation tasks in *study 3*. Previous work has shown that the amount of shape transformation was systematically related to categorisation performance for line drawings (Graf, 2002). We investigated whether these results generalize to more realistic grey-level images rendered from 3D object models (see CogVis object data base). We also studied the effects of combined shape transformations and image-plane rotations on categorisation performance. The results confirm that categorisation performance is systematically related to the amount of shape transformation, both for line drawings and grey-level images, as well as for upright and plane rotated objects. In addition, orientation dependency was corroborated with a basic level categorisation task. Finally, categorisation processes which compensate for shape changes and plane rotations seem to be independent, confirming previous evidence of independent effects for other combinations of spatial transformations (e.g., Lawson et al., 2000). Overall, the results support an image-based model of basic level categorisation, as continuous shape changes are systematically related to categorisation performance.

In *study 4* the role of context in recognition and categorisation was studied. Many recognition schemes assume a purely bottom-up processing. This assumption is questionable when one considers the fact that almost every brain area used in visual cognition sends feedback to previous areas. Indeed, using an associative priming paradigm we demonstrated that looking at an object (e.g., coffee spoon) can facilitate the recognition of noncanonical views of another object which is related in associative memory (e.g., coffee cup). This result challenges most current recognition models that are purely bottom-up and suggests an important role of context in recognizing everyday objects.

Finally, we examined the interplay between object motion and shape for categorisation decisions and perception (*study 5*). Previous studies on categorisation often have focused either purely on the static or on the dynamic domain (e.g., point-light walkers). In a first study we have taken a further step and investigated the relevance of static *and* dynamic cues in a categorisation task. Interestingly, we found that subjects

could readily use both types of cues and that there was *no* significant advantage of one cue type over the other. These results argue strongly for a cue-integrating processing strategy for categorisation, in which shape and rigid motion information is combined. In addition to rigid motion also non-rigid motion could play an important role for certain stimulus classes like living things or faces. Indeed, one of the most relevant types of information processing in everyday life is that of emotional expression. Interestingly, inverting the eyes and mouth in an upright faces creates a very bizarre facial expression which disappears when the face is turned upside down (Thatcher Illusion, Thompson, 1980). In the last study we investigated the interplay between nonrigid motion and shape for perception using "thatcherized" faces.

In summary, we investigated different aspects of visual object recognition and categorisation. The findings both support and extend an image-based model of recognition and categorisation.

2.1 Study 1: Structural aspects of recognition memory: a new method for measuring the role of featural and relational information

As explained in the introduction, object recognition and categorisation theories differ with regard to their assumptions on the representations used. On one hand, invariant property approaches, traditional feature models and structural description theories assume that local features or parts play an important role for recognition and categorisation. On the other hand, view-based schemes have often been associated with holistic processing in which objects are processed as wholes without explicitly encoding and representing parts.

Faces are one of the most relevant stimulus classes in everyday life. Moreover, they have been claimed by several authors to be the example for exclusive holistic processing (e.g., Biederman & Kalocsai, 1997; Farah, Tanaka, & Drain, 1995; Tanaka & Farah, 1993). These two properties make faces a very interesting stimulus class to examine the role of featural and relational information in recognition and categorisation.

In computer vision many face recognition algorithms process the whole face without explicitly processing facial parts. Some of these algorithms have been thought of being particularly useful to understand human face recognition and were cited in studies that claimed faces to be the example for exclusive holistic processing (e.g., Lades, Vorbrüggen, Buhmann, Lange, Malsburg, Würtz, & Konen, 1993 and Wiskott, Fellous, Krüger, & von der Malsburg, 1997 cited in Biederman & Kalocsai, 1997, or the computation models cited in Farah et al., 1995, p. 496). In contrast to these and other holistic algorithms like principal components analysis or vector quantization, recent computer vision approaches have started using local part-based or fragment-based information in faces (Heisele, Serre, Pontil, Vetter, & Poggio, 2001; Lee & Seung, 1999; Ullman & Sali, 2000). Since human observers can readily tell the parts of a face such algorithms bear a certain intuitive appeal. Moreover, potential advantages of such approaches are greater robustness against partial occlusion and less susceptibility to viewpoint changes.

In the present study we used psychophysics to investigate whether human observers only process faces holistically, or whether they encode and store the local

information in facial parts (featural information) as well as their spatial relationship (configural information). In contrast to previous studies, we developed a method that did not alter configural or featural information, but eliminated either the one or the other. Previous studies have often attempted to directly alter the facial features or their spatial positions. However, the effects of such manipulations are not always perfectly selective. For example, altering featural information by replacing the eyes and mouth with the ones from another face could also change their spatial relations (configural information) as mentioned in Rhodes, Brake, and Atkinson (1993). Rakover has pointed out that altering configuration by increasing the inter-eye distance could also induce a part-change, because the bridge of the nose might appear wider (Rakover, 2002). Such problems were avoided in our study by using scrambling and blurring procedures that allowed investigating the role of featural and configural information separately. The current study extends previous research using these manipulations (e.g., Collishaw & Hole, 2000; Davidoff & Donnelly, 1990; Sergent, 1985) by ensuring that each procedure does effectively eliminate configural or featural processing.

Experiment 1

The first experiment was designed to investigate whether human observers store featural information *independent* of configural information. In the first condition configural information was eliminated by cutting the faces into their constituent parts and scrambling them. If the local information in parts (featural information) is encoded and stored, it should be possible to recognise previously learnt intact faces even if they are scrambled. In condition 2 the role of configural information was investigated. Previously learnt faces had to be recognised when they were shown as greyscale low-pass filtered versions. These image manipulations destroyed featural information while leaving the configural information intact. In a control condition we confirmed that performance is reduced to chance when faces are low-pass filtered and scrambled, thus showing that our image manipulations eliminate featural and configural information respectively and effectively. Examples of the stimuli are shown in Figure 2.

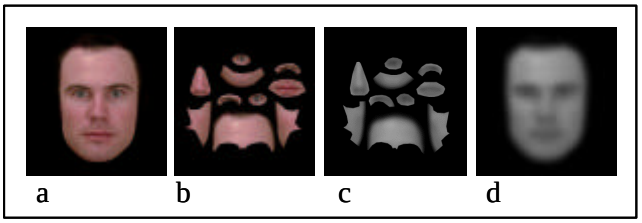

Figure 2. Sample Stimuli. a) intact face, b) scrambled, c) scrambled-blurred, d) blurred face. (From Schwaninger, Lobmaier, & Collishaw, 2002)

Thirty-six participants were randomly assigned to the different experimental conditions. Recognition performance was calculated using signal detection theory (Green & Swets, 1966). Face recognition performance was measured by calculating d' using an old-new judgement task (McMillan & Creelman, 1992).[5] d' was calculated for each participant and averaged across each group (Figure 3, black bars).

[5] This measure is calculated by the formula d' = z(H) – z(FA), whereas H denotes the proportion of hits and FA the proportion of false alarms. A hit was scored when the target button was pressed for a previously learned face (target) and a false alarm was scored when the target button was pressed for a new face

Scrambled faces were recognised above chance, suggesting that local part-based information has been encoded in the learning phase. These findings are contradictory to the view that faces are only processed holistically (Biederman & Kalocsai, 1997; Farah et al., 1995; Tanaka & Farah, 1991; Tanaka & Farah, 1993). The recognition of blurred faces was also above chance, confirming an important role of configural information for face recognition (Rhodes et al., 1993; Sergent, 1985). The control condition revealed that the blur filter eliminated all featural information since recognition was at chance when faces were blurred and scrambled.

Our results suggest that unfamiliar face recognition in humans entails separate representations for featural information and for configural information. The aim of Experiment 2 was to investigate changes of recognition memory due to familiarity. More specifically, we were interested whether there is a quantitative or a qualitative shift in processing strategy when faces have become familiar.

Experiment 2

Thirty-six participants participated in this experiment. The materials and procedure were the same as in Experiment 1, but all the targets were faces of fellow students and thus familiar to the participants.

The results of Experiment 2 replicated the clear effects from Experiment 1 and suggest an important role of local part-based and configural information in both unfamiliar and familiar face recognition. By comparing recognition performance from both experiments (Figure 3) we addressed the question to what extent familiar and unfamiliar face recognition differ quantitatively (e.g., generally a better performance when faces are familiar) or qualitatively (e.g., better performance for familiar faces using more accurate configural processing). To this end, a two-way analysis of variance (ANOVA) was carried out with the data from the scrambled and blurred conditions of Experiments 1 and 2 with familiarity (familiar vs. unfamiliar) and condition (scrambled vs. blurred) as between-subjects factors. There was a main effect of familiarity, $F(1,42) = 12.80$, $p < .01$, suggesting that familiar faces are more reliably recognised than unfamiliar faces (quantitative difference). There was also a main effect of condition, $F(1,42) = 6.7$, $p < .05$, indicating that blurred faces were better recognised than scrambled faces. The relative impact of blurring and scrambling did not differ between the two experiments, since there was no interaction between condition and familiarity, $F(1,42) = 1.02$, $p = 0.32$. This result suggests that

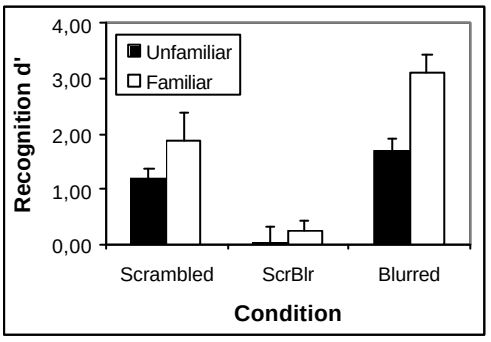

Figure 3. Recognition performance in unfamiliar and familiar face recognition across the three different conditions at test. ScrBlr: scrambled and blurred faces. Error bars indicate standard errors of the mean. (From Schwaninger, Lobmaier, & Collishaw, 2002)

(distractor). In the formula z denotes the z-transformation, i.e. H and FA are converted into z-scores (standard-deviation units).

there are no qualitative differences between familiar and unfamiliar face recognition on the basis of configural and featural information. In both cases both types of information are of similar importance.

Conclusion

In this study we have investigated the role of local part-based information and their spatial interrelationship (configural information). We used faces for two reasons: First, they are one of the most relevant stimuli in everyday life. Second, they have often been cited as the example for exclusive holistic processing (Biederman & Kalocsai, 1997; Farah et al., 1995; Tanaka & Farah, 1993).

The results of our experiments provided clear evidence for the view that human observers process familiar and unfamiliar faces by encoding and storing configural as well as local information of facial parts. Moreover, when faces are familiar both featural and configural processing becomes more accurate. Interestingly, there is no qualitative change, i.e. the relative balance between the two types of processing remains the same.

These results challenge the assumption that faces are processed only holistically and suggest a greater biological plausibility for recent machine vision approaches in which local features and parts play a pivotal role (e.g., Heisele et al., 2001; Lee & Seung, 1999; Ullman & Sali; 2000).

Neurophysiological evidence supports part-based as well as configural and holisitic processing assumptions. In general, cells responsive to facial identity are found in inferior temporal cortex while selectivity to facial expressions, viewing angle and gaze direction can be found in the superior temporal sulcus (Hasselmo, Rolls, & Baylis, 1989; Perret, Hietanen, Oram, & Benson, 1992). For some neurons, selectivity for particular features of the head and face, e.g., the eyes and mouth, has been revealed (Perret et al., 1992; Perret, Mistlin, & Chitty, 1987; Perret, Rolls, & Caan, 1982). Other groups of cells need the simultaneous presentation of multiple parts of a face and are therefore consistent with a more holistic type of processing (Perret & Oram, 1993; Wachsmuth, Oram, & Perret, 1994). Finally, Yamane, Kaji, and Kawano (1988) have discovered neurons that detect combinations of distances between facial parts, such as the eyes, mouth, eyebrows, and hair, which suggest sensitivity for the spatial relations between facial parts (configural information).

In order to integrate the above mentioned findings from psychophysics, neurophysiology and computer vision we propose the framework depicted in Figure 4. Faces are first represented by a metric representation in primary visual areas corresponding to the perception of the pictorial aspects of a face. Further processing entails extracting local part-based information and spatial relations between them in order to activate featural and configural representations in higher visual areas of the ventral stream, i.e. face selective areas in temporal cortex[6]. In a recent study, repetition priming was used in order to investigate whether the outputs of featural and configural representations converge to the same face identification units (Schwaninger, Lobmaier, &

[6] Although a role of the dorsal system in encoding of metric spatial relations has been proposed for object recognition it remains to be investigated, whether it does play a role for the processing of configural information in faces.

Collishaw, 2002). Since priming was found from scrambled to blurred faces and vice versa we propose that the outputs of featural and configural representations converge to the same face identification units.

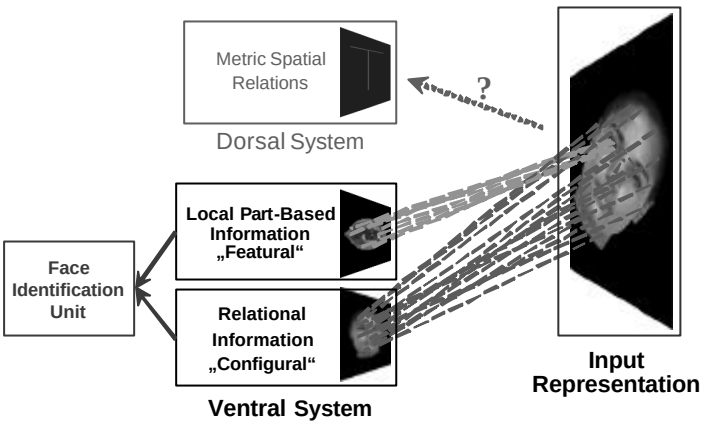

Figure 4. Integrative model for unfamiliar and familiar face recognition. (From Schwaninger, Lobmaier, & Collishaw, 2002)

Although first used for faces, the method described above provides a general tool for separately investigating featural and spatial relational processing. It would be interesting to investigate whether features, parts and configuration play a different role for recognizing objects on superordinate, basic or subordinate levels. Furthermore, it would be worthwhile to explore whether there are qualitative differences dependent on the task or context.

2.2 Study 2: Object similarity in the context of categorisation

Similarity has a central status in categorisation: Most of the otherwise distinct theories of categorisation share the assumption that the likelihood of assigning an object to a category depends on the similarity of the object to the category representation (e.g., McClelland & Rumelhart, 1985; Medin & Schaffer, 1978; Nosofsky, 1986; Reed, 1972; Rosch & Mervis, 1975; Smith & Medin, 1981; Smith, Shoben & Rips, 1974; for a review see Hahn & Chater, 1997). Moreover, objects that are visually similar to each other are more likely perceived categorically (Newell & Bülthoff, 2002). But how can similarity be accounted for? In the last decade a host of evidence has accumulated which suggests a structural alignment model of similarity and categorisation (e.g., Gentner & Markman, 1994, 1995; Goldstone, 1994a, 1994b, 1996; Goldstone & Medin, 1994; Markman & Gentner, 1993a, 1993b, 1997; Medin, Goldstone & Gentner, 1993; for a review see Markman, 2001): The basic idea is that similarity needs to be understood as a process, which brings aspects of the entities that are to be compared into correspondence. This *alignment process* is hypothesized to be a dynamic process, which supplies constraints for the similarity comparison, because it determines corresponding features or parts – and thus defines which aspects will be compared in the similarity matching.

It is still unclear, however, how the *shape* of common and familiar objects can be described on the basis of – usually more or less abstract – properties or features, which underlie most present approaches of similarity in the categorisation literature, and also the structural alignment approach. These rather abstract descriptions seem not adequate to capture object shape. Moreover, the existing recognition models in the recognition

literature either do not deal with class level recognition and the similarity of different (deformed) shapes (e.g., Poggio & Edelman, 1990; Poggio & Girosi, 1990), do not account for the similarity of deformed objects (e.g., Biederman, 1987; Hummel & Biederman, 1992), or just presuppose similarity – without explaining it (e.g., Edelman, 1998; Perrett et al., 1998). Therefore, it is still an open question how shape similarity can be conceptualised. This shortcoming is critical, because object shape is a central determinant in categorisation at the basic and subordinate level (Rosch et al., 1976). For these reasons, an account of shape similarity seems to be essential to understand basic level and subordinate level categorisation.

Most previous studies on shape similarity that investigated systematic changes of object shape used unfamiliar shapes. Several studies employed novel blob-like shapes with closed contours, which were produced on the basis of radial frequency components (Fourier descriptors) (e.g., Cortese & Dyre, 1996; Op de Beeck, Wagemans and Vogels, 2001; Shepard & Cermak, 1973). Psychophysical studies found an ordinal agreement between parametric configurations and their perceptual representation (as determined by multidimensional scaling), i.e. detected the same number of dimensions and the same stimulus order (Cortese & Dyre, 1996; Shepard & Cermak, 1973). In a recent study, this finding was confirmed both at the behavioural and neuronal level (Op de Beeck et al., 2001). Only few studies were performed in which more common and natural stimuli were used: Cutzu and Edelman (1998, 1996) investigated the similarity of parametrically deformed animal-like computer-generated objects. Their results showed that planar configurations in a low-dimensional shape space were recovered by MDS from proximity tables derived from the subject data. This suggests that deforming shape transformations are systematically related to human performance. Overall, these studies indicate that shape similarity decreases monotonically with increasing amount of elastic deformation. However, it is not clear whether these results transfer to the more complex shapes of common and familiar objects, over a large set of categories.

How can shape similarity within common and familiar basic level categories be explained? It was proposed that the shapes of different objects from the same basic level category can usually be aligned by rather simple deforming transformations which continuously transform object shape (so-called topological transformations) (Graf, 2001). By extending the image-based alignment approach (e.g., Ullman, 1989) to shape transformations, the similarity of common and familiar object shapes can be explained (see Figure 5). The following predictions can be derived from this transformational (or alignment) model of similarity: The similarity of objects should decrease in a systematic way with increasing amount of topological transformation, leading to a high negative correlation between transformational distance and perceived similarity. Moreover, the decrease in similarity is expected to be monotonic.

Experiment

In the experiment, subjects were asked to rate the similarity of objects from the same basic level category on a scale from *1* (very dissimilar) to *9* (very similar). Outline shapes (line drawings) of object pairs from 25 common and familiar categories (12 biological and 13 artefact categories) were presented in a booklet. The amount of topological transformation between objects was manipulated.

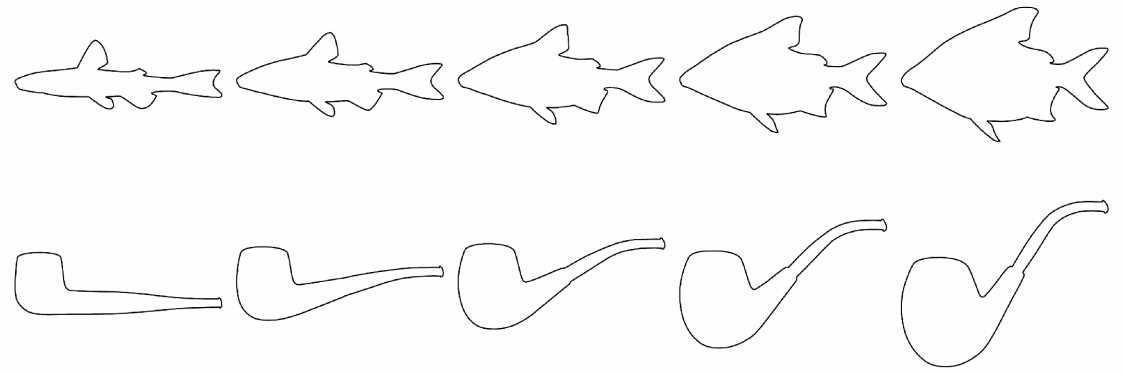

Figure 5. Members of a specific category are produced by continuously transforming the shape of existing members. This method works for biological categories and for many artefact categories, as exemplified by line drawings of objects from the categories *fish* and *pipe*. The underlying morphing (or warping) transformations are well suited to describe the shape variability within basic level categories.

Mean similarity ratings were computed, averaged over categories. The factor distance was highly significant (F (4,140) = 1848.75; $p < .001$). Mean similarity ratings decreased with increasing transformational distance, as can be seen in Figure 6. This is confirmed by a very high negative Spearman correlation between transformational distance and the similarity ratings (r_s = -.967, $p < .001$). The amount of topological transformation accounts for 93.5% of the variance in the similarity ratings. Also a highly significant linear trend was found in the data (F (1,35) = 5280.08; $p < .001$), which indicates a monotonic decrease.

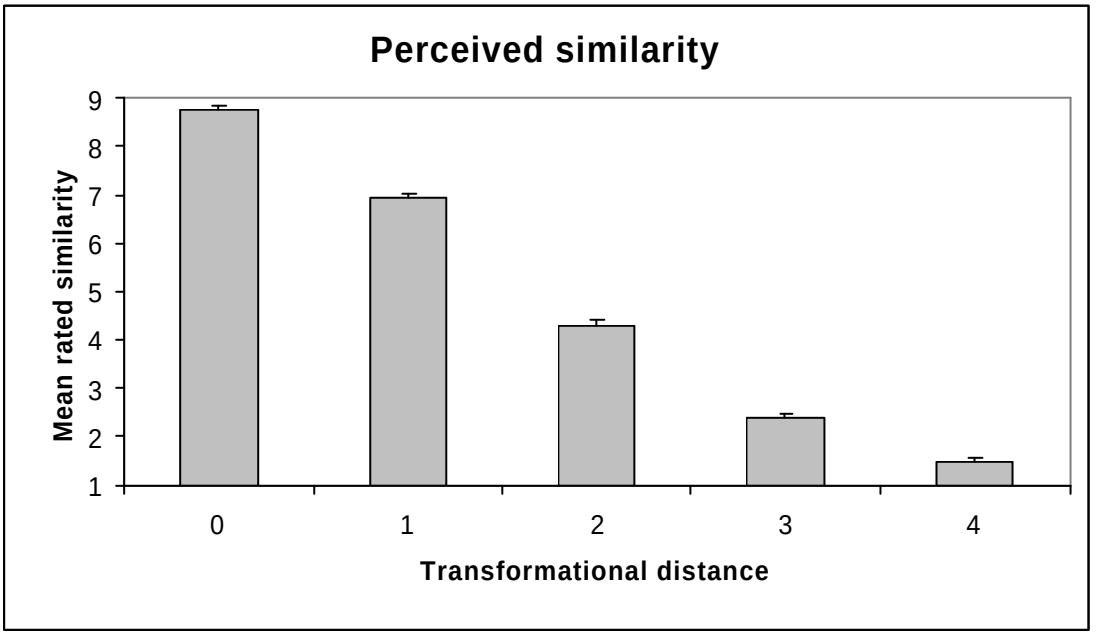

Figure 6. Mean similarity ratings, averaged over categories and subjects. Perceived similarity decreased with increasing amount of topological transformation between the two objects (as specified in the morphing procedure).

In a next step, it was tested whether this decrease in rated similarity can be found also for individual categories. For every category, an ANOVA with distance as factor was performed. For all 25 categories, the main effect distance was highly significant ($p < .001$). In all categories a linear trend proved to be highly significant ($p < .001$ in all categories), suggesting a monotonic decrease. In addition, high negative Spearman correlations between transformational distance and rated similarity were found for all categories, ranging from $r_s = -.850$ to $r_s = -.935$ (see Graf, 2002). Thus, perceived similarity decreased with increasing topological distance in each of the 25 categories.

Does the systematic relation actually result from topological transformations, or can alternative models account for it, without having to refer to topological transformations? First, can the effect be reduced to simple affine changes? In order to investigate this question, data analysis was repeated for those categories with little affine change (*dinosaur*, *glass*, *bell*, *head*, *starfish*). Again, similarity decreased with increasing transformational distance (F (4,140) = 1050.66; $p < .001$). A very high negative correlation between transformational distance and the similarity ratings was found ($r_s = -.951$, $p < .001$), and a linear trend was highly significant (F (1,35) = 3697.19; $p < .001$). Thus, the pattern of results is highly similar to that for all objects, indicating that the systematic relation also holds for categories with little affine change. Second, can the systematic relation also be detected for categories in which the spatial configuration of parts is highly similar? Also for these categories (*bell*, *bottle*, *light-bulb* and *turnip*) perceived similarity decreased with increasing transformational distance (F (4, 140) = 986.70; $p < .001$). This is confirmed by a very high negative Spearman correlation ($r_s = -.959$, $p < .001$); the linear trend was highly significant, as well (F (1,35) = 7949.64; $p < .001$). Therefore, the effect is not simply due to changes in the configuration of parts. To sum up, the systematic decrease of rated similarity with increasing amount of transformation cannot be reduced to affine transformations, or to the changes in the part configuration, but reflects the amount of topological transformation.

The increase of similarity with increasing transformational distance was found for both biological categories (F (4,140) = 1447.32; $p < .001$) and artefact categories (F (4,140) = 1552.78; $p < .001$). The Spearman correlations between transformational distance and rated similarity showed similar results: Very high negative correlations resulted both for biological objects ($r_s = -.965$, $p < .001$) and for artefact objects ($r_s = -.962$, $p < .001$). Also highly significant linear trends were detected, for biological categories (F (1,35) = 4510.74; $p < .001$) and artefact categories (F (1,35) = 5731.31; $p < .001$).

Conclusions

The results matched well with the predictions. A very high negative correlation between similarity and transformational distance was found, and the results clearly showed that perceived similarity decreased monotonically with increasing amount of shape transformation. The effect of topological distance could not be reduced to simple affine changes, or to changes in the configuration of parts. Consequently, it is actually the deforming (or space-curving) nature of the transformations that influences perceived similarity. The decrease of similarity with increasing transformational distance was found for both biological and artefact categories, which suggests commonalities in shape processing for biological and artefact categories on this fundamental (and

"prefunctional") level of visual similarity. Overall, the findings suggest that topological transformations are well suited to account for the shape similarity of basic level category members, both for biological and artefact categories.

The findings are in agreement with the proposal that similarity is determined by a dynamic alignment process (e.g., Medin et al., 1993). The proposed model may be regarded as an image-based extension of the structural alignment approach to similarity, and complements the structural alignment approach with a model that accounts for the similarity of shapes (see also Basri et al., 1998). Moreover, this account fits nicely with recent findings that suggest a transformational model of similarity (Hahn, Chater & Richardson, in press).

2.3 Study 3: Shape transformations and image-plane rotations in object categorisation

For quite some time there is evidence that category representations up to the basic level of categorisation are image-based (e.g., Rosch et al., 1976), but there is still relatively little research on the question how image-based categorisation is achieved in the human visual system (for exceptions see Edelman, 1998; Graf, 2002). One of the central questions in object categorisation is how the shape variability of different category members can be accounted for. Recently, Graf (2001, 2002) proposed that the shape variability within categories up to the basic level of categorisation can be described by topological transformations, i.e. by nonrigid continuous deformations which can be illustrated by locally deforming a rubber sheet on which the shape of an object was printed.

A database of basic level categories was created by constructing morphable 3D models (using 3D Studio Max Version 4.2) of exemplars from 29 different categories, covering both biological and artefact categories. New members of each category can be produced by morphing between two category exemplars. The objects can be morphed and also rotated in space; they can be rendered as grey-scale images (see Figure 7).[7]

The central question now is in what way categorisation performance is related to these shape transformations. As delineated in chapter 1.4, recognition performance (RTs and error rates) deteriorates with increasing amount of rotation, size-scaling or translation of the object. If performance for shape changes in categorisation was similar to that for orientation and size changes, then categorisation performance should deteriorate systematically with increasing topological distance. A systematic increase of categorisation latencies and error rates with increasing topological distance was found in a sequential matching experiment, using 2D outline shapes (line drawings) as stimuli (Graf, 2001; 2002). These findings provided first evidence that categorisation performance is systematically related to shape transformations. However, further research is needed: Can these findings be replicated with grey-scale images of familiar objects, using different morphing software? How is categorisation performance influenced by

[7] We thank Christoph Dahl for substantial help in generating object models and morphs.

other transformations, like rotations, especially when objects are both deformed and rotated?

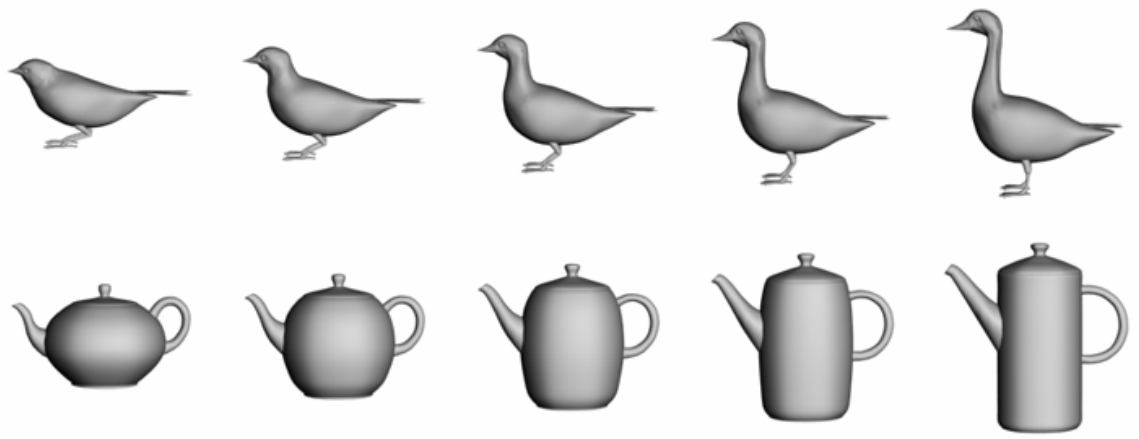

Figure 7. Shape transformation by morphing, using 3D objects rendered as grey-level images. New category members can be created by morphing between existing members from the same basic level category, both for biological and artefact categories, as exemplified by the categories *bird* and *pot*. Intermediate morphs are surprisingly good category members.

Experiment 1: Speeded categorisation task with morphed objects.
In a first experiment it was attempted to replicate the systematic relation between the amount of shape transformation and categorisation performance with more realistic grey-scale images. Objects from 29 basic level categories were employed, of which two were used only in the practice phase. The stimuli were rendered from 3D models (CogVis object database). These objects are highly realistic compared to the stimuli which are typically employed in the categorisation literature (e.g., Goldstone, 1996; Maddox et al., 1998; Nosofsky, 1986). Two objects were presented sequentially (backward masked), and subjects were required to decide whether both objects belonged to the same basic level category, or to a different category. The amount of topological transformation (morphing distance) between members of the same category was varied. 12 subjects participated in the experiment; every subject had to perform two sessions.

Reaction times increased systematically with increasing amount of shape transformation (F (3,33) = 59.44, p < .001) (see Figure 8). The increase showed a highly significant linear trend (F (1,11) = 69.82, p < .001). Thus, the relation between topological distance and categorisation latencies was replicated for the more realistic grey-level objects.

Even though RTs decreased with practice (F (3,33) = 16.26, p < .001), the systematic effect of shape transformation did not diminish with practice (interaction not significant: F (9,99) = 1.13, p = .35). A similar increase of RTs was found for those categories which undergo only little affine change (< 20%) (F (3,33) = 21.87, p < .001), or which have a highly similar part configuration (F (3,33) = 33.74, p < .001). Thus, the results are not simply due to affine changes, or to changes in the configuration of parts. The systematic effect was found for both artefact categories (F (3,33) = 43.28, p < .001)

and biological categories (F (3,33 = 37.46, $p < .001$). In contrast to the earlier study with 2D shapes (Graf, 2002), transformation effects were larger for artefact categories than for biological categories, especially for higher transformational distances (indicated by a significant interaction between transformational distance and category type: F (3,33) = 7.07, $p = .001$). This may result from larger shape changes for artefact categories than for biological categories, which were often accompanied by comparatively larger changes in shading.

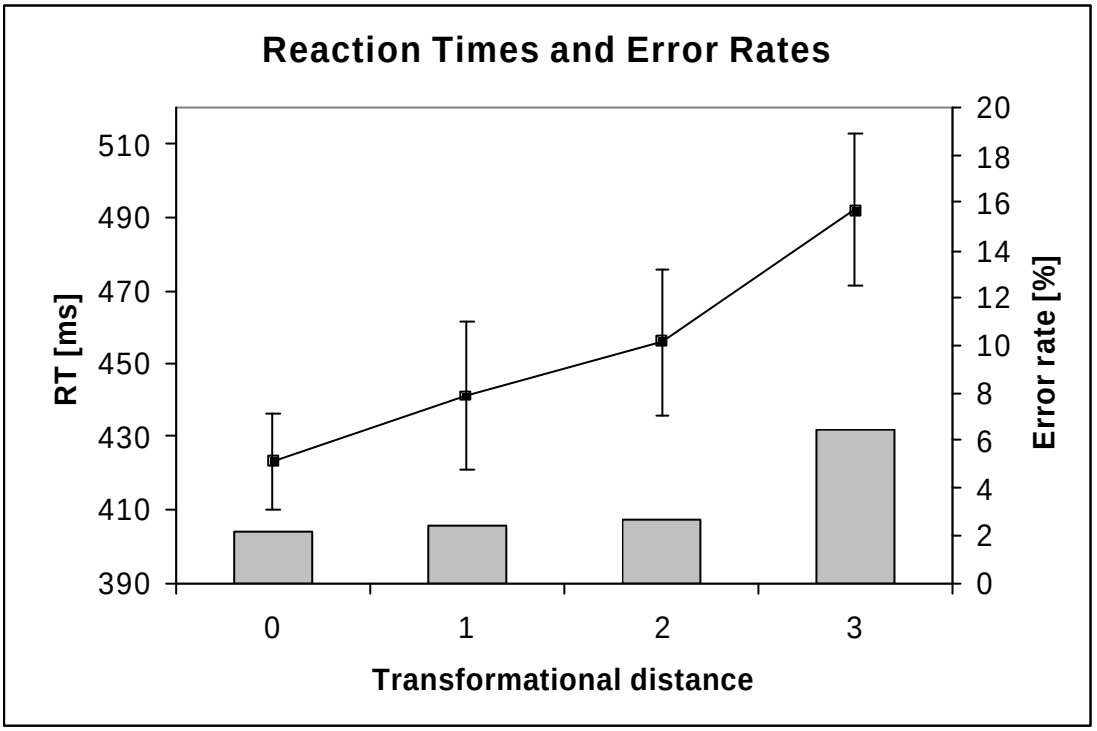

Figure 8. Basic level categorisation requires more time and is more error-prone with increasing amount of topological transformation.

Sequential additivity of transformation times was investigated, in order to test whether analogue transformation processes are involved in categorisation. Sequential additivity means that the transformation times for a given interval can be described by the sum of its partial segments. A high correlation was found between the empirical values for distance 3 and the theoretical predictions under the hypothesis of sequential additivity ($r = .87$, $p < .001$). This suggests that the transformations are analogue, i.e. traverse intermediate stages of the transformational path.

Also error rates increased with the amount of transformation (F (3,33) = 25.83, $p < .001$), indicating that the findings are not simply the result of a speed-accuracy trade-off.

In summary, the results for the more realistic 3D rendered objects nicely confirmed those of the previous study with 2D outline shapes. Categorisation takes longer and is more error-prone with increased shape transformation.

Experiment 2: Simultaneous shape and orientation changes

The objects from the CogVis database can not only be morphed, but also rotated in 3D space. This allows to investigate in a single experiment how humans categorise familiar objects when both shape and orientation changes occur. The experiment serves several purposes: First, it allows to test whether the systematic dependency on topological transformations can be replicated in more demanding circumstances, when objects are not always presented in an upright orientation. Second, the question whether basic level categorisation performance is orientation-dependent or not is further addressed. Some experiments suggest that orientation-dependency is limited to subordinate level categorisation or identification (Hamm & McMullen, 1998), while other experiments demonstrated orientation-dependency also for basic level categorisation (Hayward & Williams, 2000; Lawson & Humphreys, 1998). And third, it helps to clarify whether compensation processes for shape and orientation differences are independent, or whether they interact.

Stimuli from 26 basic level categories were created, using the CogVis object database. 24 categories were presented in the experiment; two additional categories were used in the practice phase. Objects were both morphed (within class) and rotated in the picture plane (see Figure 9). Two objects were presented sequentially, and subjects were required to decide whether both objects belonged to the same class, independent of orientation. Both the amount of shape transformation and the amount of image-plane rotation was varied. Twelve subjects participated in the experiment.

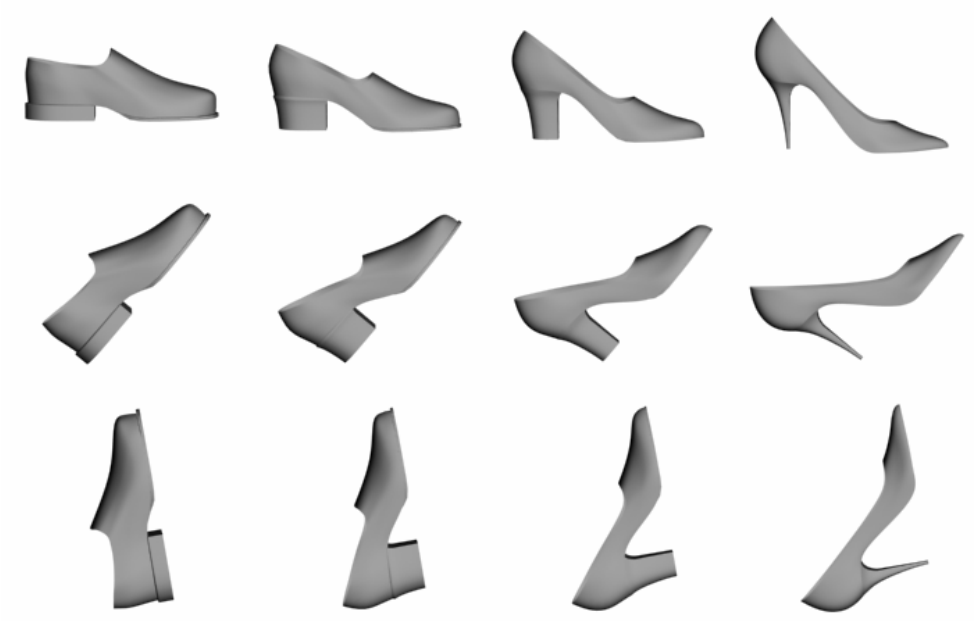

Figure 9. In Experiment 2, both the amount of morphing and of image-plane rotation was varied systematically, exemplified by the category *shoe*. Objects are morphed within rows, and are rotated in the image-plane within columns.

First, RTs increased systematically with increasing topological distance (F (3,33) = 22.38, $p < .001$). A linear trend was highly significant (F (1,11) = 37.08, $p < .001$), and also a quadratic trend was significant (F (1,11) = 10.69, $p = .007$) (see Figure 10). These

findings confirm and extend the results from Experiment 1: Categorisation performance deteriorates systematically with increasing amount of shape transformation, even when objects were presented in different orientations. Second, categorisation required more time with increasing orientation distance between the objects (F (2,22) = 15.81, p < .001). Again, a linear trend was significant (F (1,11) = 23.49, p < .001). Thus, orientation dependency was found in basic level categorisation, even in a task which required matching different members from the same category. The orientation effect was small compared to other studies, which is in accordance with the finding that orientation effects on basic level usually are smaller when the distractors are not similar (see Lawson, 1999, p. 231-232). Third, there was no significant interaction between topological and orientation distance (F (6,66) = .67, p = .67), which suggests that both processes are independent (see Figure 10). This is in agreement with evidence for independent compensation processes for rotations in the picture plane and size-scalings (Bundesen et al, 1981), and for rotations in the picture plane and in depth (Lawson et al., 2000). Consequently, the independence of both compensation processes suggests that topological transformations of shape in categorisation are processed by the brain in a similar way as other spatial transformations (like rotations).

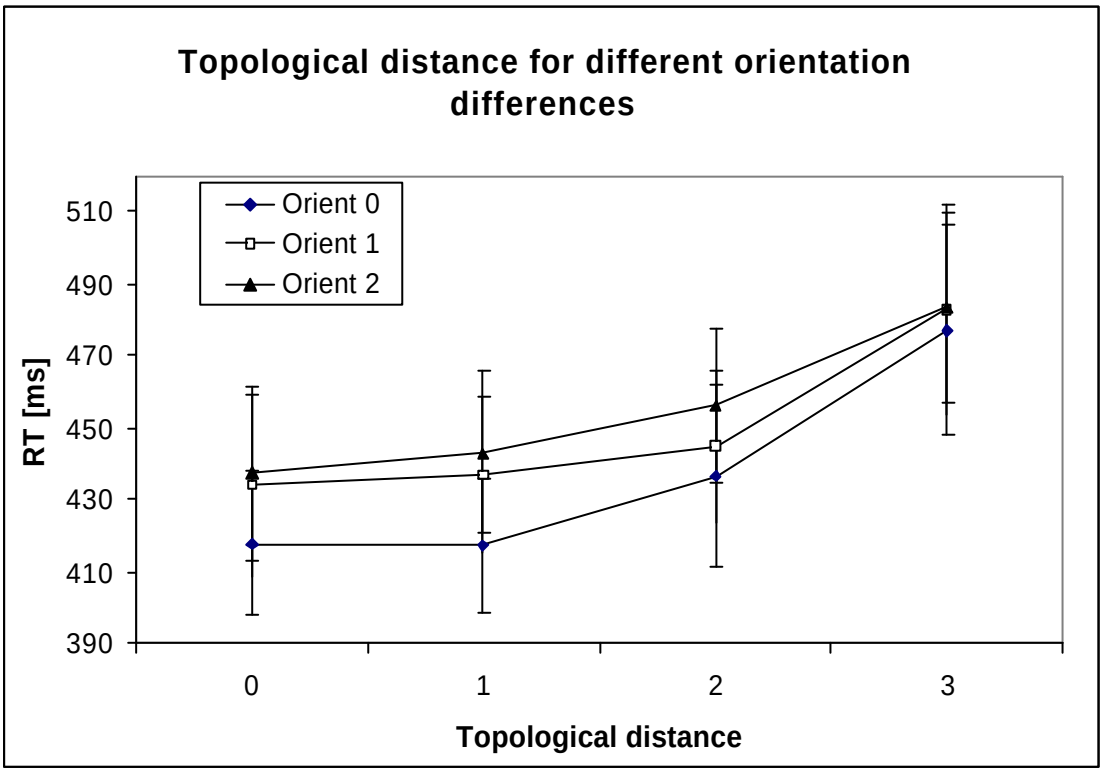

Figure 10. RTs increase with increasing amount of topological transformation for all 3 orientation conditions. RTs increased also with increasing orientation distance. Shape and orientation changes did not interact, which suggests that they are compensated by independent processes.

In this aspect categorisation is different from discrimination tasks, where an interaction between view and shape changes was demonstrated (Lawson, Bülthoff & Dumbell,

2002). In summary, both shape changes and orientation changes require processing time. Shape and orientation changes do not interact, and therefore seem to be compensated for independently, at least for plane rotations.

Error rates increased systematically with increasing topological distance (F (3,33) = 12.49, $p < .001$). Errors tended to increase also with increasing orientation difference, but this was not significant (F (2,22) = 1.86, $p = .18$). There was no speed-accuracy trade-off.

Conclusions and Outlook

Both experiments demonstrated that categorisation performance depends systematically on the amount of shape transformation, also for quite realistic grey-scale stimuli, and even when objects were rotated in the picture plane. Compensation processes for shape deformations and disorientation seem to be independent, confirming previous findings for other types of transformations. The results suggest an image-based model of categorisation. There is some evidence that analogue transformation processes are involved in categorisation, which favours the transformational (alignment) account of categorisation (Graf, 2001, 2002). However, other image-based models cannot be excluded with certainty. In order to further investigate the nature of human basic level categorisation, a study with functional MRI was conceptualised; first scans will start this year. Using fMRI we try to find evidence which allows to differentiate better between different image-based models of categorisation. Moreover, further psychophysical experiments are planned which will investigate categorisation performance for simultaneous topological transformations and rotations in depth.

2.4 Study 4: Role of context in recognition and categorisation: top-down processing and viewpoint-dependence

The overview of different object recognition theories in the introduction illustrates that most theoretical approaches to recognition and categorisation assume a serial bottom-up process. For some researchers this is consistent with recent evidence for an ultra-rapid categorisation. Van Rullen & Thorpe (2001) found evidence that the categorisation of objects can be very fast, i.e. within 350-400 ms. According to Fabre-Thorpe, Richard, & Thorpe (1998), monkeys seem to be even faster than humans, at least in some tasks. Event-related studies are consistent with the assumption of very fast bottom-up processing. For example Thorpe, Fize, & Marlot (1996) showed that event-related potentials (ERPs) to target and distractor images diverge strongly at roughly 150 ms in humans. Interestingly, for these authors the processing is so fast, that they conclude that categorisation and perception occur in parallel, whereas some processes involved may be shared. Moreover, recurrent top-down processing for categorisation seems to be implausible when these fast response times are taken into account. Note however, that deciding whether something is living or non-living could be based on elementary features like curvature and certain textures. For object recognition or naming such features rarely provide enough information. In fact, Humphreys, Van Ridden, & Price (1997) propose that bottom-up activation of semantic knowledge from vision may be insufficient to

invoke a name. Object naming requires recurrent activation of stored perceptual knowledge to differentiate activation from a target object from that present in other representations. In the same paper several results form from neuropsychology and functional imaging are discussed that provide further evidence in favour of top-down processing in object recognition.

Another interesting result was revealed by Liter and Bülthoff (1997). They found that object naming was orientation sensitive, but name verification was not. This result could be related to top-down activation, too. If it is assumed that an object consists of a collection of more or less viewpoint-specific descriptions, showing the name in advance could activate many views of the object. Such a pre-activation would help to resolve the competition with another object that has similar viewpoint-specific descriptions.

In the present study we further investigated this hypothesis using an associative priming paradigm.[8] Instead of showing the name prior to the object that has to be recognised (name verification) a more natural situation would be to show an object that is associated with the object to be recognised. For example tea spoons tend to be near to tea cups. Thus, it could be expected that looking at a tea spoon activates the viewpoint-specific descriptions of a tea cup, which would help to recognise it even in unusual views.

Experiment

Twelve participants (6 females, 6 males) had to name 64 common objects as fast and as accurately as possible. Half of the objects served as priming stimuli, whereas the other half was defined as target stimuli. Of each target stimulus, a canonical and a noncanonical view was used. For the priming stimuli, only one view was used and it was chosen to be between the two views of the target stimulus. Each trial started with a 1000 ms fixation cross followed by the priming stimulus. After the participants named the prime, a masking stimulus was presented for 1000 ms, followed by the target object. Half of the trials were consistent, i.e. the prime was associated with the target (e.g., tea spoon and tea cup). The remaining trials were inconsistent, i.e. the prime was not related to the target (e.g., tea spoon and car). There were 4 blocks of 32 trials each. In each block all 32 target stimuli were presented randomly and the experimental condition (consistent-canonical, inconsistent-noncanonical, consistent-noncanonical, inconsistent-canonical) was counterbalanced so that in one block each condition occurred 8 times. There was a total of 128 trials per experiment: 2(consistent vs. inconsistent prime) * 2 (canonical vs. noncanonical target) * 32(target stimuli).

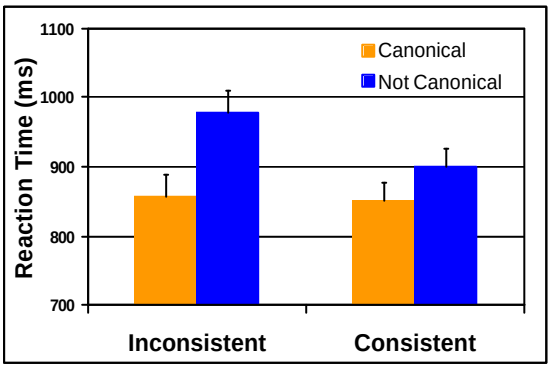

Figure 11 Reaction times for the four conditions used in Experiment 1. Error bars represent SEM.

[8] We thank Franziska Hofer, Stefan Michel, and Gisela Schoch, Department of Psychology, University of Zurich for their help with this study.

Individual data were averaged across different objects in order to eliminate an item-specific factor. Trials in which naming or technical errors occurred were excluded from the analysis and outliers were eliminated[9]. The means and standard errors for the different experimental conditions are depicted in Figure 11. A two factor analysis of variance (ANOVA) with consistency (consistent vs. inconsistent prime) and view (canonical vs. noncanonical target) as within-subjects factors was carried out on reaction times of the correct trials. There were reliable main effects of consistency, $F(1, 11) = 9.44$, $p < .05$, and view $F(1, 11) = 56.71$, $p < .001$, and there was an interaction between consistencies and view $F(1, 11) = 10.60$, $p < .01$. In order to investigate the time course of the priming effect, a three factor analysis of variance (ANOVA) with block, consistency and view as within-subjects factor was computed. There was a significant main effect of block $F(1,77, 19.48) = 37.05$, $p < .001$, consistency $F(1, 11) = 9.69$, $p < .01$, and view $F(1, 11) = 67.40$, $p < .001$. In addition, there was a significant interaction between consistency and view $F(1, 11) = 15.03$, $p < .01$, and a marginal interaction between block, consistency and view $F(1.95, 21.42) = 3.05$, $p = .07$.

Conclusion

We found clear evidence for top-down influences in object recognition. The interaction between consistency and view found in Experiment 1 shows that a priming stimulus, which is related to a target stimulus, can minimize the viewpoint dependency in recognizing the target object. Note that there are several earlier studies related to the question whether object recognition is facilitated when an object is semantically consistent rather than inconsistent with the scene in which it appears (see Hollingworth & Henderson, 1999 for a review). The main new finding by our own study and Liter and Bülthoff (1997) is however, that associated objects can activate a multiple views representation of another object. Such a 'pre-activation' could be especially helpful for recognizing objects in non-canonical views, because it can help to resolving the competition with another object that has similar viewpoint-specific descriptions (Liter & Bülthoff, 1997).

2.5 Study 5: Interplay between object motion and shape for recognition and categorisation decisions

(a) The role of motion and shape for object categorisation (Huber, Newell, Wallraven)

In this study we investigated, which perceptual cues play a role in forming perceptual categories and in the task of categorizing new objects. Object recognition and object categorisation have been typically studied either with static objects (i.e., with no dynamic

[9] From the total of 1536 trials (12 participants*128 trials) 112 trials (7.3%) were excluded from the statistical analysis because of naming errors. Additionally, 73 trials (4.8%) were discarded either because the microphone didn't record the voice or because participants started with some lutes prior to the name of the target. To eliminate possible effects of name retrieval problems reaction times greater than the mean + 2 standard deviations were also discarded (30 trials or 2%). Taken these exclusions together, 1321 trials (86%) were analyzed.

information) or with dynamically presented point-light displays (i.e., with almost no shape information). But how are dynamic and static features integrated in object representation? In particular, what is the role of movement for object recognition and categorisation when the object is presented in its entirety, i.e. when static and dynamic features are presented together?

We report three experiments where novel objects were categorised on the basis of two spatial properties (colour and shape), and two dynamic properties (action and path). The 'action' of an object referred to its intrinsic motion pattern, whereas 'path' referred to an object's extrinsic motion pattern, i.e. the route an object took. The task for the participant was to first learn to categorise prototype objects and then categorise new exemplar objects, which varied in number and type of properties which they shared with the prototype.

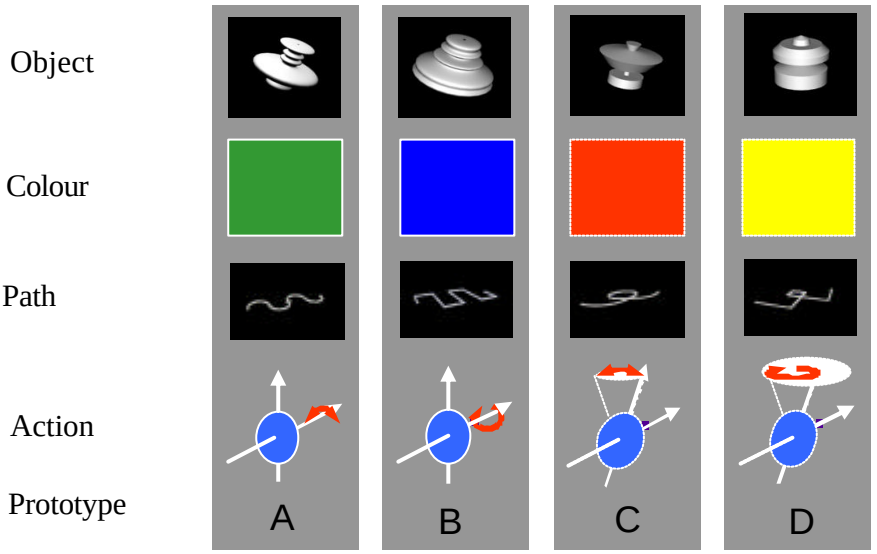

Figure 12 This figure shows the four prototypes used in our experiments together with their constituting properties. Stimuli were created using 3D Studio Max 3.0 and rendered as 320x160 pixel avi sequences (Indeo-Codec) consisting of 300 frames with 30 frames/sec. The shapes were defined by discontinuous and continuous curves in two dimensions, which were rotated around the upright axis to create three-dimensional objects. All paths had equal length and were determined by a rectangular and sinusoidal wave pattern and by a smooth and sharp loop, respectively. The four colours were pure red, green, blue and yellow. Four types of actions were defined by either a swinging or a continuous rotation of the objects around the upright or horizontal axis such that each action was completed four times during the sequence. Objects were placed in a 'room' consisting of a checkerboard-pattern floor and two grey walls with the start point of the sequence in one corner of the room and the end point in the opposite corner. A spot-light illuminated the scene from above to create a shadow of the object on the floor in order to facilitate perception of depth and object motion. Prototypes were defined by selecting four sets of four features. Exchanging one or more features between prototypes yielded the whole set of stimuli for the experiments.

Experiment 1: The design of experiment 1 was based on a two-way mixed design with one between-subjects factor (paired prototypes learned) and one within-subjects factor (feature changes from prototype). The between group factor had two levels (AB, CD prototype pairings or AC, BD prototype pairings). The within group factor had five levels indicating the feature differences between the exemplar and the prototype (shape, colour, path, action and shape+colour/path+action).

The experiment consisted of two phases: a learning phase with feedback followed by a test phase without feedback. In the learning phase, each subject was shown a movie file and instructed to learn the object and press one of two buttons indicating one of the two learned prototypes. Six trials were presented and subjects received feedback for each trial. Test stimuli were derived from the prototype either by changing a dynamic feature or a static feature. Feature changes were counterbalanced across all participants. In the test phase, the task for the subject was to correctly associate each exemplar under either one or two feature changes with the appropriate category. The experiment consisted of two blocks and participants could take a self-timed break between blocks. Each block consisted of a pair of prototypes. The order of the blocks was counterbalanced across subjects. Participants were told to categorise each object as accurately as possible and to consider all information present in the stimulus as relevant for categorisation.

Results: The error rates across subjects for all trials were then calculated as a bias from the actual percentage difference between the exemplar and the prototype. The mean percentage bias for each feature change is presented in Figure 13. A positive bias means that the participants were sensitive to this feature and tended to over-estimate changes from the prototype with changes to that feature. A negative bias, on the other hand, meant that the participants did not use changes to this feature in their categorisation decisions. The only statistically significant bias we could find was a negative bias for the path feature ($Z = 3.95$, $p < 0.001$).

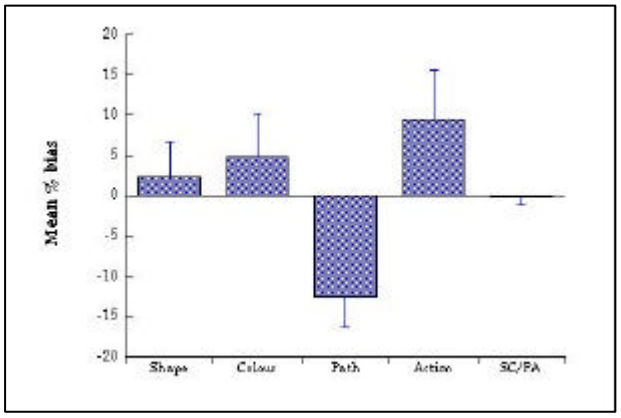

Figure 13. Results from Experiment 1. The bars depict the mean percent bias from the expected results.

Discussion: The two major findings from this experiment were that path was effectively ignored and that participants are as likely to use dynamic cues as static cues for object categorisation. One reason for the negative path bias might be that path *per se* is not as indicative of category membership as the other features (we seldom would classify e.g., an animal as a predator by using its walking path). In order to test whether path as a feature is used at all, we conducted a second experiment.

Experiment 2: In the second experiment we tested for differences in perceptual saliency of the four features in order to ensure that no single feature (in particular path) would be overpowered by the other features. The experiment was based on a 2 way, repeated measures design using number of feature differences (1 or 3) and feature type (shape,

colour, action and path) as factors. In any one trial, subjects had to rate the similarity of two objects using a scale from 1 to 7, where a rating of 1 indicated a high degree of similarity. One of the objects was always a prototype object and the other an exemplar object. The exemplar differed from the prototype either in 1 feature or 3 features. The two objects were presented next to each other and started moving at the same time. The position of the prototype (left or right) was counterbalanced. Participants had to participate in four blocks which were presented in random order. The blocks differed in the pair of prototypes used (AB, AC, BD, or CD). In each block participants conducted 2 similarity ratings for each 1 and 3 feature change and each prototype resulting in 32 trials per block. In particular, if all features have a similar perceptual saliency, the similarity ratings for single feature change as well as for a three feature change should show *no* difference between feature types.

Results: The mean ratings per feature for 12 participants are shown in Figure 14. We found a significant difference between the one and three feature changes ($F(3,33) = 3.89$, $p < 0.05$). The only significant effect for one feature changes was that colour changes were rated as more similar than action changes ($p < 0.05$). There were no significant differences between the three feature changes, however.

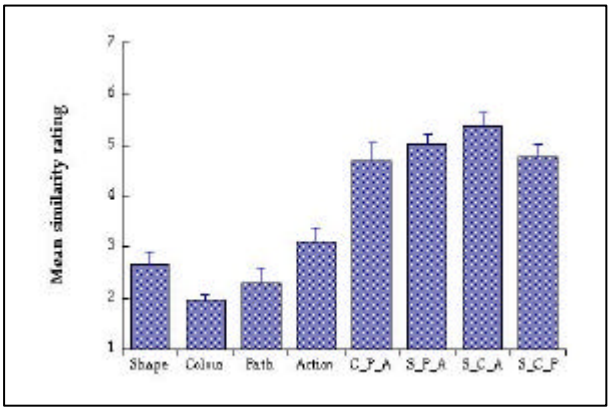

Figure 14. Results from Experiment 2. The bars show the mean similarity ratings for either one- or three-feature changes.

Discussion: The main result of the second experiment was that no feature can overshadow all the other features in the categorisation task. Thus there seems to be a largely uniform saliency for all four features. The difference for the colour change seems not to have an overall effect as it is not present in the three feature changes.

Experiment 3: Experiment 2 revealed that path was perceptually salient. Thus, the cause for not using path for categorisation in Experiment 1 was *not* due to the fact that path was not perceived by the participants. Instead, it could be possible that path was overshadowed by a stronger action feature. To test this hypothesis, we conducted a third experiment, in which action was the same for all prototypes and thus reduced the set of discriminating features to colour, shape and path. The design followed that outlined in Experiment 1 for the AC, BD group of participants only. Feature changes were therefore counter-balanced across prototype pairs for each subject. The experiment was based on a repeated measures design with feature type as the main factor (shape, colour, path). Trials were randomly presented across subjects according to the constraints of feature allocation to the learning and test trials described in Experiment 1 above.

Results: The mean percentage bias for 16 participants calculated as in experiment 1 for each feature change is presented in Figure 15. We found no evidence of a bias for any of the features (shape, $Z = 0.25$, n.s.; colour, $Z = 1.25$, n.s.; path, $Z = .25$, n.s.). However, an

analysis of only the test trials revealed that colour was significantly different from path ($p < 0.05$).

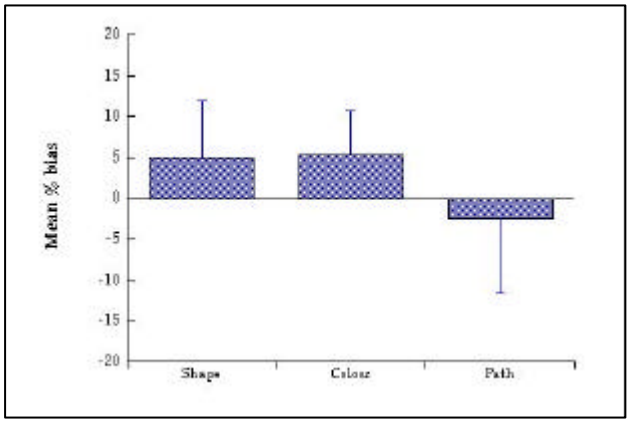

Figure 15. Results from Experiment 3. Shown here is the mean percent bias for the three types of feature changes

Discussion: The major result of the third experiment was that with a redundant feature action, path was used as readily as shape and colour to categorise the objects. Again, the dynamic cue was as readily used as the static cues. Thus, path is not a feature that is not used *per se* for categorisation. The significant difference between colour and path, however, seems to indicate that participants still have a tendency to use object intrinsic features such as colour and shape (and also action, see experiment 1) more often than object extrinsic features such as path.

General Discussion

Traditionally, categorisation and recognition have often been studied exclusively in the static or dynamic domain. We presented results from experiments on categorisation, which aimed to bring both types of cues together in a perceptually relevant and realistic task of categorisation of novel objects. Our main result is that when presented with a number of static and dynamic cues, participants readily made use of both types of information. First of all, this shows that all of these features are accessible to visual memory and thus are part of the stored and learned representations of these objects. This again supports the results from Wallis et al. (2001), where it was shown that the dynamic properties of objects were encoded during learning. It also supports the view that – without any prior information about cue saliency – there is no intrinsic advantage of static cues in such a task (this is e.g., in contradiction to results from Mak et al., 1999). Our second main finding was that there seems to be a slight disadvantage in cue saliency for extrinsic cues such as the path an object takes, which was found in experiments 1 and 3. In an ecological perspective such a strategy makes sense as extrinsic properties of an object are less salient with regard to its identity than intrinsic ones (the path an animal takes will be less indicative than the way it moves or its shape). In general, it can be said that this line of experiments speaks in strong favour of a cue integration model of object recognition – more specifically, the cues present in our experiments seem to have been represented *independently* (we found virtually no interactions between features) as largely orthogonal dimensions in the object representation. Our findings thus explicitly support the computational research on cue integration strategies, which is an integral part of the CogVis project.

(b) Effects of motion and orientation upon featural and configural processing of emotional perception

In addition to rigid motion also non-rigid motion could play an important role in processing the shape of an object. For example Knappmeyer, Thornton and Bülthoff (submitted) have shown recently that during the computation of identity the human face recognition system accesses and integrates individual non-rigid facial motion and individual facial form. Another highly relevant example for an interaction between non-rigid motion and shape is the processing of emotional expression in faces. We have shown in study 1 that part-based information and their spatial relations (configural information) play both an important role in face recognition (Schwaninger et al., 2002). Interestingly, inverting the eyes and mouth in an upright faces creates a very bizarre facial expression which disappears when the face is turned upside down (Thatcher illusion, Thompson, 1980). In this study we investigated to what extent non-rigid motion and shape information interact in affecting this illusion (Schwaninger, Cunningham, & Kleiner, 2002).

It was revealed in study 1 that faces are processed by encoding part-based information as well as spatial relational information between many facial features (configural information). According to the model presented in study 1 (see Figure 4) both types of information are integrated in order to activate identification units. As a consequence, thatcherized eyes and a mouth presented in isolation should be rated as less bizarre than when they are shown within the facial context because configural information is drastically reduced. It has been shown by several previous studies that processing configural information is impaired when faces are rotated (for a review see Valentine, 1988; Schwaninger, Carbon, & Leder, in press). Therefore, it is very difficult to perceive that the eyes and mouth have been altered in an inverted Thatcher face. In other words, an interaction between condition (isolated parts vs. whole Thatcher face) and orientation is predicted. By comparing static vs. moving conditions we could investigate the interaction between non-rigid motion on one hand and processing featural and configural information on the other hand.

Experiment 1

Twenty undergraduates (10 females) from the University of Zürich participated in this study. Stimuli were created by recording a single subject with a stereo camera system while talking and smiling. A 3D facial form algorithm was used to fit a three dimensional morphable model (Blanz & Vetter, 1999) to the video images in order to extract images of the facial texture and eliminate effects of rigid head motion and orientation. In each texture image, thatcherized versions were created by mirroring the regions of the eyes and mouth around the horizontal axis. The modified texture was reapplied to the subjects 3D head model and the model was rendered in front of a black background using standard computer graphics. For Experiment 1, 4 sequences were selected. Each sequence contained one smile and lasted 1.5 s (30 frames). Static versions were created by selecting the frame representing the peak of the smile. Moving and static stimuli were shown upright and inverted and as whole faces as well as parts (i.e., just the eyes and mouth). Examples of the static stimuli are shown in Figure 16. Each trial started with a 1

s fixation cross, followed by a 1.5 s stimulus presentation. Apparent bizarreness was rated from 0 to 9. There were 32 trials per block: motion (moving vs. static) * orientation (upright vs. inverted) * information type (parts vs. wholes) * 4 sequences. These trials were repeated four times in separate blocks resulting in a total of 128 trials. The order of trials was randomized within each block.

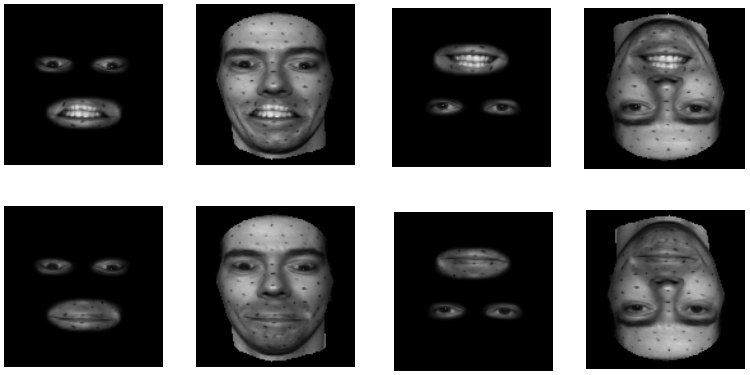

Figure 16. Sample stimuli (static condition only).

Static Thatcher faces were rated as more bizarre upright than when inverted (7.73 vs. 2.68, $p < .001$). This large difference validates the use of bizarreness ratings for investigating the Thatcher illusion. Static upright Thatcher faces were rated as more bizarre than the isolated parts (7.73 vs. 6.18, $p < .01$). This result is consistent with the assumption that for upright faces two sources of information contribute to perceived bizarreness, namely processing featural as well as configural information.

The bizarreness ratings of static stimuli were subjected to a two-way ANOVA with orientation and information type as within-subjects factors. There was a main effect of orientation, and an interaction between orientation and information type (all F's (1, 9) > 17, p's < .01). This interaction is consistent with the assumption that inversion impairs configural processing more than processing featural information (for a review see Valentine, 1988; Schwaninger et al, in press).

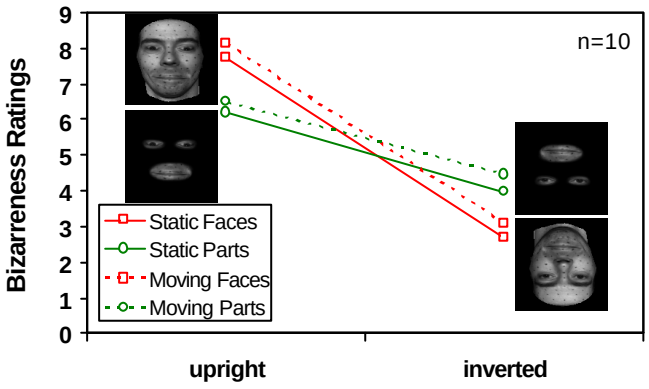

Figure 17. Results from Experiment 3. Shown here is the mean percent bias for the three types of feature changes

The effect of motion was investigated using a three-way ANOVA with orientation, information type, and motion as within-subjects factors. There was a main effect of motion, $F(1,9) = 12.59$, $p < .01$, indicating that motion increases bizarreness. Interestingly, there were no significant interactions (all F's (1,9) < .3). That is, motion increased perceived bizarreness by about the same amount regardless of orientation or information type.

Experiment 2

Experiment 2 replicates and extends Experiment 1 to talking faces. The materials and procedure were identical to Experiment 1 except that talking instead of smiling sequences were used. The static faces were created by selecting one of the 30 frames on a random basis.

The data were analyzed using the same methods as in Experiment 1. The main results were replicated. The two-way ANOVA on bizarreness ratings of static stimuli revealed a main effect of orientation and an interaction between orientation and information type (all F's (1, 9) > 10.03, p's < .05). The effects of motion were again analyzed using a three-way ANOVA. There was again a main effect of motion, $F(1,9) = 10.91$, $p < .01$. As in Experiment 1, neither the information type by motion interaction nor the three-way interaction were significant (both F's (1, 9) > 1.52, p's > .24). There was, however, an interaction between orientation and motion, $F(1, 9) = 13.88$, $p < .01$: Motion increased perceived bizarreness for all conditions, although not by much in the inverted whole talking faces condition.

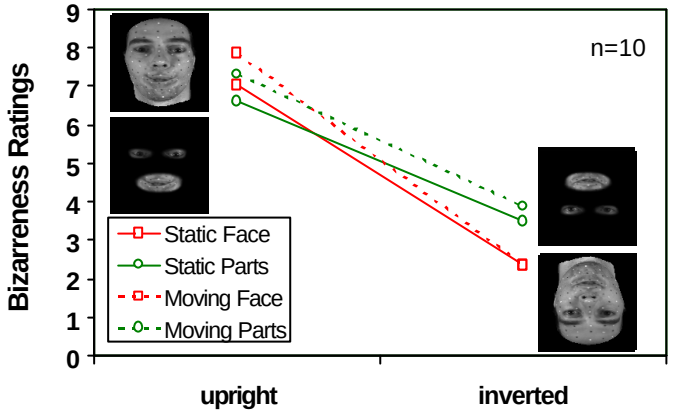

Figure 18 Results from Experiment 2 (talking faces). Mean bizarreness ratings for upright and inverted faces and parts presented in motion or static.

Conclusion

Face context can increase the perceived bizarreness of the parts, which is especially evident in smiling faces. Moreover, in both experiments inversion reduced bizarreness of whole Thatcher faces more than for part versions. These results are consistent with the idea of configural and featural processing for upright faces and impaired processing of configural information when faces are inverted (for a review see Schwaninger et al., in press).

Motion seemed to increase bizarreness similarly in nearly all conditions. There are at least two general classes of explanations for this. First, motion provides some form of global conflict. For example, eye blinks consist primarily of upper lid motion. When the eyes were rotated, however, the eyes closed primarily from the direction of the mouth. Second, it has been proposed that static facial information is processed in the ventral stream and motion in the dorsal stream (O`Toole, Roark, & Abdi, 2002). The addition of motion, then, could simply increase static bizarreness.

3 Conclusions: Our view on recognition and categorisation

In summary, the main aim of our studies was to investigate a number of issues in the field of object recognition and categorisation, which are of central importance in order to build a successful cognitive vision system. We focused on the questions whether parts play a role in recognition, how shape similarity and basic level categorisation can be accounted for, whether context and top-down information is important, and how object motion and shape interact. Our results confirm and complement existing image-based model of recognition and categorisation. In this chapter we present our view on recognition and categorisation, which extends existing models summarized in chapter 1, and is based on our psychophysical studies described in chapter 2. Our view is illustrated in Figure 19, which depicts an integrative framework that serves as a theoretical basis for a computational recognition system grounded in cognitive research. We will first present the foundations of our basic approach, and provide a more detailed picture afterwards, based on the results of our psychophysical studies.

Visual recognition and categorisation is achieved by matching the visual input to stored memory representations. The input can be envisaged as a pictorial, appearance-based and dynamic flow of visual information which is induced by the object or scene. A preprocessing stage extracts low-level features like colour, orientation, motion, texture and other properties. These features can be used to select possible candidate representations from visual memory in order to match them against the input for recognition and categorisation.

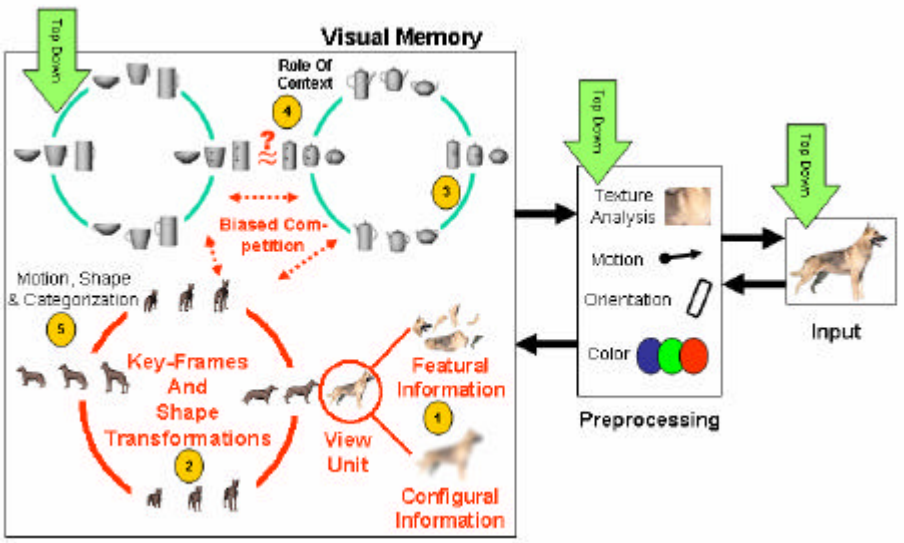

Figure 19. Integrative model of recognition and categorisation, based on the results of our studies (indicated by numbers 1-5).

The matching can occur both on the basis of featural and configural information (study 1). The outputs of featural and configural matching is pooled by view-units. Up to the basic level of categorisation, recognition and categorisation rely on image-based or

pictorial category representations. The recognition of disoriented objects is based on distributed and view-based representations. The temporal coherence of the visual input plays a fundamental role in the learning of object representations (e.g., Wallis & Bülthoff, 2002). Basic level categorisation is achieved by an alignment process which brings corresponding parts of the stimulus representation and memory representation into correspondence. The alignment involves time-consuming and error-prone shape transformations (which can be conceptualised as spatial transformations). Thus, the matching requires more time with increasing amount of shape transformation which is necessary for an alignment. Processes which compensate for changes of shape and image-plane orientation seem to be independent (study 2, 3). Moreover, the recognition and categorisation process can be modulated by expectations which are provided by the scene context. Especially unfamiliar views are facilitated by contextual information, which suggests that top-down processing can bias the competitive interactions between groups of neurons that encode object representations, or pre-activate specific memory representations (study 4). Recognition and categorisation can also be mediated by rigid and non-rigid motion information in addition to shape cues (study 5).

In the following we will provide a more detailed description of this basic model of recognition and categorisation, based on the results of our psychophysical studies.

The models described in the introduction (chapter 1) differ with regard to their assumptions on structural and temporal aspects of recognition memory. On one hand, invariant property models, classical feature models and structural description models generally assume viewpoint-independent performance. In contrast, most image-based models predict viewpoint dependent performance. In the last two decades a remarkable amount of psychophysical evidence showed that performance usually is viewpoint dependent. This was confirmed for different object classes as well as different tasks like priming, naming, old-new or matching tasks (for a review see Jolicoeur & Humphreys, 1998; Lawson, 1999; Tarr & Bülthoff, 1998). Our own studies are consistent with the large body of evidence suggesting that recognition and categorisation performance is essentially viewpoint dependent (see our studies 1, 3, and 4). Although models predicting view-invariant performance bear little psychophysical plausibility, we do not claim that all aspects of these models are wrong. Instead, we actually provided evidence that part-based information is important for recognition (study 2), which is in accordance with several other studies (Biederman & Cooper, 1991; Goldstone, 1996; Tversky & Hemenway, 1984).

One potential caveat of view-based schemes is that most of them are holistic, i.e. they process the whole object without representing parts or features explicitly. In our study 1 we revealed that even the perception of faces – a stimulus class that has often been cited as the example for exclusive holistic processing in adults - relies on featural information in parts and relational information specifying the position of the parts (configural information). According to the model proposed by Schwaninger et al. (2002) both types of information are processed, represented in memory, and used for recognition. The idea of representing objects by their parts and spatial relations has been proposed many years ago by structural description theories (e.g., Biederman, 1987; Marr, 1982). Note however, that several important differences exist between these structural description models and our concept of featural and configural information (see also Wallraven, Schwaninger,

Schuhmacher & Bülthoff, 2002). First, in contrast to the traditional approaches by Marr and Biederman, our model does not rely on edge-based representations. Second, the parts we propose are completely different both conceptually and computationally from the geometrical primitives (geons) used in the approaches in Biederman (1987) and Hummel and Biederman (1992). Geons are defined by using Lowe's (1987) nonaccidental properties and are meant to be viewpoint-independent (or at least for a certain range of views, see Biederman & Gerhardstein, 1993). In contrast to view-invariant geons, we propose that part-based representations are formed by grouping image features, which often are viewpoint dependent. Indeed, there are several lines of evidence against the concept of viewpoint-invariant recognition based on geons. For example Tarr, Bülthoff, Zabinski, & Blanz, (1997) have shown reliable effects of viewpoint for the recognition of objects that were made of 1, 3, or 5 different geon-like parts. Moreover, Hayward and Tarr (1997) have found viewpoint dependent performance for geon-like objects although view-invariant performance would have been predicted for such objects even according to an extended version of RBC proposed by Biederman and Gerhardstein (1993). Most importantly, Tarr, Williams, Hayward, & Gauthier (1998) have revealed that already the processing of one geon is dependent on viewpoint, which accounted for sequential matching, matching to sample and naming tasks. Finally, another aspect in which our view differs from structural description models is the number of parts and how they are acquired. According to RBC theory a small and fixed set of geons suffices to explain the relevant aspects of human object recognition (36 geons according to Biederman, 1987, and 24 geons according to Biederman, 1995). From our point of view, human recognition performance relies on many more features, which often are determined by perceptual learning, and therefore are not fixed, but dependent on the history of the subject and the task (Schyns, Goldstone & Thibaut, 1998; Schyns & Rodet, 1997).

Another important question to be addressed with regard to structural aspects of recognition memory is how featural information and information about their spatial relations (configural information) can be combined to an integrated image-based representation. Wallraven et al. (2002) have shown that human recognition performance can be modelled by a slightly modified version of the key-frame model proposed by Wallraven and Bülthoff (2001). In this scheme, objects are represented by a set of temporally associated key-frames. Key-frames consist of a number of salient features *and* their spatial relationship that are relatively stable for a limited set of neighbouring views. This framework thus explicitly represents featural and relational information in a viewpoint-dependent manner. The recognition process is modelled by matching the input representation to key-frames using n-nearest neighbours and is therefore similar to interpolation models of recognition discussed in chapter 1.4.2. The recognition process in the key-frame approach combines featural information and information about their spatial layout (configural information). In addition, the key-frame framework can also account for the findings of Bülthoff and Edelman (1992), Edelman and Bülthoff (1992), and Wallraven et al. (2002), who compared different view-based approaches with regard to their predictive validity for recognizing unfamiliar objects (wire-frame like objects and amoebae) as well as highly over learned stimuli (faces). The results of all three studies provided converging evidence against view-independent models. The results were compatible with interpolation models and were highly consistent with the detailed predictions derived from the key-frame model proposed by Wallraven et al. (2001, 2002).

However, the findings could not be explained by a linear combination of 2D views (Ullman & Basri, 1991), or by a 3D alignment model (Lowe, 1987; Huttenlocher & Ullman, 1990) which relies on an alignment process that is not error-prone.

Studies 2 and 3 dealt with the question how a model of basic level categorisation should be conceptualised. Starting point was the observation that different members of the same basic level category usually can be aligned by rather simple deforming shape transformations. Thus, shape variability within common and familiar basic level categories can be described well with continuous deforming shape transformations (see also the morphed objects we created for the CogVis object data base). We investigated whether categorisation performance and perceived similarity are systematically related to these deforming shape transformations. Study 2 showed that perceived similarity of line drawings is systematically related to the amount of shape transformation. These results were not simply due to affine transformations or changes in the configuration of parts, because highly similar results were found also for those categories with little affine change, or with small changes in the configuration of parts. Study 3 investigated performance in a speeded categorisation task. Experiments with grey-level images from the CogVis object data base showed that categorisation performance deteriorated systematically with increased shape transformation, confirming previous findings with line drawings. Moreover, the systematic deterioration of categorisation performance with increased shape transformation was also found when objects were rotated in the image-plane. Overall, the systematic dependency on the amount of shape transformation was demonstrated for rating tasks and speeded categorisation tasks, for line drawings and grey-level images, as well as for upright and plane rotated objects. When both topological distance and image-plane orientation were manipulated, the two effects did not interact, which suggests that they were compensated independently. This is in accordance with independent effects for other combinations of spatial transformations.

In summary, these findings suggest an image-based model of categorisation, which is in accordance with earlier findings (e.g., Rosch et al., 1976). The systematic relation between categorisation performance and the amount of shape transformation is reminiscent of the dependency of recognition performance on the amount of rotation and size-scaling. In principle, different image-based models may account for these results. However, for a number of reasons an alignment model of categorisation seems best suited, in which categorisation is achieved by an alignment of memory representation and stimulus representation. First, sequential additivity of transformation times provides some evidence that transformations are analogue processes, i.e. pass through intermediate points on the transformational path (study 3). This result is best compatible with an alignment model of categorisation, which relies on analogue spatial compensation processes. Second, the alignment model also explains the systematic relation between similarity and the amount of shape transformation (study 2), while other image-based models only assume this relation, but cannot explain it. Moreover, the proposed model is compatible with a number of studies from the categorisation literature which advocated a structural alignment model of similarity and categorisation (for review see Markman, 2001). Overall, the alignment model seems to provide the most parsimonious account, but at present, other image-based models of categorisation – like the interpolation or threshold model – cannot be excluded. We are currently planning to perform further

investigations in order to distinguish between different image-based models of categorisation.

Categorisation within the alignment approach can be conceptualised on the basis of 2.5D representations (Graf, 2002). The category representation can be conceptualised as a superposition composite (i.e. average) of the category exemplars, containing information both about the exemplars and the prototype at the same time. A category can be defined by an averaged shape and a range of tolerable shape transformations. The range of tolerable transformations can be influenced by context and top-down processing (e.g., Labov, 1973). To extend the alignment approach to deforming transformations has further advantages, because this allows to account for the recognition of deformable objects (like animals), or articulated objects, like scissors (see already Ullman, 1989).

It should be noted that alignment models are not limited to holistic processing of objects. Even though Hummel (2000) argued that image-based models with linear compensation processes are in principle not compatible with structured representations, he had to admit that his arguments do not apply to models that involve deformable transformations. The proposed alignment model with deforming transformations implies that corresponding parts or features are identified and brought into alignment, and thus already entails the notion of structured representations. Therefore, an alignment that allows for deforming transformations is compatible with the notion of structured category representations, and – unlike the structural description approach – does not imply the problematic notion of invariance regarding spatial transformations. Similarly, Basri et al. (1998) argued that structured representations can be combined with elastic matching methods. Moreover, structural alignment models of similarity and categorisation were suggested in which categorisation is achieved by aligning corresponding object parts in structured representations (e.g., Markman, 2001; Markman & Gentner, 1993; Markman & Wisniewski, 1997; Medin, Goldstone & Gentner, 1993). Current structural alignment models are still propositional and not image-based, but they indicate that the alignment approach is nicely compatible with structured representations (which entail featural and relational information, see study 1), and offer an interesting alternative to structural description models (Biederman, 1987; Hummel & Biederman, 1992; Marr, 1982).

Flexible (or deformable) template models from the computer vision literature are highly similar to the alignment approach, as they are based on flexible deformation processes prior to matching (e.g., Jain, Zhong & Lakshmanan, 1996; Yuille, 1991; Yuille, Ferraro & Zhang, 1998). These models are in agreement with the finding that categorisation performance depends systematically on shape transformations – if they assume time-consuming deformation processes. One of the best known flexible template models was developed by v.d. Malsburg and collaborators (Lades et al., 1993). It was designed as a general model of object recognition, but not as a model of basic level categorisation. Recently, however, deformable template models of categorisation were suggested (e.g., Belongie, Malik & Puzicha, 2002).

The matching of stimulus representation and memory representation is usually performed first at the basic level of categorisation, i.e. the stimulus representation is aligned with a rather coarse shape representation of the category. Thus, basic level categorisation is achieved before the object is identified as an individual (e.g., Liu, Harris & Kanwisher, 2002). For subordinate level classification additional perceptual processing

is necessary, while superordinate level categorisation requires additional conceptual exploration (Jolicoeur et al., 1984; Rosch et al., 1976; Tanaka, Luu, Weisbrod & Kiefer, 1999). However, the entry point of recognition can shift down to the subordinate level for atypical objects (e.g., penguin instead of bird, see Jolicoeur et al., 1984) or – with increasing expertise – even to the level of the individual exemplar (e.g., Tanaka, 2001).

Recognition and categorisation performance also depends on the orientation of the objects. Moreover, study 3 indicated that the effects of shape and orientation changes are independent – and thus are compensated for independently. A possible neurophysiological model that accounts for these findings can be created according to the findings of Wang et al. (1998), which indicate that the orientation of objects is continuously mapped in the visual cortex: The positions of activation spots changed gradually along the cortical surface as the stimulus face was rotated in depth. Intermediate orientations are coded in intermediate locations on the cortical surface. This organisation of views in the cortex also corresponds to the spatio-temporal view patterns of the visual input, and is compatible with the key-frame approach (Wallis & Bülthoff, 2002; Wallraven et al., 2001, 2002). When an object in an unusual orientation has to be recognised, the activation may have to proceed along these cortical paths, such that more time is required for increasing orientation differences. In an analogous manner, shapes and shape transformations may be organised systematically in the cortex: Similar shapes may be located in nearby positions in the cortex, in such a way that shapes of intermediate transformational distances are located in intermediate positions.[10] Again, the matching of similar shapes may proceed along these neuronal pathways, corresponding to analogue compensation processes. Compensation processes for orientation and shape may then be processed independently, i.e. by different modules in the cortex. This model provides a possible implementation for our integrative model depicted in Figure 19.

Moreover, this framework is compatible with findings which indicate the importance of temporal order and temporal binding in object recognition: Memory representations of different images of the visual input stream, that were perceived in temporal contiguity, are bound together to the same category or object identity (Wallis & Bülthoff, 2001; Wallis et al., 2001). It also seems that the direction of time flow is encoded in the object representation and is actively used during recognition (Stone, 1998, 1999). This can be accounted for in two ways – within the key-frame model and within a transformational approach. First, the key-frame model (see Wallraven et al., 2001, 2002) provides a temporal binding of spatio-temporal input. In this approach, the dynamic visual input is processed on-line in order to segment it into coherent parts. The segments are defined by how long a set of visual features can be tracked reliably across the sequence. In this way, one can use the temporal contiguity of the input to automatically learn important views of the sequence which are given by the endpoints of the segments. The final representation of the system then consists of a number of *temporally linked* key-frames, each of which contains a set of salient visual features. These key-frames can also be seen as a set of connected view-tuned units, which represent a specific exemplar of a category and incorporate explicit temporal information. Second, the temporal aspects can be accounted for within a transformational (alignment) approach by assuming that knowledge about perceived transformations (e.g., Landau, 1994; Zaki & Homa, 1999) is

[10] This does not imply that categories are represented by separate modules in the cortex. Instead, we assume a distributed representation (e.g., Ishai, Ungerleider, Martin & Haxby, 2000).

encoded and used for recognition. Both approaches are difficult to distinguish experimentally.

The role of context was explored in study 4. Objects in the real world do not occur in a random manner. Looking for coffee is more successful in kitchens than in woods and tea spoons tend to be near tea cups. A cognitive system, which is adapted to the environment, would take such co-occurrences into account and use top-down driven expectancies for faster and less viewpoint dependent recognition. This could be achieved by priming image-based representations in visual memory (top-down arrow and green area in top left of Figure 19). Such pre-activation would be especially helpful when non-canonical views have to be recognised, which often tend to be similar to views of other objects. Indeed, we found in our study 4 that non-canonical views of an object were recognised much faster, when they were primed by another object that tends to co-occur in the same scene. Top-down processing could be related to competitive interactions between object representations. The biased competition model, which has been derived from neurophysiological evidence, explains how this could be implemented in the brain (Chelazzi, Duncan, Miller, & Desimone, 1998; Chelazzi, Miller, Duncan, & Desimone, 2001; Luck, Chelazzi, Hillyard, & Desimone, 1997). According to this model, object representations compete for input. Competitive Interactions are strongest when stimuli activate cells in the same local region of cortex (V4 and IT) and the competitive interactions are biased in favour of one stimulus. This bias not only includes stimulus-driven bottom-up process but also top-down processes. The main source of the top-down bias is thought to derive from prefrontal cortex (working memory). The feedback bias is not purely spatial, i.e. processing can be biased in favour of stimuli possessing a specific behaviourally relevant shape, texture, colour, and so on, in parallel throughout the visual field. Thus, top-down effects are not restricted to pre-activations in visual memory. They could also be used to prime preprocessing mechanisms and even influence what is being encoded from the input representation (for a review see Humphreys et al., 1997; Kosslyn, 1994).

In another series of experiments (study 5) we showed that both static and dynamic cues are important for categorisation of objects. We have found evidence for an increased perceptual weighting of object intrinsic cues (shape, colour and action) over extrinsic cues (path). In addition, however, we have demonstrated that in the absence of a stronger cue, the weaker cue is readily used in the categorisation task. Again, this argues for an explicit cue integration strategy, which in addition should be task-dependent and constantly updating its cue saliencies. Furthermore, the results from the experiments hint at the richness of object and category descriptions in visual memory, which one has to take into account when developing truly cognitive systems. This seems to be particularly true for objects with diagnostic non-rigid motion as illustrated by the perception of moving thatcherized stimuli.

To summarize, our investigations both confirm and extend existing image-based models of recognition and categorisation, providing the following main results: First, recognition involves both featural and relational information. Second, similarity ratings and basic level categorisation performance are systematically related to shape variations within basic level categories. Third, top-down contextual expectations play a role in object categorisation. Fourth, motion cues can be integrated with form cues for

categorisation decisions. Image-based models can account for these findings, and they are in accordance with previous findings showing viewpoint dependency of recognition and categorisation.

References

Ashbridge, E., & Perrett, D.I. (1998). Generalizing across object orientation and size. In V. Walsh & J. Kulikowski (Eds.), *Perceptual constancy. Why things look as they do* (192-209). Cambridge: Cambridge University Press.

Ashby, F.G., & Maddox, W.T. (1994). A response time theory of perceptual separability and perceptual integrality in speeded classification. *Journal of Mathematical Psychology, 38*, 423-466.

Ashby, F.G., & Maddox, W.T. (1996). Perceptual separability, decisional separability, and the identification – speeded classification relationship. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 795-817.

Barsalou, L.W. (1992). Frames, concepts, and conceptual fields. In E. Kittay and A. Lehrer (Eds.), *Frames, fields, and contrasts: new essays in lexical and semantic organization*. Hillsdale, N.J.: Erlbaum.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577-660.

Barsalou, L.W., & Hale, C.R. (1993). Components of conceptual representation: from feature lists to recursive frames. In I. Van Mechelen, J. Hampton, R.S. Michalski, & P. Theuns (Eds.), *Categories and concepts. Theoretical views and inductive data analysis* (97-144). London: Academic Press.

Bartlett, J.C., & Searcy. J. (1993). Inversion and configuration of faces. *Cognitive Psychology, 25(3),* 281-316.

Basri, R. (1996). Recognition by prototypes. *International Journal of Computer Vision, 19*, 147-167.

Basri, R., Costa, L., Geiger, D., & Jacobs, D. (1998). Determining the similarity of deformable shapes. *Vision Research, 38*, 2365-2385.

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 509-522.

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

Biederman, I. (1995). Visual object recognition. In S. M. Kosslyn & D. N. Osherson (Eds.), *An Invitation to Cognitive Science: Visual Cognition*, 2nd ed., Vol. 2, pp. 121-165. Cambridge, MA: MIT Press.

Biederman, I., & Cooper, E.E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology, 23*, 393-419.

Biederman, I., & Gerhardstein, P.C. (1993). Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 1162-1182.

Biederman, I., & Gerhardstein, P.C. (1995). Viewpoint-dependent mechanisms in visual object recognition: reply to Tarr and Bülthoff (1995). *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1506-1514.

Biederman, I., & Ju, G. (1988). Surface vs. edge-based determinants of visual recognition. *Cognitive Psychology, 20*, 38-64.

Biederman, I., & Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society of London, B, 352*, 1203-1219.

Biederman, I., Subramaniam, S., Bar, M., Kalocsai, P., & Fiser, J. (1999). Subordinate-level object classification reexamined. *Psychological Research, 62*, 131-153.

Blanz, V., & Vetter, T. (1999). A model for the synthesis of 3D faces", *Computer Graphics Proceedings SIGGRAPH 99*, 187-194.

Bonmassar, G., & Schwartz, E.L. (1998). Representation is space-variant. *Behavioral and Brain Sciences, 21*, 469-470.

Brooks, L.R. (1978). Non-analytic concept formation and memory for instances. In E. Rosch & B.B. Lloyd (Eds.), *Cognition and concepts* (169-211). Hillsdale, N.J.: Erlbaum.

Bülthoff, H.H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences of the United States of America, 89*, 60-64.

Bundesen, C., & Larsen, A. (1975). Visual transformation of size. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 214-220.

Bundesen, C., Larsen, A., & Farrell, J.E. (1981). Mental transformations of size and orientation. In J. Long & A. Baddeley (Eds.), *Attention and Performance, IX* (279-294). Hillsdale, N.J.: Erlbaum.

Cassirer, E. (1944). The concept of group and the theory of perception. *Philosophy and Phenomenological Research, 5*, 1-35.

Cave, C.B., & Kosslyn, S.M. (1989). Varieties of size-specific visual selection. *Journal of Experimental Psychology: General, 118*, 148-164.

Cave, K.R., Pinker, S., Giorgi, L., Thomas, C.E., Heller, L.M., Wolfe, J.M. & Lin, H. (1994). The representation of location in visual images. *Cognitive Psychology, 26*, 1-32.

Chelazzi, L., Duncan, J., Miller, E.K., & Desimone, R. (1998). Responses of neurons in inferior temporal cortex during visual search. *Journal of Neurophysiology, 80*, 2918-2940.

Chelazzi, L., Miller, E.K., Duncan, J., & Desimone, R. (2001). Responses of neurons in macaque area V4 during memory-guided visual search. *Cerebral Cortex, 11*, 761-772.

Chen, L. (1982). Topological structure in visual perception. *Science, 218*, 699-700.

Chen, L. (1985). Topological structure in the perception of apparent motion. *Perception, 14*, 197-208.

Chen, L. (2001). Perceptual organization: To reverse back the inverted (upside-down) question of feature binding. *Visual Cognition, 8*, 287-303.

Cohen, A.L., & Nosofsky, R.M. (2000). An exemplar-retrieval model of speeded same-different judgments. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 1549-1569.

Collishaw, S.M., Hole G.J. (2000). Featural and configurational processes in the recognition of faces of different familiarity. *Perception, 29*, 893-910.

Cortese, J.M., & Dyre, B.P. (1996). Perceptual similarity of shapes generated from fourier descriptors. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 133-143.

Cutting, J.E. (1986). *Perception with an eye for motion*. Cambridge, MA.: MIT Press.

Cutzu, F., & Edelman, S. (1996). Faithful representation of similarities among three-dimensional shapes in human vision. *Proceedings of the National Academy of Sciences, 93*, 12046-12050.

Cutzu, F., & Edelman, S. (1998). Representation of object similarity in human vision: Psychophysics and a computational model. *Vision Research, 38*, 2229-2257.

Davidoff, J., & Donnelly, N. (1990). Object superiority: a comparison of complete and part probes, *Acta Psychologica, 73*, 225–243.

Dill, M., & Edelman, S. (2001). Imperfect invariance to object translation in the discrimination of complex shapes. *Perception, 30*, 707-724.

Dill, M., & Fahle, M. (1998). Limited translation invariance of human visual pattern recognition. *Perception & Psychophysics, 60*, 65-81.

Edelman, S. (1995). Representation of similarity in three-dimensional object discrimination. *Neural Computation, 7*, 408-423.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences, 21*, 449-498.

Edelman, S. (1999). *Representation and recognition in vision*. Cambridge, MA: MIT Press.

Edelman, S., & Bülthoff, H.H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research, 32*, 2385-2400.

Edelman, S., & Duvdevani-Bar, S. (1997). Similarity, connectionism, and the problem of representation in vision. *Neural Computation, 9*, 701-720.

Edelman, S., & Intrator, N. (2000). (Coarse coding of shape fragments) + (retinotopy) ≈ representation of structure. *Spatial Vision, 13*, 255-264.

Edelman, S., & Intrator, N. (2001). A productive, systematic framework for the representation of visual structure. In T.K. Lean, T.G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13* (10-16). Cambridge, MA: MIT Press.

Erickson, M.A., & Kruschke, J.K. (1994). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127*, 107-140.

Estes, W.K. (1986). Array models for category learning. *Cognitive Psychology, 18*, 500-549.

Estes, W.K. (1994). *Classification and cognition.* Oxford: Oxford University Press.

Fabre-Thorpe, M., Richard, G., & Thorpe, S. J. (1998). Rapid categorisation of natural images by rhesus monkeys. *Neuroreport, 9*, 303-308.

Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance, 21, (3)*, 628-634.

Foster, D.H., & Kahn, J.I. (1985). Internal representations and operations in visual comparison of transformed patterns: effects of pattern point-inversion, positional symmetry, and separation. *Biological Cybernetics, 51,* 305-312.

Gentner, D., & Markman, A.B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science, 5*, 152-158.

Gentner, D., & Markman, A.B. (1995). Similarity is like analogy. In C. Cacciari (Ed.), *Similarity in language, thought, and perception* (111-148). Brussels, Belgium: Brepols.

Gibson, J.J. (1950). *The perception of the visual world*. Boston: Houghton Mifflin.

Gibson, B.S., & Peterson, M.A. (1994). Does orientation-independent object recognition precede orientation-dependent recognition? Evidence from a cuing paradigm. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 299-316.

Goldstone, R.L. (1994a). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 3-28.

Goldstone, R.L. (1994b). The role of similarity in categorization: providing a groundwork. *Cognition, 52*, 125-157.

Goldstone, R.L. (1996). Alignment-based nonmonotonicities in similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 988-1001.

Goldstone, R.L., & Medin, D.L. (1994). Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 29-50.

Graf, M. (2001). Analog topological transformations in basic level object recognition [Abstract]. *Journal of Vision, 1(3)*, 98a, http://journalofvision.org/1/3/98, DOI 10.1167/1.3.98.

Graf, M. (2002). *Form, space and object. Geometrical transformations in object recognition and categorization*. Wissenschaftlicher Verlag Berlin, Berlin.

Graf, M., & Schneider, W.X. (2001). Structural descriptions in HIT - a problematic commitment. *Behavioral and Brain Sciences, 24*, 483-484.

Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Gregson, R.A.M. (1998). Metric assumptions are neither necessary nor sufficient to describe similarities. *Behavioral and Brain Sciences, 21*, 473.

Hahn, U., & Chater, N. (1997). Concepts and similarity. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (43-92). Cambridge, MA: MIT Press.

Hahn, U., Chater, N., & Richardson, L.B.C. (in press). Similarity as transformation. *Cognition*.

Hamm, J.P., & McMullen, P.A. (1998). Effects of orientation on the identification of rotated objects depend on level of identity. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 413-426.

Hasselmo, M.E., Rolls, E.T., & Baylis, C.G. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Experimental Brain Research, 32*, 203-218.

Hayward, W.G., & Tarr, M.J. (1997). Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 1511-1521.

Hayward, W.G., & Williams, P. (2000). Viewpoint dependence and object discriminability. *Psychological Science, 11*, 7-12.

Heisele, B., Serre, T., Pontil, M., Vetter, T., and Poggio, T. (2001). Categorisation by learning and combining object parts. *NIPS proceedings*.

Henderson JM, Hollingworth A. (1999). High-level scene perception. *Annual Review of Psychology, 50,* 243-271.

Hill, H., Schyns, P.G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition, 62*, 201-222.

Humphreys, G. W., Riddoch, M. J., & Price, C. J. (1997). Top-down processes in object identification: evidence from experimental psychology, neuropsychology and functional anatomy. *Philosophical Transactions of the Royal Society of London, B, 352,* 1275-1282.

Hummel, J.E. (2000). Where view-based theories of human object recognition break down: the role of structure in human shape perception. In E. Dietrich & A.B. Markman (Eds.), *Cognitive Dynamics: conceptual change in humans and machines* (157-185). Hillsdale, NJ: Erlbaum.

Hummel, J.E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99,* 480-517.

Huttenlocher, D. P., & Ullman, S. (1990). Recognizing solid objects by alignment with an image. *International Journal of Computer Vision, 5*, 195-212.

Ishai, A., Ungerleider, L.G., Martin, A., Haxby, J.V. (2000). The representation of objects in the human occipital and temporal cortex. *Journal of Cognitive Neuroscience, 12*, 35-51.

Jain, A.K., Zhong, Y., & Lakshmanan, S. (1996). Object matching using deformable templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*, 267-278.

Jolicoeur, P. (1985). The time to name disoriented natural objects. *Memory & Cognition, 13*, 289-303.

Jolicoeur, P. (1987). A size-congruence effect in memory for visual shape. *Memory & Cognition, 15*, 531-543.

Jolicoeur, P. (1990). Identification of disoriented objects: a dual-systems theory. *Mind & Language, 5*, 387-410).

Jolicoeur, P. (1992). Orientation congruency effects in visual search. *Canadian Journal of Psychology, 46*, 280-305.

Jolicoeur, P., Corballis, M.C., & Lawson, R. (1998). The influence of perceived rotary motion on the recognition of rotated objects. *Psychonomic Bulletin & Review, 5*, 140-146.

Jolicoeur, P., Gluck, M.A., & Kosslyn, S.M. (1984). Pictures and names: making the connection. *Cognitive Psychology, 16*, 243-275.

Jolicoeur, P., & Humphrey, G.K. (1998). Perception of rotated two-dimensional and three-dimensional objects and visual shapes. In V. Walsh & J. Kulikowski (Eds.), *Perceptual constancy. Why things look as they do* (69-123). Cambridge: Cambridge University Press.

Knappmeyer, Thornton, & Bülthoff (submitted). Interactions between facial form and facial motion during the processing of identity. *Vision Research.*

Kosslyn, S. M. (1994). *Image and Brain. The resolution of the imagery debate*. Cambridge, Massachusetts: MIT Press.

Kourtzi, Z., & Shiffrar, M. (2001). Visual representation of malleable and rigid objects that deform as they rotate. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 335-355.

Kruschke, J.K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Kurbat, M.A. (1994). Structural description theories: is RBC/JIM a general-purpose theory of human entry-level object recognition? *Perception, 23*, 1339-1368.

Labov, W. (1973). The boundaries of words and their meanings. In C.-J.N. Bailey & R.W. Shuy (Eds.), *New ways of analyzing variations in english*. Washington, D.C.: Georgetown University Press.

Lades, M., Vorbrüggen, J.C., Buhmann, J., Lange, J., v.d. Malsburg, C., Würtz, R.P., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers, 42*, 300-311.

Lamberts, K. (1994). Flexible tuning of similarity in exemplar-based categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*, 1003-1021.

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24*, 695-711.

Landau, B. (1994). Object shape, object name, and object kind: representation and development. In D.L. Medin (Ed.), *The Psychology of Learning and Motivation, 31* (253-304). New York: Academic Press.

Landau, B., Smith, L.B., & Jones, S.S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*, 299-321.

Larsen, A., & Bundesen, C. (1978). Size scaling in visual pattern recognition. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 1-20.

Larsen, A., & Bundesen, C. (1998). Effects of spatial separation in visual pattern matching: evidence on the role of mental translation. *Journal of Experimental Psychology: Human Perception & Performance, 24*, 719-731.

Lawson, R. (1999). Achieving visual object constancy across plane rotation and depth rotation. *Acta Psychologica, 102*, 221-245.

Lawson, R., Bülthoff, H.H., & Dumbell, S. (2002). Interactions between view changes and shape changes in picture-picture matching. *Technical Report No. 95, Max Planck Institute for Biological Cybernetics, Tübingen*, Germany. http://www.kyb.tuebingen.mpg.de/publications/pdfs/pdf1609.pdf.

Lawson, R., & Humphreys, G.W. (1996). View-specificity in object processing: Evidence from picture matching. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 395-416.

Lawson, R., & Humphreys, G.W. (1998). View-specific effects of depth rotation and foreshortening on the initial recognition and priming of familiar objects. *Perception & Psychophysics, 60*, 1052-1066.

Lawson, R., Humphreys, G.W., & Jolicoeur, P. (2000). The combined effects of plane disorientation and foreshortening on picture naming: One manipulation are two? *Journal of Experimental Psychology: Human Perception and Performance, 26*, 568-581.

Lee, D.D., & Seung, S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*, 788-791.

Leyton, M. (1992). *Symmetry, causality, mind*. Cambridge, MA.: MIT Press.

Liter, J. C., & Bülthoff, H. H. (1997). View Canonicality Affects Naming But Not Name Verification Of Common Objects. *Technical Report No. 50. Max Planck Institute for Biological Cybernetics, Tübingen, Germany.*

Liu, J, Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. *Nature Neuroscience, 5*, 910-916.

Lowe, D.G. (1985). *Perceptual organization and visual recognition.* Boston, MA: Kluwer.

Lowe, D.G. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence, 31,* 355-395.

Luck, S.J., Chelazzi, L., Hillyard, S.A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology, 77*, 24-42.

Maddox, W.T., Ashby, F.G., Gottlob, L.R. (1998). Response time distributions in multidimensional perceptual categorization. *Perception & Psychophysics, 60*, 620-637.

Mak, Benise S.K., Vera, Alonso H. (1999). The role of motion in children's categorisation of objects. *Cognition, 71,* 10.

Markman, A.B. (2001). Structural alignment, similarity, and the internal structure of category representations. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (109-130). Oxford, Oxford University Press.

Markman, A.B., & Gentner, D. (1993). Splitting the differences: a structural alignment view of similarity. *Journal of Memory and Language, 32*, 517-535.

Markman, A.B., & Gentner, D. (1993b). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*, 431-467.

Markman, A.B., & Gentner, D. (1997). The effects of alignability on memory. *Psychological Science, 8*, 363-367.

Markman, A.B., & Wisniewski, E.J. (1997). Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 54-70.

Marko, H. (1973). Space distortion and decomposition theory. A new approach to pattern recognition by vision. *Kybernetik, 13*, 132-143.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Marr, D., & Nishihara, H.K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society, London, B, 200*, 269-294.

McClelland, J.L., & Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*, 159-188.

McMillan, N.A., & Creelman, C.D. (1992). *Detection theory: A user's guide.* New York: Cambridge University Press.

Medin, D.L. (1989). Concepts and conceptual structure. *American Psychologist, 44*, 1469-1481.

Medin, D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*, 254-278.
Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.
Milliken, B, Jolicoeur, P. (1992). Size effects in visual recognition memory are determined by perceived size. *Memory & Cognition, 20*, 83-95.
Murray, J.E. (1997). Flipping and spinning: Spatial transformation procedures in the identification of rotated natural objects. *Memory & Cognition, 25*, 96-105.
Murray, J.E. (1998). Is entry-level recognition viewpoint invariant or viewpoint dependent? *Psychonomic Bulletin & Review, 5*, 300-304.
Murray, J.E. (1999). Orientation-specific effects in picture matching and naming. *Memory & Cognition, 27*, 878-889.
Newell, F.N. & Bülthoff, H.H. (2002). Categorical perception of familiar objects. *Cognition, 85*, 113-143.
Newell, F.N., & Findlay, J.M. (1997). The effect of depth rotation on object identification. *Perception, 26*, 1231-1257.
Niall, K.K. (1992). Projective invariance and the kinetic depth effect. *Acta Psychologica, 81*, 127-168.
Niall, K.K., & Macnamara, J. (1990). Projective invariance and picture perception. *Perception, 19*, 637-660.
Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization-relationship. *Journal of Experimental Psychology: General, 115*, 39-57.
Nosofsky, R.M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 14*, 54-65.
Nosofsky, R.M., & Palmeri, T.J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266-300.
Op de Beeck, H., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience, 4*, 1244-1252.
O'Toole, A.J., Roark, D.A., & Abdi, H. (2002). Recognizing moving faces: a psychological and neural synthesis. *Trends in Cognitive Sciences, 6(6)*, 261-266.
Palmer, S.E., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and Performance IX* (135-151). Hillsdale, N.J.: Erlbaum.
Perret, D.I., Hietanen, J.K., Oram, M.W., & Benson, P.J. (1992). Organization and functions of cells in the macaque temporal cortex. *Philosophical Transactions of the Royal Society of London, B, 335*, 23-50.
Perret, D.I., Mistlin, A.J., & Chitty, A.J. (1987). Visual neurones responsive to faces. *Trends in Neuroscience, 10*, 358-364.
Perret, D.I., & Oram, M.W. (1993). The neurophysiology of shape processing. *Image and visual computing, 11*, 317-333.
Perrett, D.I., & Oram, W.M. (1998). Visual recognition based on temporal cortex cells: viewer-centred processing of pattern configurations. *Zeitschrift für Naturforschung, C, 53*, 518-541.

Perrett, D.I., Oram, W.M., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalization of recognition without mental transformations. *Cognition, 67*, 111-145.

Perret, D.I., Rolls, E.T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research, 47*, 329-342.

Pitts, W., & McCulloch, W.S. (1947). How we know universals: The perception of auditory and visual forms. *Bulletin of Mathematical Biophysics, 9*, 127-147.

Poggio, T. (1990). A theory of how the brain might work. *The Brain: Cold Spring Harbour Symposia on Quantitative Biology*, N.Y.: CSH Laboratory Press, 899-910.

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

Poggio, T., & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science, 247*, 978-982.

Rakover, S. S. (2002). Featural vs. configurational information in faces: A conceptual and empirical analysis. *British Journal of Psychology, 93*, 1-30.

Reed, S.K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382-407.

Rhodes, G., Brake, S., & Atkinson, A.P. (1993). What's lost in inverted faces? *Cognition, 47*, 25-57.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*, 1019-1025.

Riesenhuber, M., & Poggio, T. (2000). CBF: A new framework for object categorization in cortex. In S.-W. Lee & H.H. Bülthoff (Eds.), *Biologically motivated computer vision*. Berlin: Springer.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology, 12*, 162-168.

Rosch, E., & Mervis, C.B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382-439.

Schwaninger, A., Carbon, C.C., & Leder, H. (in press). Expert face processing: Specialisation and constraints. In Schwarzer, G. & Leder, H. (Eds.), *The Development of Face Processing.*

Schwaninger, A., Cunningham, D.W., & Kleiner, M. (2002). Moving the Thatcher illusion. *Poster presented at the 10th Annual Workshop on Object Perception and Memory, Kansas City*, November 21, 2002.

Schwaninger, A., Lobmaier, J.S., & Collishaw, S.M. (2002). Role of Featural and Configural Information in Familiar and Unfamiliar Face Recognition. *Lecture Notes in Computer Science, 2525*, 643-650.

Schwaninger A., Lobmaier, J., & Collishaw, S, M. (2002). Role and interaction of featural and configural processing in face recognition. *Vision Sciences Society, 2nd annual meeting, Sarasota*, Florida, May 10-15, 2002.

Schyns, P.G., Goldstone, R.L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21*, 1-54.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(3)*, 681-696l.

Selfridge, O.G., & Neisser, U. (1963). Pattern recognition by machine. In E.A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Sergent J. (1985). Influence of task and input factors on hemispheric involvement in face processing. *Journal of Experimental Psychology: Human Perception and Performance, 11(6)*, 846-61.

Shaw, R., & Pittenger, J. (1977). Perceiving the face of change in changing faces: implications for a theory of object perception. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing. Toward an ecological psychology.* Hillsdale, N.J.: Erlbaum.

Shepard, R.N. (1957). Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. *Psychometrika, 22,* 325-345.

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

Shepard, R.N., & Cermak, G.W. (1973). Perceptual-cognitive explorations of a toroidal set of free-form stimuli. *Cognitive Psychology, 4*, 351-377.

Smith, E.E., & Medin, D.L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smith, E.E., Shoben, E.J., & Rips, L.J. (1974). Structure and process in semantic memory: a featural model for semantic decisions. *Psychological Review, 81*, 214-241.

Stone, J. V. (1998). Object recognition using spatio-temporal signatures. *Vision Research, 38*, 947-951.

Stone, J. V. (1999). Object Recognition: View-Specificity and Motion-Specificity, *Vision Research, 39*, 4032-4044.

Sutherland, N.S. (1968). Outlines of a theory of visual pattern recognition in animals and man. *Proceedings of the Royal Society, London, B, 171*, 297-317.

Tanaka, J.W. (2001). The entry point of face recognition: evidence for face expertise. *Journal of Experimental Psychology: General, 130*, 534-543.

Tanaka J. W. & Farah, M. J. (1991). Second-order relational properties and the inversion-effect: Testing a theory of face perception. *Perception & Psychophysics, 50*, 367-372.

Tanaka, J. W. & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology, 79*, 471-491.

Tanaka, J., Luu, P., Weisbrod, M., & Kiefer, M. (1999). Tracking the time-course of object categorization using event-related potentials. *NeuroReport, 10*, 829-835.

Tarr, M.J. (1995). Rotating objects to recognize them: a case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review, 2*, 55-82.

Tarr, M.J. (1999). News on views: Pandemonium revisited. *Nature Neuroscience, 2*, 932-935.

Tarr, M.J., & Bülthoff, H.H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1494-1505.

Tarr, M.J., & Bülthoff, H.H. (1998). Image-based object recognition in man, monkey and machine. In M.J. Tarr & H.H. Bülthoff (Eds.), *Object recognition in man, monkey, and machine* (1-20). Cambridge, MA: MIT Press.

Tarr, M. J., Bülthoff, H. H., Zabinski, M., & Blanz, V. (1997). To what extent do unique parts influence recognition across changes in viewpoint? *Psychological Science, 8*, 282-289.

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

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience, 1*, 275-277.

Thompson, D'A.W. (1917). *On growth and form*. Second edition, 1942: Cambridge: Cambridge University Press.

Thompson, P. (1980). Margaret Thatcher -- A new illusion. *Perception, 9*, 483-484.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.

Todd, J.T, Chen, L., & Norman, J.F. (1998). On the relative salience of Euclidean, affine, and topological structure for 3-D form discrimination. *Perception, 27*, 273-282.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General, 113*, 169-193.

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

Ullman, S. (1995). Sequence seeking and counter streams: A computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex, 5*, 1-11.

Ullman, S. (1996). *High-level vision. Object recognition and visual cognition*. Cambridge, MA: MIT Press.

Ullman, S., & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*, 992-1006.

Ullman, S., & Sali, E. (2000). Object Classification Using a Fragment-Based Representation. BMCV 2000. *Lecture Notes in Computer Science, 1811*, 73-87.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*, 682-687.

Valentine, T. (1988). Upside-down faces: a review of the effect of inversion upon face recognition. *British Journal of Psychology, 79*, 471-491.

Van Gool, L.J., Moons, T., Pauwels, E., & Wagemans, J. (1994). Invariance from the Euclidean geometer's perspective. *Perception, 23*, 547-561.

Van Leeuwen, C. (1998). Regular spaces versus computing with chaos. *Behavioral and Brain Sciences, 21*, 482-483.

Van Rullen, R., & Thorpe, S. J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects, *Perception, 30*, 655-668.

Wachsmuth, E., Oram, M.W., & Perret, D.I. (1994). Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque. *Cerebral Cortex, 4*, 509-522.

Wagemans, J., Van Gool, L., & Lamote, C. (1996). The visual systems measurement of invariants need not itself be invariant. *Psychological Science, 7*, 232-236.

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

Wallis, G. M., Bülthoff H.H. (2001). Effect of temporal association on recognition memory. *Proceedings of the National Academy of Science USA, 98*, 4800-4804.

Wallis, G., & Bülthoff, H.H. (2002). A brief introduction to cortical representations of objects. *Technical Report No. 97, Max Planck Institute for Biological Cybernetics, Tübingen, Germany*. http://www.kyb.tuebingen.mpg.de/publications/pdfs/pdf1646.pdf.

Wallraven, C. & Bülthoff, H.H. (2001). Automatic acquisition of exemplar-based representations for recognition from image sequences. *CVPR 2001 - Workshop on Models vs. Exemplars*

Wallraven, C., Schwaninger, A., Schuhmacher, S., & Bülthoff, H.H. (2002). View-Based Recognition of Faces in Man and Machine: Re-visiting Inter-Extra-Ortho. *Lecture Notes in Computer Science, 2525*, 651-660.

Wang, G., Tanifuji, M., & Tanaka, K. (1998). Funcitional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neuroscience Research, 32*, 33-46.

Wiskott, L., Fellous, J.M., Krüger, N., & von der Malsburg, C. (1997). Face Recognition by Elastic Bunch Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7)*, 775-779.

Yamane, S., Kaji, S., & Kawano, K. (1988). What facial features activate face neurons in inferotemporal cortex of the monkey. *Experimental Brain Research, 73*, 209-214.

Yuille, A.L. (1991). Deformable templates for face recognition. *Journal of Cognitive Neuroscience, 3*, 59-70.

Yuille, A.L., Ferraro, M., & Zhang, T. (1998). Image warping for shape recovery and recognition. *Computer Vision and Image Understanding, 72*, 351-359.

Zaki, S. R. & Homa, D. (1999). Concepts and transformational knowledge. *Cognitive Psychology, 39*, 69-115.