Aus dem Institut für Medizinische Psychologie

der Ludwig-Maximilians-Universität München

Vorstand: Prof. Dr. Ernst Pöppel

Relational Strategies for the Study of Visual Object Recognition

Dissertation
zum Erwerb des Doktorgrades der Humanbiologie
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität zu München

vorgelegt von

Erol Osman

aus

München

Jahr

2008

Mit Genehmigung der Medizinischen Fakultät der Universität München

Berichterstatter:	Prof. Dr. Ingo Rentschler
Mitberichterstatter:	Prof. Dr. Peter Grafe Priv. Doz. Dr. Oliver Ehrt
Mitbetreuung durch den promovierten Mitarbeiter:	Dr. Martin Jüttner
Dekan:	Prof. Dr. med. D. Reinhardt
Tag der mündlichen Prüfung:	24. Juli 2008

Contents

List of Figures

List of Tables

Acknowledgements

I would like to use the opportunity to thank some of the persons who contributed to this thesis. First of all I would like to thank my supervisors Martin Jüttner and Ingo Rentschler. Without their support on every stage of this thesis and without their unending patience I wouldn't have succeeded. Terry and his wife Kate received me in Australia with great hospitality, for which I shall always be grateful. Terry is just sparkling with ideas and you never knew where a discussion with him could lead to. As someone put it: '15 minutes of talk with Terry mean 3 months of work'. I would like to thank "The Nerds" from Terry's lab, Tom Drummond, Craig Dillon, Andy McCabe, David Squire, Gary Briscoe, Adrian Pearce, late arrival Stuart Campbell, and their respective partners. They helped me to see what an adventure and how much fun science can be. Adrian introduced me to the mysteries of CLARET. I hope he likes what I've done with his algorithm.

Here in Munich, I was confined to "the dungeons" of our lab. Whenever I ascended to the "higher plains" of the first floor, it was nice to find persons with whom to discuss problems or just chat. I would like to thank Bernhard Treutwein, Christoph Zetzsche, Markus Gschwind and Alexander Müller for that. I've shared some sweat, in the sauna and in the lab with Volker Baier and Andreas Eisenkolb, who have become dear friends to me. Andreas was one of the persons at the lab pressing me most relentlessly to end my thesis.

I would also like to thank the "Bosch Stiftung" and the "Graduiertenkolleg für sensorische Interaktion in biologischen und technischen Systemen", where I received scholarships. The secretary of the latter, Isolde von Bülow, deserves special mentioning. I could always rely on her help and advice, whether it concerned funding hardware or discussing career options.

Finally, I would like to thank my partner and my family, who endured me and never lost faith, even when I stopped responding to questions concerning deadlines.

Chapter 1

Introduction

Aristotle saw the world composed of a distinct number of objects provided with two types of attributes. One type of attribute was accidental and fleeting, the other type of attribute was fixed or slowly changing. The peculiarity of objects was made up by some attributes, whereas other attributes determined the categories to which the objects belonged. Therefore, the possession of a common set of attributes was an invariant property over the objects within one category. That is, in the Aristotelian view objects were assigned to categories according to attributes they have in common with other occurrences (see Russell, 1962; Watanabe, 1985, chap. 2).

Ludwig Wittgenstein was the first to point out that a category like *game* has neither clear boundaries nor is it defined by a set of common properties, thus challenging the 2000 year old Aristotelian concept of categories (see Glock, 1996). He wondered what a game of darts, for instance, might have in common with the game of soccer. This led Wittgenstein to conceive of the concept of "family resemblance", according to which attributes are distributed across the members of a family, or category, in a probabilistic fashion. Thus, games, tables, and trees are natural families, each constituted by a criss-cross network of overlapping resemblances. In particular, categories of natural languages are characterised by family resemblance (see Lakoff, 1987).

In accord with the Aristotelian view of categorisation, strategies of pattern and object recognition by machines typically relied on the representation of patterns and object views (images) as vectors of characteristic features, or attributes. Such vectors may establish representational uniqueness of views of objects from different categories. If so, object recognition, or classification, is achieved by partitioning feature space into regions associated with different object classes. This approach, of which neural networks (e.g. Haykin, 1994) are prominent examples, worked well for simple and complete objects

occurring in isolation but not for complex and maybe only partially visible objects embedded in scenes. To solve the latter type of problems, strategies of recognition-by-parts were developed in the domain of machine intelligence. Here, patterns or views are decomposed into constituent parts and represented in terms of attributes of parts and part relations, thus generating relational structures that define their correspondents uniquely. Classification is then achieved using relational graph matching, where similarity functions between sample structures and model structures are maximised through learning. In brief, formal strategies based on the notion of family resemblance enable the recognition of objects or structures composed of parts in complex scenes as well as image understanding in general (see Bischof and Caelli, 1997a; Caelli and Bischof, 1997b).

In the medical domain, strategies of relational or structural pattern and object recognition are of interest for a number of reasons. First, the question concerning the strategies underlying human object recognition is generally unresolved (e.g. Osaka et al., 2007). Second, disturbances of the visual sense of form recognition, object recognition, and scene understanding are frequently encountered as a consequence of brain pathology (e.g. Grüsser and Landis, 1991). Third, imaging technologies are of ever increasing importance for medical diagnosis, but theories of image understanding by humans are few and typically restricted to a type of schematic drawings that have little resemblance with, say, photographs made by X-rays (e.g. Biederman, 1987). For these reasons, the present dissertation explores the potential that computer-aided relational strategies have for studying visual object recognition and image understanding by humans.

Chapter 2

Neuroanatomy and Neurophysiology of Visual Object Recognition

Neuroanatomical and neurophysiological data is mainly based on the examination of mammals, especially of monkeys. Modern imaging techniques, such as functional magnetic resonance imaging (fMRI), have only recently allowed imaging of human brain functions with sufficient resolution in time and space. Other techniques are PET, evoked potentials, and transcranial magnetic stimulation.

2.1 Vision as Analysis

2.1.1 Precortical Visual Processing

The first step of processing visual stimuli already takes place in the retina. The retina is a part of the brain, which has been secluded early in development, but has kept its connections in a bundle of fibres – the optic nerve. The retina consists of several layers, three neural layers and two separating layers containing synapses.

At the rear of the retina one finds the light receptors, the rods and cones. Three types of cones, which are sensitive to different wavelengths of light, provide the basis for colour vision. Cones, however, do not function well in dim light. Rods are responsible for vision in dim light, but in turn they do not contribute to colour vision. Rods and cones are not distributed equally over the retina. Cones are densely packed within a central region of the retina, the fovea, whereas the majority of receptors in the periphery are rods.

The light receptors are connected to the bipolar cells, which in turn are connected to the retinal ganglion cells. The ganglion cells determine the "output" of the eye so providing the connection between eye and brain. Bipolar and ganglion cells already possess a very important property of neurons in the visual system – a receptive field. Loosely speaking, the receptive field defines the type of stimuli and the area of presentation in the visual field, which evoke a response of a neuron. It is a powerful but shorthand description of the behaviour of a neural cell. The two main types of receptive fields of retinal ganglion cells are on-centre and off-centre. They are circular with a centre and an outer ring. On-centre cells respond to a bright central spot with a dark surrounding, whereas off-centre cells respond to a dark centre with bright surround. The size of receptive fields varies systematically being smallest in the centre of the visual field, within the fovea, and increasing in size as the retinal eccentricity grows.

The optic nerve connects the retina with subsequent stages of the visual brain. In the optic chiasm half of the fibres of each eye cross over. The fibres from the medial area of the retina lead to the contralateral hemisphere of the brain, the fibres from the lateral area of the retina lead to the ipsilateral side. This implies that the information from either the left or the right half of the visual field is transmitted to the contralateral hemisphere of the brain, a fact which is exploited in experiments on cerebral lateralisation.

There are two major pathways, the retino-collicular pathway containing about 10% of the afferent axons and the retino-geniculate pathway comprised of the remaining 90%. The retino-collicular path leads via the superior colliculus to the pulvinar, which in turn is connected to many cortical areas. Its function is mainly the control of eye movements and attention. The retino-geniculate path leads via the lateral geniculate nucleus (LGN) to the primary visual cortex. It is associated with visual pattern- and movement-analysis.

Already in the precortical processing of visual information four main principles become obvious. (1) There is clearly a hierarchical processing, starting in the retina and leading to the cortex. (2) There is a division in several processing streams – besides the aforementioned split into two major pathways – filtering and structuring the visual information in different ways. This starts with different types of ganglion cells – alpha, beta and gamma (Boycott and Wässle, 1974) – and continues with the magno- and parvo-cells in LGN. (3) The retinal ganglion cells and the cells in LGN are organised retinotopically – neighbouring neurons correspond to neighbouring locations on the retina and, therefore, in the visual field. (4) The majority of connections between LGN and primary visual cortex are not afferent but on the contrary corticofugal.

2.1.2 Primary Visual Cortex

The main pathway for the cortical processing of visual information leads via the LGN to the primary visual cortex, which is also called striate cortex and corresponds to Brodmann area 17, also called V1.

In V1 the processing of visual information is continued. The neurons are mainly divided into three different classes (Hubel and Wiesel, 1962, 1968):

Simple cells respond best to lines of a certain location and orientation.

Complex cells respond best to lines moving in a certain direction. They are sensitive to orientation of lines and edges, but they ignore the spatial position of such stimuli.

Hypercomplex or endstopped cells respond best to line segments of a certain length, which move in a specific direction.

It should also be noted that there is an increasing number of reports in the literature about cells in V1 displaying a far more complex behaviour – thus giving rise to the concept of "extraclassical receptive fields". (see p. 23).

Cells in V1 again show a retinotopic organisation and furthermore a columnar functional organisation, which extends vertically to the cortical surface. Orientation sensitive neurons of similar orientation selectivity are organised in vertical columns. These columns are organised as bands extending tangentially to the cortical surface. The orientation selectivity of neurons changes continually, as one moves perpendicularly to these bands. A complete orientation sequence of 180° forms a hypercolumn, which is about 1mm in size. Further cytoarchitectural structure is added by the presence of columns of ocular dominance. Finally, V1 can be divided into blobs, assuming a high level of cytochrome oxidase in appropriate staining, and interblob regions. Blob neurons are predominantly monocular, sensitive to colour, but not to orientation. Interblob neurons, on the other hand, are mostly binocular, sensitive to orientation, but not to colour. This suggests that a single area can be heterogeneous regarding its function.

V1 is the first stage in the hierarchy of cortical visual processing, projecting to all other occipital areas. Lesions in V1 tend to cause a total loss of conscious vision, although patients may still react to visual stimuli (blindsight). There are three major projections for further visual analysis: (1) The blob regions project to V4, which can be viewed as a colour (and form) processing area. (2) The regions with magnocellular input project to V2 and to V5 – the latter area being mainly concerned with motion processing. (3) Finally, V1 projects to V3, which is considered to be a form processing area.

The next stage in the hierarchy of visual processing is V2 (Brodmann area 18), also projecting to all other occipital areas. Anatomically V2 also shows regions with different concentrations of cytochrome oxidase – now appearing in stripes. Thick stripes contain neurons, which are largely selective to orientation, movement and disparity. In thin stripes one mostly finds neurons sensitive to colour information. Finally, the interstripe regions contain end-stopping cells, spot-cells, sensitive to dimension and wavelength, and also cells reacting to illusory contours (von der Heydt et al., 1984).

Following V1, projections to extrastriate areas split into a multitude of different pathways, leading to about 30 visually engaged centres for further processing (Desimone and Ungerleider, 1989; Felleman and van Essen, 1991). Although these paths are mostly hierarchically organised, there are many cross-connections and back-projections to "lower" areas.

2.1.3 Inferior Temporal Cortex (IT)

Based on the work of Ungerleider and Mishkin (1982) two major pathways of visual processing have been identified beyond V1 and V2: (1) The dorsal stream, leading via V3, MT (middle temporal) and MST (medial superior temporal), and VIP (ventral inferior parietal) to area 7a in the parietal cortex. This system is mainly concerned with the location of visual information in space – the "where pathway". (2) The ventral stream leads via V4 to regions in the inferotemporal cortex – TEO and TE. It is largely concerned with the recognition of objects – the "what pathway".

The inferotemporal cortex is a visual area known to be essential for object vision. Patients with lobectomies in the temporal lobe usually show no signs of loss of visual abilities such as contrast sensitivity, acuity or colour discrimination. Nevertheless they develop deficits in visual perception (Milner, 1958, 1968), among them impairment in recognising objects, i.e. object agnosia (Farah, 1990).

The posterior part of IT, called TEO, receives its input primarily from V4, but also directly from V2 and V3. Feedback connections to TEO exist, among others, from TE, from the parahippocampal area TH, and the perirhinal cortex. TEO itself projects largely into area TE; back projections to V2, V3, V4 exist, and TEO is interconnected with a number of areas of the dorsal stream. Input to TE comes primarily from TEO, but also from V4.

The inferotemporal cortex has a large number of interconnections. Four major cortical connections, most of them reciprocal, can be identified: (1) Projections from the primary visual areas in the occipital lobe, leading via TEO to TE. They can be identified with the stream of visual information down the processing hierarchy. Neurons tend to react only to increasingly

complex stimuli, they display invariance to position, size, and other types of invariance become more common and more pronounced. (2) Projections to the multimodal areas of STS. Tanaka (2000) calls TE in the monkey the "final purely visual stage" of the ventral pathway. (3) Projections to the medial temporal region, including amygdala and hippocampus. They are important for affective associations and long-term memory functions. The anatomical connection between IT and memory structures indicates the importance of IT for object recognition. (4) Projections to the prefrontal cortex. They are important for solving visual tasks accessing short-term memory – delayed matching to sample for instance. Neurons in the prefrontal cortex, the association cortex of the frontal lobe, react to stimuli depending on their relevance for behaviour.

Significant contribution to the knowledge about IT was achieved by the single-cell studies of Tanaka and co-workers on anaesthetised monkeys (Tanaka, 2000). They found that cells in TEO responded to simple stimuli, for instance geometrical figures of certain size, orientation or colour, or textures (Tanaka, 1996). TEO is characterised by a mixture of cells with various levels of selectivity to moderately complex stimuli. Neurons in TE are rather activated by complex stimuli, i.e. stimuli possessing a spatially complex structure. After isolating spike activities from a single cell, a set of plant and animal models was presented, to find the optimal stimulus. In a heuristic manner the optimal stimuli were simplified step by step to determine which feature or combination of features was essential for maximal activation (Fig. 2.1). Tanaka and co-workers found that the critical features were in general of moderate complexity, being not sufficient to code complete objects. Exceptions were the discovery of cells reacting specifically to faces and hands. Tanaka (2000) claims that invariance in object recognition can in part be explained by the invariant properties of single cell responses in TE. The receptive fields, with sizes from 10° to 30°, are larger than in earlier stages allowing for positional invariance. Other types of invariance found in a varying number of TE-cells are stimulus size, contrast polarity, and aspect ratio.

Using single cell recording and optical imaging a columnar organisation was found in TE with a horizontal extent of about 0.5mm and a vertical extent from layers 2 to 6 (Fujita et al., 1992). Neighbouring cells respond to similar stimuli. It was found that columns with similar properties can overlap. Sometimes columns can map transformations of stimuli such as rotation. This was most clearly observed for faces (Fig. 2.2).

These observations provide remarkable insight into the functional characteristics of visual object recognition, but it has to be kept in mind that the selection of optimal stimuli and their reduction as well as the definition

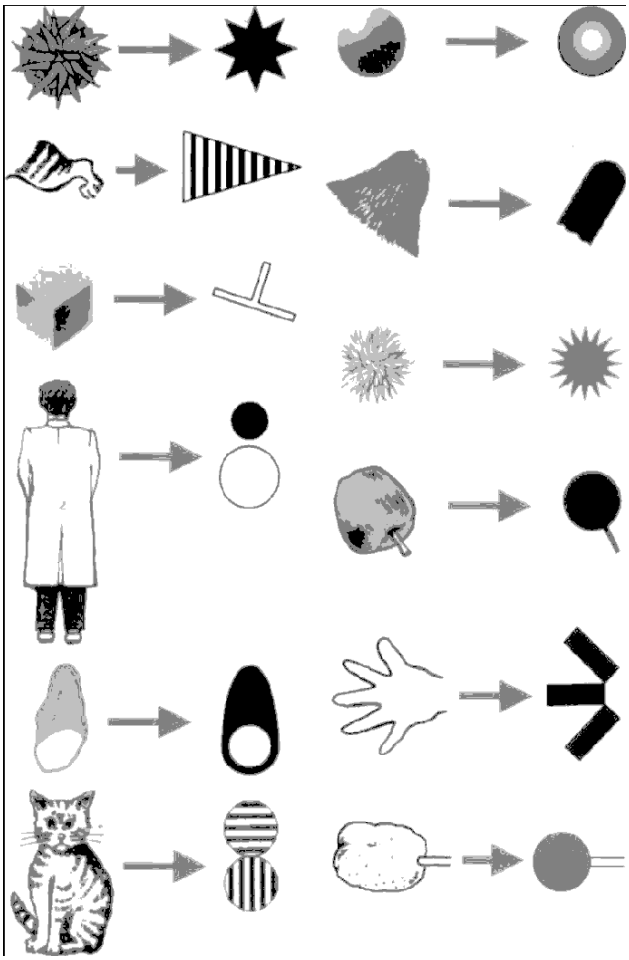


Figure 2.1: 12 Examples for the reduction process applied to stimuli to determine the critical feature for the activation of individual TE cells in monkeys. The images to the left of the arrows represent the images of the most effective object stimuli, the corresponding reduced stimuli are shown right of the arrows. Further reduction reduced the activity of the measured cells. (From Tanaka, 2000, p. 149)

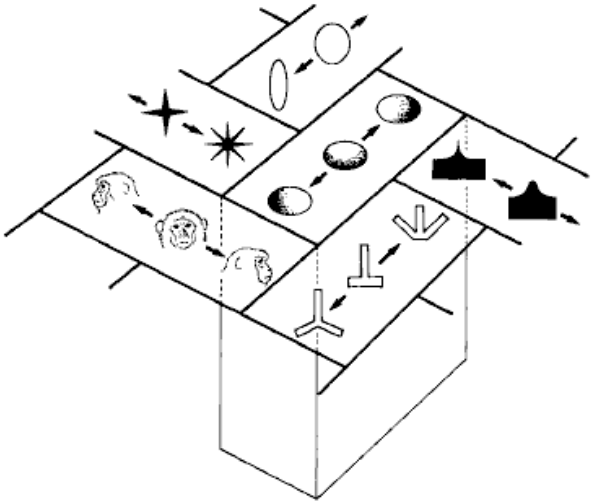


Figure 2.2: A model of columnar organisation in TE for non-face stimuli (from Tanaka, 2000, p. 154). Multiple columns representing different but related features could be partially overlapping, the borders not necessarily being distinctive.

of similarity for stimuli has no formal basis. It is done by using heuristic methodologies.

Tanaka (2000) further found that the selectivity of cells in TE in adult animals can be changed by learning. Training monkeys to solve a categorisation task enhanced the representation in IT of task relevant features relative to irrelevant features (Sigala and Logothetis, 2002).

Other researchers have found similar results. After extensively training monkeys with views of artificial 3D objects in several thousand trials over several months, recording from the anterior IT revealed a number of cells that were highly selective to familiar views of those objects (Logothetis et al., 1995; Logothetis, 2000). Responses decreased when the objects were rotated away from the learned viewpoint. Only a minority of the cells showed viewpoint-invariant object-specific responses. Since the objects used – wire- and amoeba-like objects – were novel for the animals, these results indicate a high degree of plasticity in IT. In contrast to the results obtained by Tanaka and co-workers, the results from trained monkeys showed similarities between face-specific cells and neurons selective for wire- or amoeba-like objects. They both showed selectivity to complex configurations, which cannot be further reduced, at the same time being invariant to changes in position and size (Logothetis et al., 1995; Logothetis and Pauls, 1995). It is unclear

whether these differences to Tanaka's results come from using awake animals, from using different recording locations, or from the type of stimuli used.

An approach different to those used by Tanaka (2000) and Logothetis (2000) was used by Booth and Rolls (1998). Their monkeys were allowed to familiarise themselves with a set of real world objects for a certain amount of time. Thereafter single cells in IT were recorded, while the monkeys performed a fixation task. Booth and Rolls (1998) found both neurons selective to features or single views of objects, but also a significant number of object-selective (view-independent) cells. The latter did not form a separate population, but were intermingled with view-dependent cells, supporting the hypothesis, that their responses are built by associating the outputs of several feature- and view-selective cells. These object selective cells are not to be identified as "grandmother cells" (Barlow, 1972), since they do not respond to a single object exclusively, but to views of one or more objects, or also to single views of other objects. The information-theoretic analysis performed by Booth and Rolls (1998) suggests that the coding of objects, as well as faces, in IT uses a sparse, distributed representation, where a great deal of information is conveyed by the firing rates of the neurons.

A central stage in the ventral visual pathway in humans is the "lateral occipital complex" (LOC), an occipito-temporal region. It is seen as a putative homologue in humans to monkey IT and can be loosely defined as the cortical region responding more strongly to views of objects and object fragments than to textures or scrambled objects (for a review see Grill-Spector et al. (2001)). LOC is a largely non-retinotopic area, activated by both the contralateral and ipsilateral visual fields. Recent studies using fMRI show, that regions in LOC are to a certain extent invariant to changes in stimulus size, position, and contour cues (Grill-Spector et al., 2001; Vuilleumier et al., 2002). It is not clear whether there exist several modules, which are specialised for different categories. A region has been identified, reacting selectively to faces, the fusiform face area (FFA) (Puce et al., 1995), but there is also evidence that activation in FFA might just reflect a high level of expertise for a certain category (Gauthier et al., 2000). There is further evidence, that object representations are distributed and overlapping, since regions responding maximally to a certain category also respond significantly to views from other categories (Ishai et al., 1999). Moore and Engel (2001) have found that activity in LOC increased with the presentation of 3D volumes relative to 2D shapes. The increase also occurred when 2D shapes were perceived as volumetric objects. Further, data exists linking activity in LOC with behavioural performance in various recognition tasks (Avidan-Carmel et al., 2000; Grill-Spector et al., 2000; Duhoux et al., 2005). LOC certainly plays an important role in object recognition, although the fact that

both novel and known objects activate LOC, suggests this cortical area to be a kind of general system for perceiving the shape of objects (Grill-Spector et al., 2001).

2.1.4 Prefrontal Cortex (PFC)

The frontal lobe can be seen as the target region for the spatial and object recognition pathways originating in the occipital lobe. Its function is the selection of behaviour depending on context and internal knowledge (for an overview see Kolb and Whishaw, 1996).

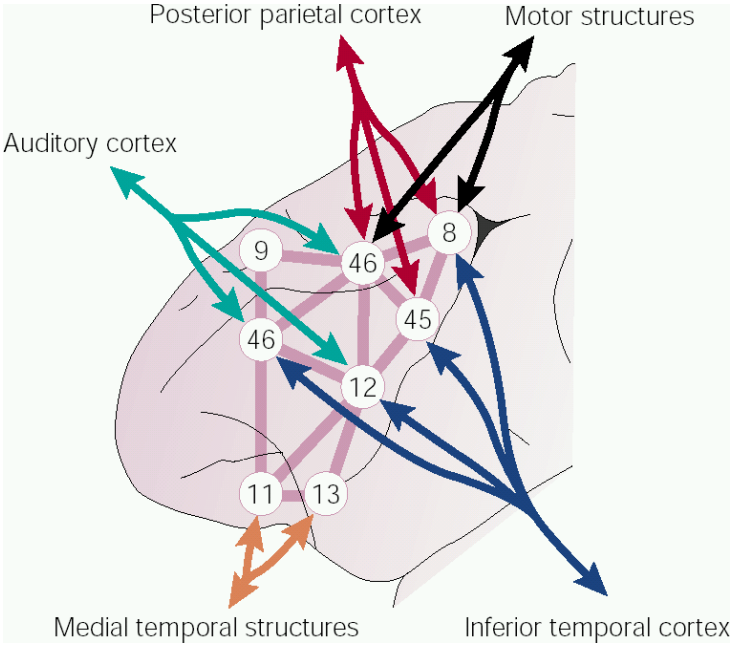


Figure 2.3: Integrative anatomy of the macaque monkey prefrontal cortex (From Miller, 2000). Numbers refer to Brodmann areas.

The PFC, the association cortex of the frontal lobe, is connected with most of the neocortical sensory systems, motor systems, and a wide variety of subcortical structures (see Fig.2.3). It thus integrates sensory information from different modalities with information about the internal environment – arousal, drives, and motives – as well as motivational significance and emotional content of sensory input. This puts the PFC in an ideal position to control the cognitive processes resulting in the correct action being taken at the right time and place (for reviews see Fuster, 2001; Miller, 2000). Projections back to these areas allow the PFC to exert a "top-down" influence

on a range of brain processes. In monkeys for instance, where the posterior part of the corpus callosum was severed, leaving only the anterior connection between the cortex hemispheres, an activation of IT neurons by a top-down signal from PFC was found (Tomita et al., 1999). Neurons in PFC are able to sustain their activity for several seconds, forming associations between events separated in time, and ignoring potentially distracting intervening events. This active performance distinguishes working memory from simple short-term buffering. In contrast, sustained activity for instance in IT is more easily disrupted by distractors (Miller et al., 1996).

Neuropsychology of frontal lesions shows how the loss of social context integration with ensuing inappropriate behaviour is being brought about (Fuster, 2001).

2.1.5 Haptic and Visual: Multimodal Processing

Information processing in early sensory areas is believed to be strictly modality specific. Also LOC was assumed to be an area dedicated to purely visual processing, where different processing streams converge to infer object shape – or rather volume[1] – in a manner partly invariant to size, position and contour cues (Grill-Spector et al., 2001). Recent fMRI studies have shown that sensing objects in the haptic modality activates areas abutting and overlapping regions of LOC (Amedi et al., 2001; James et al., 2002). Amedi et al. (2001) compared the differential activation of areas by viewing objects and textures[2] with the difference in activation between touching objects and haptic textures. They found somatosensory activation in the occipito-temporal region, with tactile objects creating greater activation than tactile textures. Most of the voxels overlapped LOC, the others were abutting. In a further experiment it could be shown that this crossmodal overlapping is not general, but specific in several ways: (1) The voxels activated by tactile and visual stimuli tend to prefer graspable objects over faces or houses and (2) auditory stimuli – meaningless and object-related – do not elicit a substantial activation in these areas (Amedi et al., 2002). A further fMRI study investigated crossmodal haptic and visual interaction using novel clay objects (James et al., 2002). Haptic exploration produced an activation in known somatosensory regions, but also in occipital regions. Activation during visual and haptic exploration overlapped in the middle occipital area of LOC. Viewing visually and haptically primed objects produced stronger activation in the middle occipital and lateral occipital regions of LOC. The authors of

[1] findings of priming experiments support the hypothesis of the integration of volume information in LOC (Moore and Engel, 2001)

[2] This is a common method to locate LOC in fMRI studies

these studies assume the exploitation of visual object related representation systems for haptic object perception

2.2 Vision as Inference

So far, visual processing has been considered largely to be an analytical process. Information enters the cortex – largely via LGN and V1 – where it is split into specialised streams, each organised in an hierarchical manner. Neurons do increasingly more complex filter operations as one proceeds from one cortical centre to the next within the hierarchy. The neurons themselves are characterised by their receptive fields, the properties of which solely depend on how the input of neurons from lower stages is combined. The classical example for this view is the construction of orientation selective simple cell properties from the arrangement of radial symmetric neurons in LGN by Hubel and Wiesel (Hubel and Wiesel, 1977; Hubel, 1995). This view is also confirmed by properties of IT neurons, which – being at a comparatively high stage in the hierarchy – typically possess receptive fields which are large and specific to complex stimuli. These properties are assumed to result from the convergence of neurons from V4 and TEO (Tanaka, 1996).

The visual system is thus considered to be a feedforward network, where more and more complex properties of the input visual information are extracted, until finally, for instance, the identity and pose of a seen object can be handed to decision making centres, such as the prefrontal cortex. This is a functional interpretation of an arrangement of cortical centres in the visual system, defined by their connectivities (van Essen et al., 1992). Within each centre of processing within the hierarchy, there are lateral connections, tempering the purely bottom-up defined properties of receptive fields. Furthermore the effects of attention call for a certain amount of top-down control, which is not so much qualitative as quantitative in nature, modulating the activity of neurons down to V1.

This interpretation of the brain enacting feed-forward analysis largely ignores, that feedforward connections are only a small part of the overall connections. In the V1 layer receiving input from the geniculate only 5% of the total excitatory synapses on the average neuron originate from LGN (Peters and Payne (1993) as cited by Young (2000)). The others are connections from within V1 and from other subcortical and cortical centres, which are necessarily higher up in the processing hierarchy. The fact that so many resources are given to cross- and feedback-connections within a – presumably – highly optimised visual system, warrants the assumption, that these connections have more functions than merely modulating neurons depending on

their immediate surroundings (lateral connections within a cortical centre) and the given attention (feedback from higher centres).

Several authors have proposed the idea, that vision might be largely a process of inference (Young, 2000; Barlow, 2001; Neisser, 1976; Rao and Ballard, 1999). Inference is indeed mandatory to recognise 3D objects based on their 2D projections on the retina, since this is an ill-posed problem[3] (Pizlo, 2001). Vision is thus – according to some theories – largely based on the internal knowledge about objects of the outside world. Within an inferential framework neurons signal the probability of the presence of a feature in the outside world, based on the context, the statistical structure of the visual world, and internal knowledge as the memory of learned categories.

That such a framework would be possible is shown by the sometimes short latencies of neurons in higher cortical areas such as MT, MST, and FEF (for an overview see Bullier (2001b). It has further been shown that neurons in V1 can encode different information with different latencies. Whereas first local information is encoded (the orientation of bars comprising a texture), with longer latency the global property of a stimulus is encoded. Hereby information is combined over larger areas of the visual field, than can be accounted for by the known length of lateral connections within V1 (Lee et al., 1998; Bullier, 2001b). In a pop-out experiment Lee et al. (2002) showed, that neurons in V1 became significantly more sensitive to shape-from-shading stimuli, after they had been used in behaviour. Long-latency neural signals in V1 and V2 were correlated with the behavioural performance of the monkeys. The authors speculate, that the sensitivity to 3D shape, which the neurons in V1 show, "may be mediated by recurrent feedback connections from V2 and/or other extrastriate areas" (Lee et al., 2002, p.596). In a discrimination task using line segments Li et al. (2004) found that neurons in V1 took on novel functional properties related to the attributes of the trained shapes. These properties furthermore depended on the performed perceptual task, meaning that neurons in V1 responded very differently to an identical visual stimulus under different visual discrimination tasks. Similar findings have been obtained in IT, where neurons first code global properties of a stimulus (human vs. monkey face) and with a delay of about 50ms encode more detailed properties (identity and facial expression). The authors hypothesise that the global information acts as a "header" for switching the processing stream in an higher area, thus preparing cortical destination areas for the exact nature of the following more detailed information (Sugase et al., 1999). Also experi-

[3]An infinite number of different objects can create the same 2D projection on the retina. Therefore it is, in general, impossible to solve the inverse problem of determining the object creating a specific image on the retina, without applying constraints.

ments using transcranial magnetic stimulation (TMS) showed that activation of the lowest cortical levels by feedback from higher stages is necessary for conscious vision (Bullier, 2001a; Pascual-Leone and Walsh, 2001).

All these findings support the view that descriptions of purely feedforward processing based on an hierarchical filtering of increasing complexity might not be sufficient to capture essential properties of the human visual system. Instead adaptive behaviour needs to be taken into account, where the function of certain visual areas depends on context and internal knowledge, but may also change during the time-course of visual analysis.

Chapter 3

Psychophysics and Model Predictions of Visual Object Recognition

3.1 Introduction

The special difficulty of modelling object recognition and classification is its high flexibility. Observers can not only distinguish between very dissimilar objects, but also between objects that seem to be very similar. Properties of an object that are important in one context (recognising a certain car model) may be completely irrelevant in another (discriminating between a horse and a car). Further an observer can recognise objects in the early morning, at noon, with a variety of different artificial light sources and sometimes even at night (no colour vision). All this is not only done with the object held in a fixed position, but also from a variety of different viewpoints and from different distances. To this day, these demands pose considerable difficulties for models of human visual object recognition, as well as for object recognition in computer vision.

3.1.1 Common Constraints

Although existing theories of object recognition differ in many respects, they all have to address, in varying degrees, how perceptual representations are derived from visual input, how these percepts are encoded in memory and how the matching between an unknown input signal and those encoded in memory is achieved. In their review on visual object recognition Tarr and Vuong (2002) list some constraints, which, as the authors claim, all models

of human object recognition need to consider.

Object recognition models have to account for transformations of the input images. Such transformations can be purely 2D-affine, such as scale, translation and rotation in plane. More demanding are transformations based on the three dimensional character of objects, a rotation in depth, or a change of the character of the light source, especially its position.

Object recognition models need not only be able to generalise over different views, but also over different instances of a visual object class. An observer must be able to determine from learned instances of a class, whether an unknown view belongs to this class. There is no more basic process in perception, since whenever something is seen as a kind of thing, categorisation is happening (Lakoff, 1987).

With categorisation, there comes the need to define the level of specificity, at which objects are to be classified. Categorisation processes studied by cognitive scientists are often structured hierarchically (Rosch et al., 1976). In categorising his German shepard Rex, an observer can go from the superordinate level (animal), over the basic level (dog), the subordinate level (German shepard) to the individual level (Rex). In this hierarchy basic level is the default level of access. Spontaneously an observer would categorise something as a dog, not an animal or a mammal and not as a German shepard. Visual categorisation has some distinctions. For one it relies purely on visual properties and second there is a different default level of access. Joliceur et al. (1984) point out that objects are not always recognised at their basic level, but at the "entry level", where people seeing a sparrow label it "bird" and seeing a pelican label it "pelican" and not "bird".

3.1.2 Distinguishing Principles for Models

Marr (1982) defined three criteria of efficiency for object representations: Accessibility, scope and uniqueness, and stability and sensitivity. What they say in plain words is: Can a shape description be efficiently computed from an input image, what kind of shapes can be represented, does one and the same shape possess several descriptions, do different shapes always have different descriptions, and is the degree of similarity between two objects reflected in their descriptions, while at the same time subtle differences can be expressed?

Further Marr saw three design issues: the choice of coordinate system, primitives, and their organisation. Is the coordinate system used to describe objects viewer-centred or object-centred? What kind of primitives, the most elementary units, are used to describe a shape, what information do they carry and at which size? How are the primitives organised, in the simplest case not at all, all elements have equal status, or are they grouped, for

instance, into modules of spatially adjacent elements of roughly the same size?

Another distinction, which also considers how the right object can be chosen from the database of all the objects stored in memory, comes from Dickinson (1993). For different models he compares the "indexing primitives", which are the image structures that are matched to object models. He finds a number of factors, which depend on the choice of the indexing primitive: The complexity of the model itself, the complexity of the search for the right model, the reliance on verification, i.e. verifying that the correct model was indeed found, the flexibility of the model, and the ease of recovery of the primitive from visual input. As can be seen in Fig. 3.1, the tradeoff between primitive complexity and the other factors mentioned needs to be considered. As the indexing primitives for a model become more complex (Fig. 3.1, 1), the number of primitives needed to describe an object decreases, since an object can be described by a few complex parts or by many simple ones, thus the model itself is less complex (2). Also the number of possible combinations decreases for complex features, which means that the search for the correct stored model becomes easier (3). As Dickinson (1993) notes, for that reason most systems using simple indexing primitives confine themselves to small object databases (e.g. Huttenlocher and Ullman, 1990). Since simple indexing primitives allow a more ambiguous interpretation of an object view – many different objects can be composed of the same simple parts – top-down verification is necessary to disambiguate this information (4). The reliance on top-down verification also implies that the relative locations of indexing primitives are well known. The model thus becomes very sensitive to minor changes in the object shape, the flexibility decreases (5).

So far, all criteria have favoured complex parts over simple ones, but there is a price to pay: The reliable recovery of complex features, particularly from a single 2D view is a difficult problem, especially in the presence of noise or occlusions (6). This major obstacle is probably the reason, why many object recognition systems use simple indexing primitives.

Since object recognition is the process of matching the representation of a 2D image of an object to the representation of the object stored in long-term memory, another distinction can be made, according to how emphasis is placed on the two parts, which jointly determine the properties of this process: Representation and matching. The more complex the representation is, the simpler the matching process could be and vice versa (Hummel, 2000). What is not considered by many theories is the actual process used for coding the input image, which includes figure-ground segmentation, the (optional) segmentation of the actual object into parts, and the extraction of features

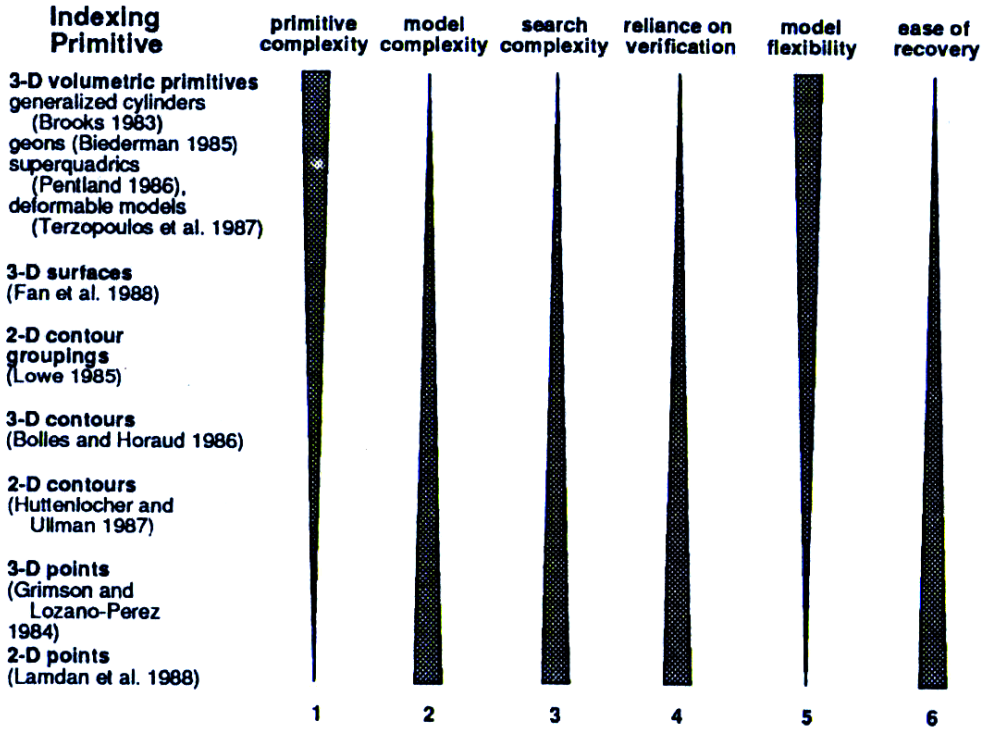


Figure 3.1: A list of possible indexing primitives ordered by their complexity and the influence they have on several properties of a recognition model: With increasing complexity of the indexing primitives (1), the model complexity decreases (2), the search complexity decreases (3), the reliance on verification decreases (4), the flexibility of a model increases (5), and the ease of recovery for such primitives decreases (6) (from Dickinson, 1993).

and their relations. Also the dependency of this coding process on the amount of information a given view of an object provides needs to be regarded (Liu et al., 1995).

The existing psychophysical models of object recognition can be roughly sorted into two classes. The first approach assumes that objects can be decomposed into their constituent parts and explicitly specifies the relations between those parts. As far as the parts of an object (and the process of decomposition generating them) and their relations are invariant to changes in viewpoint so also the model of the object will show this invariance. Since the best known examples of such a model use volumetric primitives, showing invariance to viewpoint changes, models of this type are often called viewpoint invariant, since they are able to recognise objects with approximately constant performance, regardless how the image of an object on the retina

changes – of course within certain limits (Marr and Nishihara, 1978; Biederman, 1987). The second approach represents object features as they are seen when originally viewed and do not make the relations between them explicit. Such models preserve object shape information in a viewpoint-dependent manner. Features in the input image are compared to features in learned object representations. This is done by transforming either the input image or the stored representations to produce a match (Tarr, 1995; Ullman, 1998) or by determining a statistical evidence for matching the input image to a representation (Bülthoff and Edelman, 1992; Riesenhuber and Poggio, 1999; Perret et al., 1998)

Models based on the decomposition of an object into parts and their relations are often referred to as structural description models. They often use primitives of high or very high complexity and the emphasis is on the representation part of object recognition. Models making direct use of 2D image features and their spatial locations are often referred to as view-based or image-based models. Image-based models use primitives of low complexity and accordingly the emphasis is on the matching part of object recognition (Tarr and Vuong, 2002; Hummel, 2000).

3.2 Structural Description Models

A solution to the problem of recognising a 3D object from its 2D projection on the retina, proposed by Marr (1982), is reconstruction of the 3D scene. After consideration of several efficiency and design criteria, Marr and Nishihara (1978) assumed the internal representation of 3D objects (1) to use a 3D object centred coordinate system, (2) to be assembled from volumetric primitives, and (3) to have a modular, that is hierarchical, internal organisation. The design goal was to arrive at a stable representation, being largely invariant to changes in viewpoint. For this end Marr and Nishihara (1978) designed a model based on the axis of elongation of an object. Generalised cylinders express size and orientation of these axis, which are themselves composed of smaller cylinders in an hierarchical fashion.

The process of reconstruction starts from local features and continues with combining lines into contours, contours into surfaces, and surfaces into volumes. The strength of this model is the complete algorithmic description of the recognition process, where the processing of the input image and the extraction of a 3D model are completely data-driven. Nevertheless, this is also one of its disadvantages, since this makes it a completely deterministic process. This means the same types of features are combined to complex descriptions using always the same types of rules, regardless of the nature

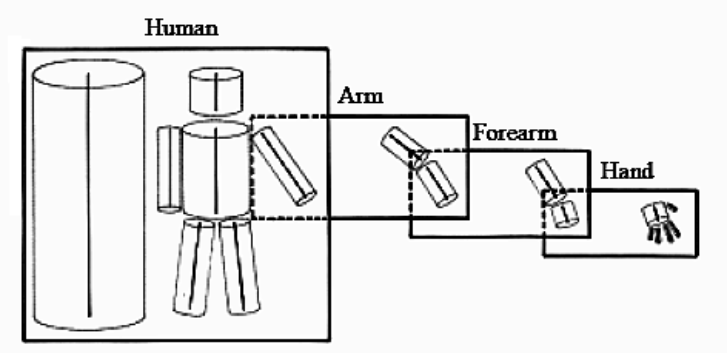


Figure 3.2: A diagram illustrating the hierarchical organisation of shape information in a 3D object-centred model. On every level of resolution the object can be approximated using generalised cones. (From Marr and Nishihara, 1978)

of the input image. Further, a feature within an image is either existent or not, there is no scope for probabilistic or partial statements. Finally, the difficulties of reliably extracting 3D generalised cones from a single 2D image have not been mastered yet (Tarr and Vuong, 2002).

Further studies of high-level vision were strongly influenced by Marr's work and one of the theories, building on Marr and Nishihara (1978) is the "Recognition-By-Components" (RBC) theory (Biederman, 1987), also termed "geon structural description" (Biederman and Bar, 1999). Much as the hierarchical model by Marr and Nishihara (1978), the geon structural description model (GSD) is a specific example for a structural description model. It also holds that objects are represented by volumetric parts, but introduces several modifications. The potentially infinite number of possible volumetric primitives has been reduced to a fixed set of 36 volumetric primitives, called "geons". The geons themselves are defined by the non-accidental properties (NAP) of 2D contours. According to Biederman and Bar (1999), NAPs (such as whether a given contour is straight or curved) are rarely produced by accidental alignments of viewpoint and object features. This implies that they are generally unaffected by slight changes in viewpoint. By enumerating the possible combinations the 36 geons are derived, each with a different signature of NAPs. With this method, the problems in the model of Marr and Nishihara (1978) like deriving depth information and extracting generalised cones from 2D images, and the transformation into a 3D object-centred coordinate system are avoided.

NAPs are assigned a special status in GSD, since they enable instant viewpoint-invariant recognition, as opposed to metric properties (such as

an object's aspect ratio or its degree of curvature), which can be affected significantly by rotations in depth. Comparing the performance of subjects in a same-different matching task, where objects either differed by a change in a metric property or in a NAP – both changes dimensioned for equal detectability – a striking advantage of NAPs for object recognition under rotations in depth was found (Biederman and Bar, 1999).

The GSD, finally, consists of a small set of geons, which are not organised hierarchically, but are represented within a "flat" structure. The relations between these geons are qualitative (e.g. "is above", "is beside") and are specified within a viewer-centred frame (see Fig. 3.3). The GSD model is therefore not really a 3D object-centred model.

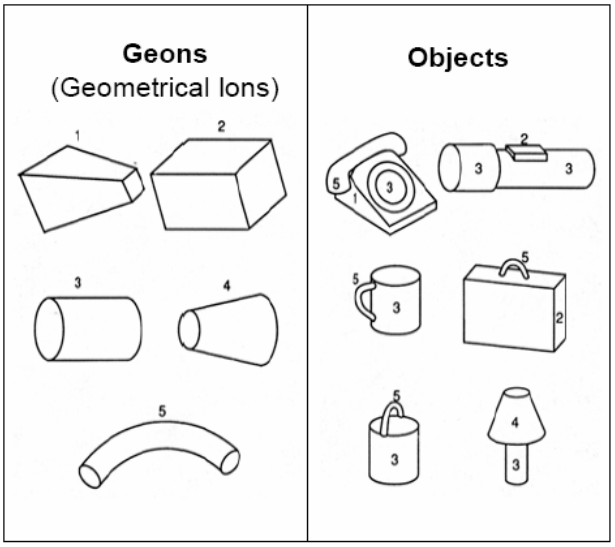


Figure 3.3: Some examples of possible geons and how the can be combined to objects. Qualitative relations can differentiate between objects composed of the same parts (e.g. "small curved cylinder above large straight cylinder" and "small curved cylinder beside large straight cylinder" describe a bucket and a cup, respectively) (From Biedermann, 1985).

To apply the geon structural description model (GSD) three preconditions for invariance have to be met, as proposed by Biederman and Gerhardstein (1993): (1) objects are decomposable into parts, (2) each object must be composed of a distinct configuration of parts, and (3) different viewpoints of an object must show the same configuration of parts. The application of these principles to stimuli, which elicited a strong viewpoint-dependent behaviour in human subjects, to so-called wire-frame or bent paperclip stimuli

(Edelmann and Bülthoff, 1992; Bülthoff and Edelman, 1992), had a striking effect. When one of the segments of the wire-frame stimuli was replaced by a distinct geon, rotation costs were dramatically reduced (Biederman and Gerhardstein, 1993).

Biederman himself has conceded, that a GSD model cannot be totally viewpoint-independent (Biederman, 1987, 2000). Rotations in depth would alter the GSD of an object, since parts can be occluded or revealed, so that a cost function is required, based on the similarity of GSDs. This implies, that objects with views which are substantially different need more than one GSD for their representation.

Major limitations of GSD were also pointed out by other authors (e.g. Tarr and Bülthoff, 1995). One is the difficulty in deriving 3D volumetric parts from a 2D projection in a reliable and stable fashion. Depending how an object is decomposed into parts, the resulting structural description can vary considerably. This argument is countered, however, by noting, that humans have no major problems with this task. So it is actually a research problem of finding and implementing an algorithm that can solve the segmentation problem (Biederman, 2000; Hummel, 2000). Another problem arises from the fact that detailed information of the shape present is lost since many different possible shapes are reduced in their description to a single label, the name of the geon (Tarr, 2002). It is claimed, that this renders GSDs useless for subordinate or individual level categorisation (Tarr and Bülthoff, 1995). However, Biederman and Bar (1999) claim, that the changes in NAPs introduced in their experiment are so small, that the respective matching task actually does concern the subordinate level. Another point of criticism is the "determinism" of GSD (Tarr and Bülthoff, 1995). Regardless of the input, the same procedures lead to an absolute statement about the presence or absence of geons and the relations between them. Consequently there is no room for representing uncertain knowledge, arising from noisy images, which also may contain occlusions. Other critics claim that the scope of the GSD model is severely limited by two factors: (1) Additional information about geons, such as size, colour, and texture is neglected. (2) Different objects must have distinct configurations of parts, excluding a large number of scenarios, where similar objects need to be distinguished (e.g. recognising different cups) (Wallis and Bülthoff, 1999).

As is acknowledged by proponents of the GSD model, there is a further problem with viewpoint-independency resulting from the type of categorical relations defined between parts. They are of the form:
"VERTICAL CYLINDER ABOVE PERPENDICULAR SMALLER THAN X", which would describe the qualitative relationship between two adjacent cylinders, which are part of the same object. Such relations are based on a viewer-

centred frame (Biederman, 2000; Hummel, 2000) and are therefore not object-centred. It is obvious, that the relation "is above" is in no way viewpoint-independent. A rotation in the picture plane of more than 45 degrees can invalidate this relation. It is claimed therefore, that by using viewer-centred relations the model can account for the sensitivity of human object recognition to rotation in the image plane, but retains invariance to rotation in depth, translations, scale changes and mirror reflections (Hummel, 2000)

An extension of the geon structural description introduced by Hummel and Stankiewicz (1996) is based on a hybrid representation of objects, where image-based representations are integrated into the structural description, making them one type of components. Since image-based components are obtained faster than the other components and are also more sensitive to viewpoint changes, this model makes, as the authors claim, correct predictions about the time dependency of object recognition performance.

A further extension of this model (Stankiewicz and Hummel, 1996; Hummel and Stankiewicz, 1998) represents the shape dimensions of the geons not in a categorical manner (e.g. curved or straight primary axis), but by continuous variables. It can thus differentiate between different degrees of primary axis curvature and the other independent dimensions used to describe geons.

3.3 Image-Based Models

The second major class of object recognition models contains image-based models. They all represent 3D objects by multiple 2D views, but there are variations regarding the primitives used, how they are organised, and in which way matching of an unknown view occurs. Since the representation of objects is tied to specific views, some mechanism of normalisation or generalisation is needed, to keep the number of views feasible. It is assumed that in human object recognition no decomposition of an object into parts takes place, with their relations being specified explicitly. Instead objects are recognised by matching their views to stored image-like views in memory.

Image-based models were inspired by the large body of experimental results reporting viewpoint-dependent performance in visual object recognition. This is expressed, especially at the subordinate level of categorisation, by the existence of so-called canonical views for familiar objects (Palmer et al., 1981). However, since the image-based models, used to explain viewpoint-dependency, tend to predict a near viewpoint invariant performance for highly familiar objects most experiments use unfamiliar stimuli which are novel to the subjects.

In a task of matching same or mirrored novel objects, a nearly linear

relation between the angular difference – either in plane or in depth – of the two objects and the reaction times of the subjects was found (Shepard and Metzler, 1971). Introspectively, the subjects felt like mentally rotating one version of the object and matching it to the other version. This finding gave rise to the "mental rotation" theory of object recognition. In a different experiment, using novel 2D "character-like" stimuli rotated in the image plane, subjects learned to recognise them from several orientations. When new orientations were introduced, naming times increased monotonically with the angular distance from the nearest familiar orientation (Tarr and Pinker, 1989). The authors hypothesised, that the same mental rotation process as found by Shepard and Metzler (1971) was used to normalise the input images. This finding was corroborated by later experiments using 3D versions of the initial stimuli (Tarr, 1995). This increase of reaction time with the angular difference, between learned and tested views of an object has also been found in other experiments (Edelmann and Bülthoff, 1992; Humphreys and Kahn, 1992), but the interpretation was different (see below).

The concept of mental rotation was criticised by some authors. One claim was that the main evidence, the increase in reaction time, could be caused by any process that takes more time the greater the differences between learned stimulus and test stimulus are (Perret et al., 1998).

The immediate and most evident problem with mental rotation that someone already needs to "know" (i.e. have recognised) an object to determine the necessary transformation is addressed by the alignment model by Ullman (1989). The approach of recognition by alignment compares the input image with a projection of a stored model. The necessary transformation of the model is computed by matching a small number of local features in the image with the corresponding features in the model. The model can be a full 3D model, or the alignment can rely on pictorial descriptions, using multiple views. In recognition every stored model is aligned with the input image and the difference between them is computed based on the pixel values. The model which minimises this difference is assumed to be the correct match. Note that this model can, in principle, achieve viewpoint-invariant recognition, though the problem of occluded features needs to be solved. The major problem is to solve the correspondence problem: Which feature in one image corresponds to which feature in another image. This is by no means a trivial task.

Ullman's work was extended into a model using a linear combination of 2D views (Ullman and Basri, 1991; Ullman, 1998). The model is based on the assumption that the locations of corresponding pixels or features of an arbitrary view of an object can be described by the linear combination of a set of suitably chosen views. The number of views depends on the type and range

of allowed transformations and the nature of the represented object. It varies from two and three to five and possibly more (Ullman, 1998). For recognition the unknown coefficients of the linear combination need to be recovered, then they are used to produce a new model image from the stored ones. Finally, this synthesised model image is compared with the input image. Methods proposed for recovering the coefficients are iterative and exhaustive search, both computationally expensive, especially, when it is considered that a large number of objects are learned. An alternative is to recover the coefficients based on a small number of matching image and model features. How the correspondence between the features is to be determined – which some would claim already solves the most difficult part of the recognition problem – is not specified.

As mentioned above, in generalising from learned views of novel objects, other experimenters found a dependency of error rates and reaction times on the angular difference. Using wire-frame objects – also termed "bent paperclips" – which were generated randomly, Edelmann and Bülthoff (1992) showed the existence of canonical views and a monotonic increase in response time with angular distance from such a view. This effect was shown to disappear with training. In a different setting, using the same stimuli, the authors found a dependence of error rates on the size and the direction of the angular difference to learned views. After training two views of an object, a novel view was presented, which was either between the two views (interpolation), outside the two views, but on the same circle (extrapolation) and, finally, outside the two views, but in a direction orthogonal to the circle, connecting the two learned views (see Fig. 3.4). Again the authors found a dependency on the angular difference between learned and novel views, but also a dependency on the direction, with interpolation yielding the smallest error rates and the orthogonal direction resulting in the highest. The results were interpreted as evidence for a theory of view interpolation. In the latter theory objects are represented by multiple views, based on the features as they appear in the learned 2D views, and recognition performance for novel views depends on the distance to the nearest learned views (Bülthoff and Edelman, 1992).

HMAX is an example for the view interpolation algorithm, and uses a multi-layered feedforward network (Riesenhuber and Poggio, 1999). Starting with local feature detectors, HMAX uses in alternation two types of operations, namely feature disjunction and feature conjunction. Feature disjunction is brought about by a maximum operation, which results in translation and size invariance. Feature conjunction results from the computation of a weighted sum and corresponds to the combination of simple features to more complex features.

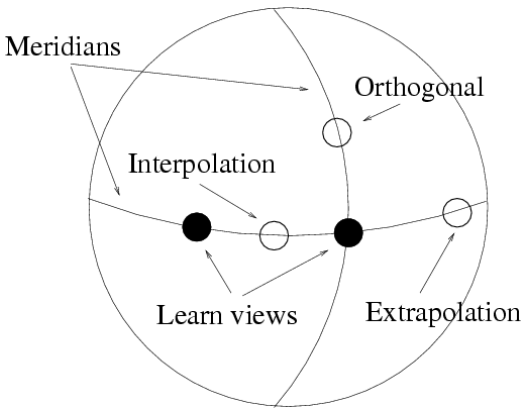


Figure 3.4: Schematic illustration of the positions of the camera on the viewing sphere relative to the two learning positions, creating orthogonal and inter- and extrapolation views. Adapted from Bülthoff and Edelman (1992)

The HMAX algorithm can be seen as an extension of Hubel and Wiesel's scheme of superimposing spatially adjacent receptive fields of simple cells of some orientation to generate receptive fields of complex cells. Repeated alternation of the operation of maximum selection and weighted summation finally leads to view-tuned units, which can then be combined for recognition or classification tasks. Although more refined than earlier approaches to view

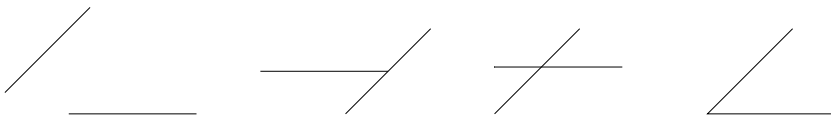


Figure 3.5: Assuming line detectors as most basic units, the HMAX algorithm by Riesenhuber and Poggio (1999) would ignore the differences between the shown object parts of facets (as long as both lines fall within the area of the maximum operation).

interpolation, which were essentially limited to template matching (Poggio and Edelmann, 1990), a number of issues remain unclear. HMAX essentially matches a list of features in the input image to the lists of features of stored views, determining the closest match. Due to the architecture of HMAX, simple features can be combined to more complex features, thus capturing the structure of the object to a certain extent. The mechanism of selecting the feature detector with the maximal activity, thus achieving translation invariance, implies the loss of information that may be essential. Furthermore the exact location of the selected feature and, even worse, its relation to

the other features it will be combined with, is lost too. Depending on the exact architecture, the model would have difficulties distinguishing between possible object parts as shown in Fig. 3.5. The shortcomings of a model based on feature lists were pointed out by by a number of authors (Hummel and Biederman, 1992; Stankiewicz, 2002b; Bischof and Caelli, 1997b; Bischof, 2000). In brief, configurational information is largely lost, by detecting the presence of features only. Simply listing the features of an object could be seen in analogy to listing the colours of a painting (Fig. 3.6). Both lists are not sufficient to specify the object or the painting respectively (Hummel, 2000).

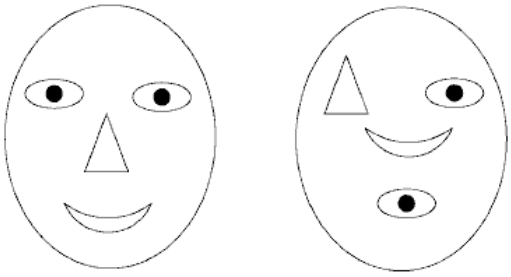


Figure 3.6: An example of two stimuli which are not distinguishable purely by feature lists without using configurational information. (From Bischof, 2000, p. 298)

Taken as a model of human object recognition, it is not specified by the authors how the deterministic connections and weights in HMAX are being established during learning. More generally, the model suffers from the typical drawbacks of purely feedforward models. It contains no mechanism for reasoning, which in this context would mean to adjust the weights and connections according to internal knowledge and the external input. An attempt is being made to construct the representation of higher level structures by building compound complex features from simpler, neighbouring feature components. Nevertheless, the representation of structural information remains poor, due to the usage of pure lists of features.

Finally, a quite different approach is the Evidence Accumulation Model, which allows the recognition of familiar objects (Perret et al., 1998). Here it is assumed that with small transformations of an object (e.g. rotation) its visible features change only gradually. During learning a connection is established between feature detectors and a decision unit, which enacts a temporal summation, thus accumulating the evidence from the feature detectors. The larger the transformation of the object from the initially learned pose, the less

this particular pool of feature detectors will be active, increasing so the time necessary to reach a threshold for the accumulated evidence. This model is a feature list model as well and suffers therefore from the same limitations as the HMAX model. Configurational information is preserved only in so far as there is a feature detector tuned to such a complex configuration. Consequently, it is possible to construct indistinguishable object views for the Evidence Accumulation Model as well (see Fig. 3.5 for comparison).

The first to demonstrate viewpoint-dependency for wire-frame stimuli were Rock and DiVita (1987). They were then considered to be "well-suited for studying the basics of object recognition" since they were missing factors such as self-occlusion (Bülthoff and Edelman, 1992, p. 63), and were widely used in psychophysical (e.g. Liu et al., 1995, 1999; Christou and Bülthoff, 1999; Foster and Gilson, 2002) and neurophysiological work (e.g. Logothetis et al., 1994, 1995; Logothetis and Pauls, 1995). Nevertheless, these stimuli have evoked some criticism from different authors. For instance it is claimed that this type of object lacks critical information for everyday shape recognition, a distinguishing geon structural description (Biederman, 2000). Indeed, bent paperclips do not conform with the criteria for viewpoint-independent recognition, as formulated by Biederman and Gerhardstein (1993). The latter authors also demonstrated, that replacement of a single segment by a distinctive geon-type component results in near viewpoint-independence. Another author, Pizlo, claims, that stimuli such as wire-frame objects force the human recognition system into a mode where vision operates in a purely feedforward way (Pizlo, 2001). He argues from experimental results, which show a recognition performance for rotating wire-frame objects at nearly chance level (Pizlo and Stevenson, 1999). Pizlo concludes, that "the constraints that are potentially useful in the case of wire objects are not used by the human visual system" (Pizlo, 2001, p. 3152).

Nevertheless, there is no final conclusion with regard to this debate. On the one hand there remain a substantial number of experiments, some of them (Hayward and Tarr, 1997) deliberately designed to conform with the conditions proposed by Biederman and Gerhardstein (1993), which show a viewpoint-dependency in object recognition. On the other hand, Hummel (2000) claims, that the viewpoint-dependency of recognition performance is not a useful criterion to differentiate between structural description models and image-based models, since both types can be tuned to show either viewpoint-dependent or viewpoint-independent behaviour.

3.4 Ideal Observer Model and Stimulus Input Information

The performance of subjects in an object recognition task depends on two factors, the internal processes of the observer and the information from the stimulus input (Liu et al., 1999). The issue of internal representation for object recognition has been addressed by a large number of publications. In contrast, the role of stimulus input information has received far less attention. However, it had been noted early, that observers rate the "canonicality" of object views according to the visibility of object information – and its subjective importance – in a given view (Palmer et al., 1981). This approach was further developed and integrated into an ideal observer model comparing the performance of human observers to different models of 3D object recognition (Liu et al., 1995). The authors distinguish two types of approaches in modelling human object recognition. The first approach makes qualitative predictions about the relative performance of subjects for different classes of stimuli and different tasks. One problem here is the lack of a "common currency" for comparing performance over different stimuli or tasks. Consequently, the effects of internal processes of a subject tend to be confounded with effects related to differences in task-relevant information available in the stimuli. As the authors say "it is quite possible that the causes of differences in performance across different stimuli or tasks ... are in the stimulus, not in the head" (Liu et al., 1995). The second approach compares the performance of particular models to that of human subjects. The problem here is that most models are constrained by the computational theory on which they are based (the interesting part) and by implementation constraints (confounding the results).

As Liu and co-workers claim, the ideal observer approach, based on signal detection theory (Green and Swets, 1966; McNicol, 1972), allows one to make qualitative predictions of performance and provides a common measure of performance across different types of stimuli and tasks. Different ideal observers can match different computational constraints, allowing for the comparison of different models of 3D object recognition. Ideal observers are the optimal implementation of a theory in the sense of giving the best possible absolute performance. This allows one to define a statistical efficiency, which is measured in terms of relative signal-to-noise ratios needed by human observers to achieve the same level of performance across different stimuli. Noise is used here in the sense of stimulus uncertainty, either inherent in the task or artificially introduced. An important aspect of the design of experiments here is that some uncertainty is needed, as the statisti-

cal efficiency of the human subject can not be computed if the ideal observer achieves perfect performance. Therefore this method of evaluating models can not be applied post-hoc to an experiment, which wasn't designed for it.

The ideal observer approach was applied to the recognition of wire-frame objects (Liu et al., 1995), which fell into four classes: (1) Only the 5 vertices were represented by rendered 3D spheres, (2) the spheres were connected with wires of the same diameter, (3) the vertices were arranged to be mirror symmetric, and (4) the vertices were further constrained to be not only mirror symmetric, but also planar. Liu and co-workers compared the recognition performance of human observers with the performance of models based on 2D templates, models using 3D information, and a neural network implementation of the view interpolation theory (Poggio and Edelmann, 1990). Their results showed that neither a model based on 2D templates, nor the neural network model could account for the performance level of human subjects under all experimental conditions. The subject seemed to make use of special aspects of 3D structure, such as symmetry and planarity. The viewpoint dependency of object recognition seemed to depend on the object structure, showing little or no viewpoint-dependency when the objects were highly regular.

Another experiment used similar stimuli and lead to the finding, that differences in human performance for object classes of different complexity are at least partly a function of the internal representations learned (Liu et al., 1999). The authors stress the point, that this conclusion could not have been arrived at without the quantitative analysis of stimulus input information. Other studies (Tjan and Legge, 1998), which also used the ideal observer approach, computed a measure termed "view complexity" for several classes of objects, such as wire-frames (Rock and DiVita, 1987), simple geometric objects, mechanical parts, and charm bracelets (Biederman and Gerhardstein, 1993), and faces. The authors found a correspondence between the view complexity of those objects and earlier behavioural data (Biederman and Gerhardstein, 1993; Edelmann and Bülthoff, 1992; Troje and Bülthoff, 1996). These results were suggestive to the authors of these studies, that the debate about the nature of human object representation is partly based on a failure to distinguish between internal processes and the properties of the stimuli used.

Chapter 4

Syntactic Approach to Visual Object Recognition

4.1 Introduction

4.1.1 Overview

CLARET-2 is an algorithm for pattern matching and object recognition which is based on multiple views, nevertheless depending on structural information. A detailed description can be found in Sec. 4.2 and in the Appendix (App. B). CLARET-2 derives from CLARET, which stands for Consolidated Learning Algorithm using Relational Evidence Theory. CLARET was initially developed by Adrian Pearce (Pearce, 1996; Pearce and Caelli, 1999). It has been extensively debugged and modified by the author and now consists of over 45000 lines of C Code (see Sec. 4.2.4 for a list of modifications). In brief, CLARET-2

- is a highly adaptive procedure of stepwise refinement, determining the similarity between views in increasing detail as processing progresses. The refinement acts on single attribute dimensions only, leading to simple rules and allowing the determination of the relative importance of each attribute dimension for the recognition task at hand.

- utilises structural descriptions, representing views of objects by constituent parts and the explicit representation of relations between those parts. A minimal structural description is guaranteed by the fact that only binary attributes are used, which describe the relations between two parts and their respective properties relative to each other. However, structure can be represented more detailed if necessary by the expanding the pairs to triplets, quadruples, and so on.

- has an inbuilt ability to generalise, since its adaptive behaviour depends not only on the properties of the previously learned views, but also on the properties of the unknown view.

- places emphasis on the matching part of the recognition process, determining likely correspondences between the parts of the learned views and an unknown view. This matching process also operates by stepwise refinement from a very coarse and imprecise matching to an increasingly detailed matching. A plausibility check is included at each step, rejecting any refinement leading to contradictory mappings of parts.

- provides an efficient method of indexing the database of stored views, by computing the probabilities of each matching between a learned view and the unknown view, discarding unlikely matchings at every refinement step. This reduces the number of possible candidates as the representations become increasingly complex.

4.1.2 Reasons for Using the CLARET-2 Algorithm

From the overview, the question might arise, of whether such a seemingly powerful but complicated algorithm is necessary for modelling object recognition. Why is a structural description model needed, since image-based models seem to be the most prominent and successful models so far? The success of image-based models (Poggio and Edelmann, 1990; Ullman and Basri, 1991; Perret et al., 1998; Riesenhuber and Poggio, 1999), which don't use structural descriptions, is largely based on the fact, that they are inherently able to model viewpoint dependency in object recognition. Nevertheless, this is a misconception, since structural description models are also able to describe viewpoint-dependent behaviour (Hummel, 2000). Further the psychophysical evidence seems to suggest that object recognition is not either viewpoint-dependent or viewpoint-independent, but can be both, depending on the stimuli and the task (see Sec. 3 and Hayward and Williams, 2000). According to Hummel (2000), the issue of viewpoint-dependency is not the core of the debate about object recognition. Instead, the important question is whether object recognition can be adequately dealt with in terms of an approach that ignores object structure. Hummel claims that this is not the case. Also Tarr and Vuong (2002), although being strong advocates of image-based models, concede that it might be better to abandon viewpoint-dependency as a guiding principle. In their view, measuring the similarity between objects could be a new guiding principle. Yet, they see the problem that "there is currently no notion of how to measure 'similarity'." (Tarr

and Vuong, 2002, p. 36). Questions about the correct feature set and how features are compared to one another are further problems which they see as unanswered. These are questions CLARET-2 could be well suited to answer, since it was designed to determine the similarity between patterns or objects, and it has a method of comparing features (or attributes) and determining, which ones are relevant or irrelevant.

CLARET-2 could also answer the question as to which parts and features in the input image correspond to which parts and features in the object model representation. Algorithmic approaches in the literature, such as the alignment model Ullman (1989) and the model of linear combination of views (Ullman and Basri, 1991), depend on a solution of the correspondence problem, but provide no answer on how it could be solved. Other algorithms assume the correspondence problem as already solved – by the operator of the algorithm – as for instance the neural network implementation of the view-interpolation model (Poggio and Edelmann, 1990). There are also models described in the literature that simply ignore the question of correspondence, such as the evidence accumulation model (Perret et al., 1998) or HMAX (Riesenhuber and Poggio, 1999), which operate on lists of, admittedly, complex features (see Sec. 3.3 and Bischof and Caelli (1994, 1997b); Hummel (2000) why this may not be sufficient). By contrast, the attempt to solve the correspondence problem is an inherent part of the CLARET-2 matching procedure.

Another problem, which has rarely been addressed since Marr (1982) in the field of human vision, is the question of how the search for the right object model, i.e. the indexing into the database of internal knowledge about the world, is done in an efficient manner. The alignment model and the model of linear combination of views assume that the differences between the input and every internal model are computed, to find the best match. Some models do not address this problem at all, or they assume the existence of only small sets of possible objects. With its procedure of eliminating unlikely models from the matching process, in step with the increasing computational demands due to its refinement, thus retaining only the most probable candidates, CLARET-2 could offer a possible solution to this problem.

There are further reasons for using adaptive algorithms for human object recognition. From an ecological point of view minimising the necessary computational power for solving the problem at hand can save time, energy, and neural resources, which are all very valuable. Also CLARET-2's adaptive algorithm returns a preliminary hypothesis as to the right solution at any given point in time, it is what is called in machine vision an "anytime-algorithm" (Dean and Boddy, 1988). For instance, if something is coming flying at you, you might want to decide very quickly, whether to catch it, or to avoid it.

In such a situation, you would not want to wait, until the object recognition problem is exactly solved. Also in the light of neurophysiological evidence an adaptive algorithm which uses stepwise refinement dependent on its internal knowledge seems plausible (see Sec. 2.2 for the importance of feedback connections, their possible utilisation for adapting visual processing according to internal knowledge, and the time-dependent processing of information, first on a coarse and later on a finer level).

4.1.3 Machine Vision Overview

Machine vision and biological vision have a number of properties in common, which any serious object recognition system must have:

- It must be possible to recognise objects with some degree of invariance.
- It should accept partial information, as, for example, single views, degraded, or partially occluded object data.
- It should be general enough to include most known stimuli belonging to an object class, but specific enough to exclude most stimuli belonging to other classes.

Nevertheless, Caelli et al. (1993) see some differences between theories in machine vision and biological vision. Theories in machine vision are always algorithmic and implementable on computer. Usually they follow a current notion of efficiency and often lack generality and proper evaluation. On the other hand psychological theories in biological vision are often described qualitatively. They are influenced by current psychophysical and neurophysiological results and often concentrate on subprocesses, offering no complete computational models. Biological vision can profit therefore from machine vision theories, if the latter are translated in a way that makes them testable within psychophysical experiments.

In computer science literature image interpretation and object recognition have nearly become synonymous with the term ‘machine learning’. Indeed, except for technical areas, where performance is of utmost importance, machine learning algorithms provide flexible algorithms, which can be applied in various fields.

There are several approaches to pattern and object recognition in the machine learning literature. One type of algorithm uses the complete given view of an object to match it to stored object models. The simplest way is to do template matching (Duda and Hart, 1973; Hayward and Tarr, 2000). After normalisation and preprocessing (i.e. for edge detection) pixel values

of the input image are compared with pixel values of stored models. Such representations have the major problem of being neither rotation nor scale invariant. They don't allow for any variations in the input data. Other methods use global features of an object view, provided by either the Fourier or the Hough transform (Ahmed and Rao, 1975) or by principal component analysis (e.g. eigenfaces, Turk and Pentland (1991)). Again there is a certain lack of invariance with these methods and the problem remains of how to extract those global features from complete images, possibly containing multiple objects, or only degraded or incomplete objects. Therefore, if the use of the complete image is problematic, some form of segmentation of the input image is necessary, a fact, which is widely accepted in the machine vision literature (e.g. Jain and Hoffman, 1988; Fan et al., 1989).

Part-Based Algorithms

Part-based algorithms depend on some sort of segmentation of the input image. The segmentation process itself is under-constrained and depends on boundary conditions, such as knowledge about physical laws, rigidity of objects or their parts, lighting properties, and context knowledge. The segmentation of the view of an object into parts can use curvature as a defining parameter (Hoffman, 1983; Hoffman and Richards, 1984). Other possibilities are visible contours or other criteria. It is not necessary to partition the complete visible surface of an object into surface parts. One type of segmentation just locates edges and corners, without determining what is a non-edge[1]. Segmentation reduces geometric information about an object into discrete, manageable chunks, the parts of the object (Caelli et al., 1993). The segmented parts are then used as indexing primitives (Dickinson, 1993) into the database of stored object models. Therefore, by extracting parts, the complexity of the problem of matching an input image to a stored model is reduced. At the same time the, robustness of the object description is increased (see Sec. 3.1.2 and Dickinson (1993) for a discussion of the tradeoff involved in using complex indexing primitives).

The importance of parts or components has also received recognition in the psychological literature. For instance in component based models of object recognition (Biederman, 1987)(see also Sutherland (1968); Barlow et al. (1972); Hoffman and Richards (1984); Tversky and Hemenway (1984)), it was recognised, that parts give structure to perceived shape and are an index to functionality and motor programs for interaction with objects. It was found that in basic-level categorisation most attributes named by subjects

[1]In this sense the HMAX algorithm, which locates features such as oriented line segments and Biederman's geon structural description are both part-based algorithms

describe objects related to their components (Tversky and Hemenway, 1984). Indeed, the early 'recognition-by-components' theory targeted the primal access – the first contact between the input of an unanticipated object with a memory representation – in basic-level categorisation (Biederman, 1987).

The next step involved in the part-based recognition process is the description of the parts by computing appropriate attributes, such as their size, the boundary length, etc. Since these attributes are bound to a single part they are also called unary attributes. The information about the parts themselves, however, is not sufficient to describe objects uniquely. To describe the configuration of the parts it is also necessary to determine the spatial relationships between parts, such as their distance. These types of attributes are also called binary attributes[2].

In the further processing of part-based descriptions of object views, Caelli and Bischof (1997b) differentiate between two general approaches, non-inductive and inductive recognition systems. There are several non-inductive methods, among them graph-matching, where parts are represented by the vertices of a graph and the relations between parts are represented by the edges connecting vertices (Grimson, 1990; Bunke, 1998). The problem with these types of approach is their difficulty in dealing with generalisation, since they do not explicitly determine parts, their relations and the associated attribute value ranges, which are necessary and sufficient for recognising training views of object classes and unseen instances of the learned classes. Accordingly these methods often need to enumerate all instances of a class, thus becoming inefficient for large sets of objects. Additionally, they have difficulties dealing with partial data, for instance, due to occlusions.

Therefore Caelli and Bischof (1997b) give more emphasis on inductive recognition systems, which all, in some way, generalise from known class samples. Using a training set of example views of objects and object classes, their parts and the relations between the parts are described by appropriate attributes. From these descriptions rules are learned, which allow first of all the correct recognition of the training views, but more importantly the generalisation to unseen views of learned objects or novel objects of learned object classes.

[2]Looking at the HMAX algorithm from a part-based perspective, the detected features are the parts, their attributes are the response strength of the feature detector. The coding of the spatial relations is implicit in the neural network structure, but severely limited by the inbuilt scale and translation invariance.

Attribute-Indexing

Learned rules in inductive recognition can become complex and the number of rules can become large. Therefore, the question needs to be addressed, how to efficiently index and evaluate the set of rules, given the properties of an unknown view, to find the correct classification. One method is indexing by attribute values. This means all attributes are used to define a point in an high-dimensional attribute space. Learning consists of defining regions in this attribute space, which give evidence for certain classes. Learning algorithms differ in the way the attribute space is partitioned, and therefore in the type of generalisations they can achieve.

One class of learning techniques uses perceptron-like linear decision functions, as for instance discriminant function methods (Duda and Hart, 1973), linear decision trees (Quinlan, 1990a) and decision trees based on neural networks (Park, 1994). Elaborations of these methods, such as radial basis functions (Poggio and Girosi, 1990), allow arbitrary partitioning of attributes, resulting in complex generalisations, which maximally evidence each class as represented in the training set, but are not easily expressed as conjunctions of bounds on attribute values.

A second class of learning algorithms constrain themselves to partitioning boundaries, which are oriented parallel to the axes, thus forming hyperrectangles in attribute space. This allows the extraction of rules of the form

> *if attribute1* $\in$ *bounds1 and attribute2* $\in$ *bounds2 ...*
> *then evidence weights for each class are ...*

where the condition (*if*-part) is defined by a conjunction of attribute bounds and the rule actions (*then*-part) are defined, for instance, as evidence weights derived from a neural network (Caelli and Dreier, 1994).

Evidence-Based System

A specific example of an attribute indexed classification method is the evidence-based system (EBS). The EBS approach is based on the decomposition of an object into its parts, described by a list of unary attributes (e.g. size, intensity, aspect ratio) and the relations between pairs of parts, described by binary features (e.g. distance, contrast). The characteristic feature of EBS is the use of bounded regions in the unary and binary attribute spaces as probabilistic evidence rules for the occurrence of objects which have attributes within those regions (Caelli and Dreier, 1994). Objects are thus represented as a set of rules, i.e. a set of regions within which the attributes, describing the object parts, fall. Each rule, which is activated by the presence

of an attribute value, provides a certain amount of evidence for each trained class. The accumulated evidence over all triggered rules yields a measure of classification probability for an unknown object. The process of recognition involves the following steps (see also Fig. 4.1):

- Segmentation of the object into its parts.
- Extraction of the attribute values for all parts and pairs of parts.
- Activation of rules by the present attributes
- Evaluation of rules using a neural network

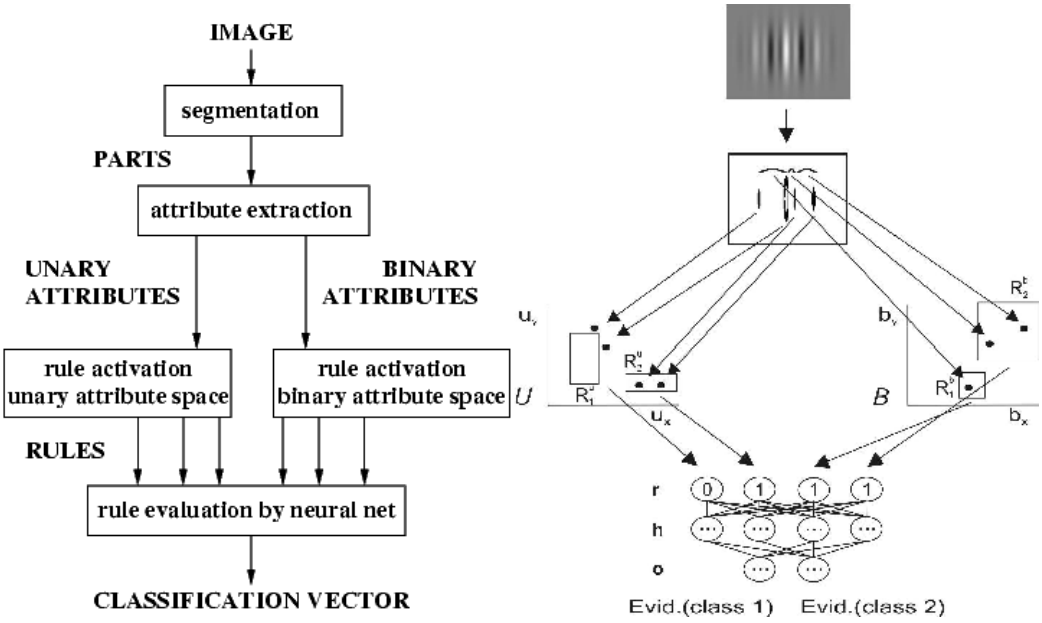


Figure 4.1: Schematic of the processing stages involved in evidence-based classification (adapted from Jüttner et al. (1997)). The input image is segmented in to parts, the attributes of these parts are computed, which activate rules in unary and binary feature spaces. The state of all rules forms the input vector for a trained neural network, which returns a vector of classification probabilities.

Rules are formed by clustering the attribute values of the training set, which are represented in two separate attribute spaces, a unary attribute space and a binary attribute space. Note that EBS uses constrained generalisation, this means that, during recognition, attributes which lie outside of rules give no evidence at all for any class.

A single rule in general gives no conclusive evidence for a class. Similar objects possess similar parts, activating largely the same set of rules, but also objects from different classes can have similar parts in common. This means that different classes are not separable by methods such as simple discriminant functions. The basic idea of evidence-based classification is that objects from the same class activate particular *patterns* of rules. Such activation patterns, which give evidence for the existence of a certain class are determined by using a neural network.

EBS was applied to modelling the behaviour of human observers in a visual pattern recognition task (Jüttner et al., 1997, 2004; Rentschler and Jüttner, 2007). Subjects were trained to classify a given set of training patterns. After successfully learning the training patterns, their ability to generalise was tested using a grey-level transformed version of the original learning set. The data of the initial learning stage provided the basis for the parameter estimation of the EBS model. For this the cumulated classification matrices of the observers were used, assuming that they reflected the formation of parameters such as attribute choices and neural network weights during learning.

The EBS analysis of the behavioural data allowed for a consistent reconstruction of the internal representation the subjects formed of the classes during learning, concerning the underlying sets of attributes, the number of extracted parts used, and the potential for generalisation.

As Bischof and Caelli (1997b) note, EBS is an efficient algorithm for learning classifications of complex patterns. Nevertheless, they see its main limitation in the fact, that structural information is only represented implicitly and in a limited form. The authors claim that the principal ways of improving systems like EBS is by representing structural information explicitly using part indexing.

Part-Indexing

The classes of learning techniques introduced until now, which use attribute indexing, are only partially adequate for object recognition systems, since they have difficulties dealing with missing parts or scenes with multiple objects. It was noted by several authors (Hummel and Biederman, 1992; Hummel, 2000; Stankiewicz, 2002a; Bischof and Caelli, 1997b; Bischof, 2000), that there is a further limitation when using lists of attributes, as is the case for attribute-indexed techniques, which is illustrated in Fig. 4.2. It is not possible to distinguish between the two patterns, indexing only by attribute values. This is a consequence of the fact, that in attribute-indexed systems structural object information is largely lost (Caelli and Bischof, 1997b). Nevertheless,

the correct labelling of unary and binary attributes in the form $u_i - b_i j - u_j$ is necessary to bind corresponding parts (u_i, u_j) and their relations $(b_i j)$.

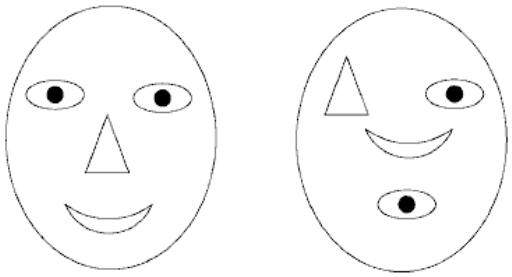


Figure 4.2: Two patterns that are indistinguishable without correct indexing of the part attributes and their respective relations (From Bischof (2000)). Both patterns contain the same sets of unary and binary attributes. To clarify, using the attribute distance between parts as an example: In both images the exact same distances occur. Without indexing, there is no information conserved as to which parts any distance is attached to.

Part-indexed systems are a type of learning system, which check for this "label-compatibility". They generate rules of the form

> *if part i has attributes in rule* u_r
> *and the relation between part i and part j has attributes in rule* b_s
> *and part j has attributes in rule* u_t
> *then part i is likely to be part x of object o*

Attribute-indexed representations rely on a single representational domain, the attribute space, where classification rules are indexed as regions in this space and all occurring attributes are evaluated using the same set of rules. In contrast, part-indexed representations rely on multiple representational domains and classification rules are indexed by sequences of parts and their relations. The attributes describing different parts and relations of an object can so each be evaluated by a different set of rules, indexed by the part label. As Caelli and Bischof (1997a) note, part-indexed systems have, in general, greater representational power, but the associated matching procedure is more complex.

Two specific methods, which combine evidence-based classification with part-indexing are the rulegraph method (Pearce et al., 1994) and conditional rule generation (CRG) (Bischof and Caelli, 1994). The rulegraph method supplements the evidence mechanism by a posterior test of structural pattern identity, using a graph matching algorithm. In contrast, CRG generates

classification rules as in EBS, which additionally include explicit structural information to the extent that is necessary to correctly classify a set of training views. CRG analyses unary and binary features of connected object components and creates a tree of hierarchically organised rules, which completely describe the training views and are then used to classify unknown views (see Fig.4.3).

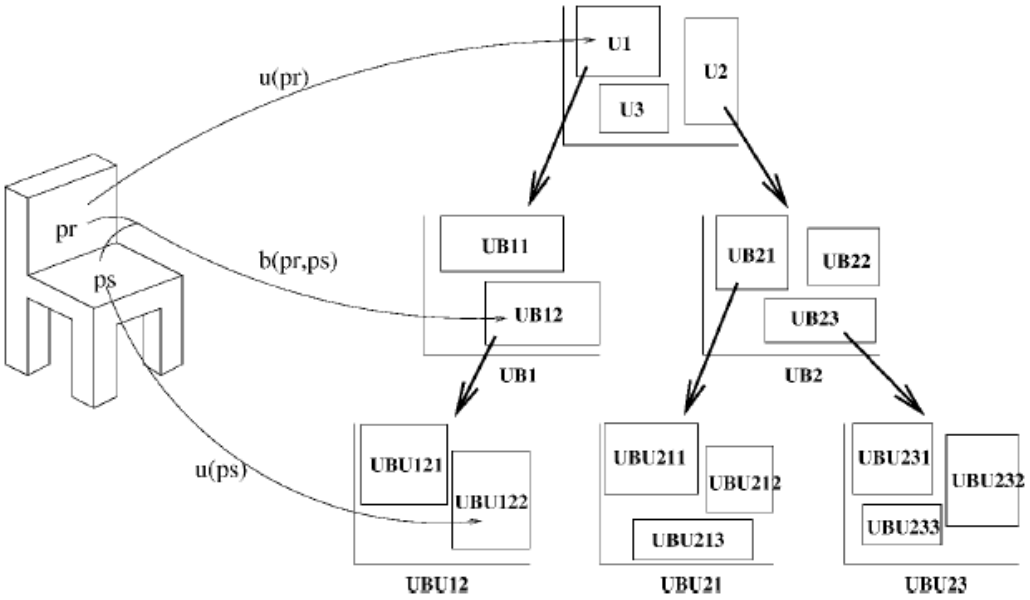


Figure 4.3: Rule tree generated by the CRG method. In the unary attribute space (top level) rules are generated (U1–U3). Some rules (e.g. U3) contain only elements of a single class. Other rules, which are yet unresolved (e.g. U1 and U2), are expanded to the binary feature spaces (e.g. UB1 and UB2), where again rules are generated (UB11–UB23). In turn, unresolved rules are extended to unary features spaces, where again rules are learned. This process of expansion and rule generation continues until all rules are resolved or the predetermined maximum rule length is reached (adapted from Bischof (2000)).

4.1.4 Features and Spatial Relations

The question of what a feature is has been answered in different ways. The use of the term "feature" is not consistent among different authors. A general description given by Tarr and Vuong (2002), citing Marr and Nishihara (1978), is that features are the elementary units used in the representation of objects. This might be a misunderstanding, since Marr and Nishihara

(1978) differentiate between features and primitives, the latter being "... the most elementary units of shape information available in a representation ...". Features describe shape at different scales, being either local or global. Tarr and Vuong (2002) make a distinction between features and spatial relations between features. For the view dependent models they cite (e.g. Riesenhuber and Poggio (1999)) features are the results of early visual processing, which are for instance the answers of line detectors or corner detectors, etc. The spatial relations between features are described by metric properties of varying flexibility (from rigid templates (Poggio and Edelmann, 1990) to no spatial relations at all (Mel, 1997)). For RBC, geons are the primitives and the identity of the geon would then be its feature (Biederman, 1987; Hummel and Biederman, 1992). The spatial relations are described quite flexibly, since only qualitative descriptions are used. There is further the distinction between "non-accidental properties" (NAP) and "metric properties" (MP). NAPs are defined to be relatively unaffected by rotation in depth, whereas MPs are. The presence of NAPs allows instant viewpoint invariant recognition, as Biederman and Bar (1999) claim.

In machine vision, the term feature (sometimes also attribute) is used to cover a larger field than in the psychological literature. Features can be local or global, they can be the answer of a filter or, given a segmentation, some sort of description of parts and their relations. Caelli et al. (1993) differentiate between morphological features, derived from the complete object, unary features, extracted from individual parts and binary features, describing relations between parts.

In this sense we will use the term "attribute", to describe the properties of parts and their relations.

4.2 Description of the CLARET-2 Algorithm

4.2.1 Introduction

The algorithm named CLARET-2 is derived from CLARET, short for Consolidated Learning Algorithm using Relational Evidence Theory, which was initially developed by Adrian Pearce as part of his PhD thesis (Pearce, 1996). It has been extensively modified by the present author for the purpose of modelling human object recognition. The procedure does pattern matching based on the principle of recognition by parts, that is patterns or object views are decomposed into parts and their relations. It relies on two techniques, graph matching and inductive logic programming. These two techniques have been widely used in the computer vision community for solving pattern

recognition tasks using recognition by parts.

Graph matching has found numerous applications in pattern recognition. They include character recognition (Messmer and Bunke, 1996; Lu et al., 1991), graphical symbol recognition (Lee et al., 1990; Jiang et al., 1999), 3D object recognition (Wong, 1992; Grimson, 1990) and others.

Graphs are suitable for the representation of structured objects. Vertices represent object parts, while edges model relationships between parts. The relations can be of spatial (e.g. distances) or other types (e.g. causal relations). The task of pattern recognition is solved using graph isomorphism. This approach seeks to find out whether two objects are the same. Subgraph isomorphism is established when two objects have not all, but only some parts in common. In this way parts and relations of an unknown pattern can be mapped to the parts and relations of a learned model. There are two drawbacks of relational graph matching. First, relational graph matching has an exponential computational complexity[3] and is therefore not feasible for typical object recognition tasks. Second, the ability to generalise is difficult to represent. Only recent developments, such as the graph edit distance (Bunke, 1998), the concept of a mean graph, the application of genetic algorithms (Jiang et al., 1999) and others, have significantly increased the ability for generalisation, for instance in case of errors due to noise or distortions, while providing computationally feasible algorithms.

Inductive logic programming in part based pattern recognition has been used to generate rules which generalise over the numerical attributes of parts and their relationships. In general, a mapping of parts and relations is not provided. Inductive logic programming has been applied in several real–world applications (Bratko and Muggleton, 1995). For example, a Chinese character recognition system based on the FOIL system (Quinlan, 1990b) has been developed (Amin et al., 1996), which is capable of learning to recognise large numbers of different characters. The Conditional Rule Generation (CRG) system, a symbolic and numerical relational learning system, has been especially developed for learning to recognise 3D objects from 2D grey scale views and complex scenes (McCane et al., 1998). Combined with a heuristic rule evaluation procedure (Scene Understanding by Rule Evaluation : SURE, Bischof and Caelli (1997a)) its use for 3D shape matching and inspection (i.e. tolerancing) has been demonstrated (Caelli et al., 1998).

CLARET-2 closely couples the process of mapping parts and relations – graph matching – with the process of generating rules which generalise over

[3]This is an NP–complete problem (Hopcroft and Ullman, 1979). The relation between computational effort and the number of nodes in the graph is at best exponential, for all known solutions.

numerical attributes of those parts and relations – inductive logic programming. This is done dynamically at run time which means, mappings and rules depend not only on the learned examples, but also on the unknown pattern.

In CLARET-2 classification is based on the similarity between an unknown pattern and learned patterns. A class is represented by a set of learned patterns, which constitute something like the prototypes for this class. A classification probability is derived from the probability that an unknown pattern matches one of the prototypes of the class. The question of selection of these prototypes belongs to a meta level and is not included in the CLARET-2 algorithm.

4.2.2 CLARET-2 Algorithm

Representation of objects and their views

Learned objects are represented as sets of example views. Each such view is decomposed into its components.[4] In general, the decomposition into parts, also called segmentation, is a non–trivial problem (Pal and Pal, 1993). A description is created by specifying the relations between such components. Relations consist of a number of attributes, which are typically numerical (e.g. distance, size, etc.), but could also be categorical (e.g. occlusion, etc.). The description of each view of an object is stored in the form of a graph (see Fig. 4.4), with vertices corresponding to labelled components and edges corresponding to the relations between them. Choosing the number of edges influences the complexity of the final description. Relations can exist between all components (fully connected graph) or only between some of the components, e.g. only neighbouring parts (partially connected graph). Also the number and types of attributes, which describe relations between components contribute to the complexity of a graph.

Matching an unknown view

An unknown view, represented by its graph, is matched to each of the learned example views to find the one it most probably fits to. The matching can be solved exactly, determining, whether a subgraph–isomorphism between the two graphs, exists. This method has two drawbacks, complexity and exactness. The problem of subgraph–isomorphism is NP–complete, meaning that the computation time increases exponentially with the size of the graph.

[4] In this application the components will be identical with the spheres, which form the objects used for experiments 5.1.

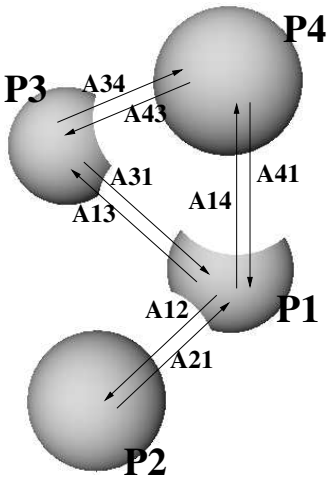


Figure 4.4: The construction of an adjacent graph. The view of an object (left side) is decomposed into its parts P1 to P4, the 4 spheres (right side). These parts represent the nodes, or vertices, of the graph. The nodes are connected through edges A_{xy} representing the relation A(x,y) between parts Px and Py. In general these relations are not symmetric, i.e. $A(x, y) \neq A(y, x)$. In this adjacent graph, only touching parts sharing a common border are connected. Adding relations between P2-P3 and P2-P4 would extend it to a fully connected graph.

This makes it necessary to find an approximation, which gives a good enough answer using less resources. The second reason, why some sort of approximation is desirable, is the fact that generally in object recognition exact solutions are of little value. There might be variations in the unknown view, errors or distortions. The sources of the variations can be noise, occlusion, lighting conditions, point of view, etc. For these reasons, we don't aim at an exact solution, but at an approximate one, where the probability of a fit is already high, if two views have, not necessarily the same structure, but a similar structure.

Determining similarity between views

Similarity of relations is expressed in terms of their distance in attribute space. Two relations are similar, if they have for instance nearly the same distance value and nearly the same size value, etc., with respect to a chosen resolution. Similarity is determined by partitioning the attribute space. Successively, a partition is divided into two new ones by splitting it perpendicular to an axis, thus gaining more information about the relations. Two relations

are similar at a certain level of resolution, if they are within the same partition of the attribute space. As the space is partitioned by more and more splits, the resolution increases with decreasing partition sizes. This process is dynamic, i.e., where the splits are, and how many there are, depends not only on the learned views, but it is different for every new unknown view. In this way the algorithm is trying find an optimal solution for every given task.

Mapping between parts of a learned and unknown view

Every partition provides information about the mapping between parts of the learned view and the parts of the unknown view. For instance, lets assume we find the relation $A(P1, P3)$ from a learned view and relation $A(X2, X3)$ from an unknown view within the same partition. In that case we can infer, that $X2$ can map to $P1$ or $P3$ and $X3$ can map to $P1$ or $P3$. As seen in this example, the mapping will be generally ambiguous in the beginning, being of the form m-to-n. As the attribute space is divided into more and more smaller partitions, the new information about the mapping of parts is combined with the existing knowledge. Thereby the initially ambiguous mappings are reduced in dimensionality, ideally arriving at mappings of the form 1-to-1 for recognised views, where learned parts uniquely match unknown parts.

Owing to influences such as changes in viewpoint or noise, a new partition, resulting from splitting an existing partition, may map parts in a way which is contradictory to the mapping already attained from the set of existing partitions. In that case the algorithm rejects such a partition as contradictory thus ignoring it in the further matching and refinement process.

The similarity of views expressed only by binary relations is high, if the fraction of similar relations is high within all partitions which provide a consistent part mapping.

Structural similarity expressed by chains of relations

Structure is expressed by building conjunctions of relations, which are just linear subgraphs or chains, taken from the graph representations. In that way new relations are formed, which are not binary anymore (i.e. between two components), but are of higher order, combining three or more parts. This is an approximation to the problem of determining a subgraph–isomorphism. The longer the chains get, the better the approximation might be. Similarity between views is again expressed in terms of the number of similar relations, now of higher order.

The dynamic nature of refining the matching

In the beginning, the matchings between an unknown view and learned examples will be very unspecific, starting with a very low granularity, or resolution, and only looking at very short chains. In the, most likely, case that no single matching is far better than all the others, the matching has to be refined. This is done by increasing the resolution in attribute space, leading to a finer discrimination between values, and by increasing the length of conjunctions of relations, thus capturing more and more of the graph structure.

In this process of refinement, learned views which are very unlikely to fit the unknown view are discarded.

The quality of a matching is defined in two ways:

- The probability of a fit, defined by the number of similar conjunctions of relations.
- The number of parts, both in the unknown and in the learned view, which can be mapped onto each other.

The refinement process stops, if a certain quality of matching is reached, for instance, when every unknown part has a corresponding known part, or a single hypothesis is much more likely, than all the others. This again shows the dynamic nature of CLARET-2, as only as many partitions and relational extensions are generated, as are necessary to reach this point for a given unknown view. Of course the refinement process also stops, if a certain level of complexity is reached, that is, when no more resources are available.

Example

In Figure 4.5 a simplified example is given for the processes of partitioning and relational extension, which constitute the main part of CLARET-2's algorithm. A single view of an object used in the later experiments (see p. 68) was chosen for this demonstration. The object is composed of four identical spheres, which are identified as the parts used here. In Figure 4.5(a) a segmented object view is shown with the spheres corresponding to the parts $P1 \cdots P4$ and arrows denoting the directed edges of the adjacency graph ($A12 \cdots A43$), where only neighbouring spheres are connected. For better readability every edge is labelled with only 2 attributes, distance and area.

Figure 4.5(d) is to be read from top left to bottom right and shows the attribute spaces with the 2 dimensions, distance and area, and the course of partitioning and extension, as the matching is refined.

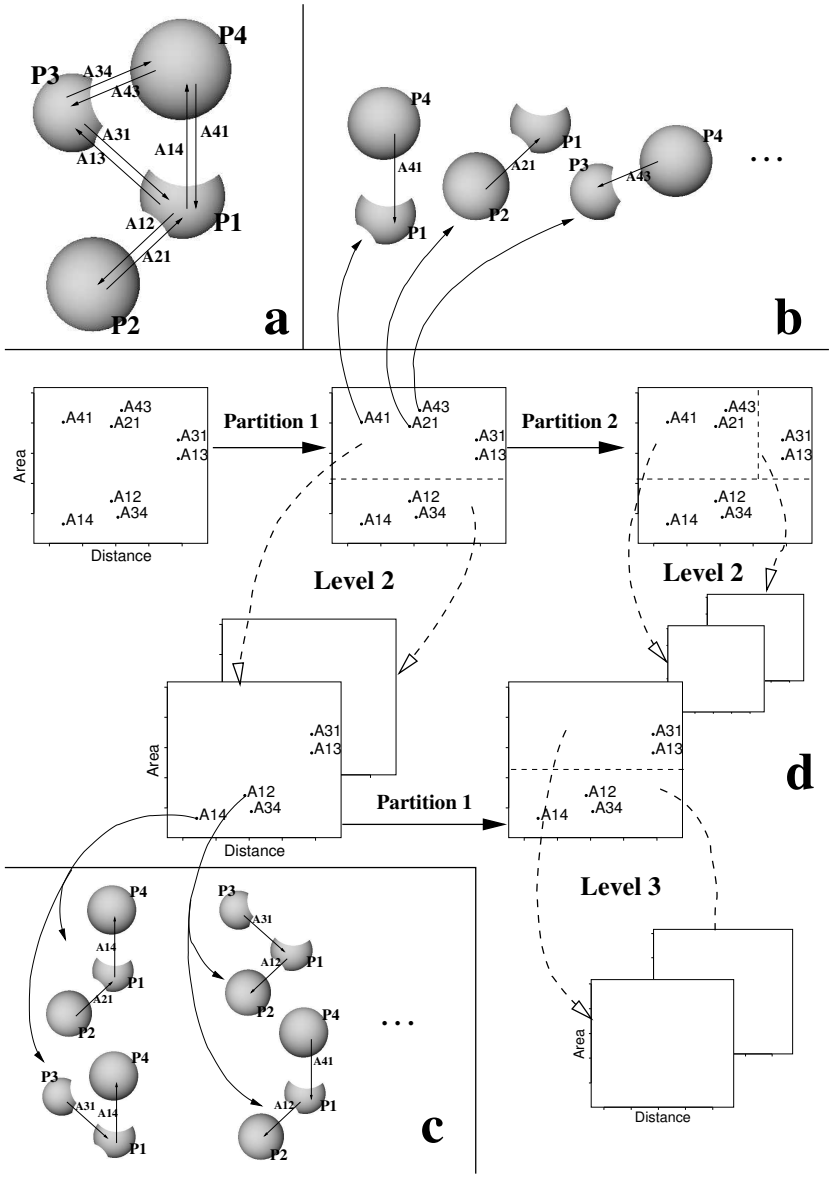


Figure 4.5: An example for the process of partitioning and relational extension employed in the CLARET-2 algorithm. Figure (a) shows an example input image, already represented as a graph; (b) shows extracts from the graph corresponding to points in the attribute space; (c) shows extracts from the input graph corresponding to points in an attribute space, which was relationally extended; (d) shows how the initial attribute space is processed, increasing the number of partition from left to right, and increasing the level of relational extension from top to bottom. For a more detailed explanation see the text.

In the beginning, every edge is represented as a point in the attribute space. Note, that for simplicity *only* the attributes of the example view are shown. In a real example, all the attributes of all learned views would also be present.

The first action is a partitioning on level 1 (partition 1, going right), splitting the initial attribute space in two halves. Figure 4.5(b) shows some of the "dipoles", the pairs of spheres, connected by a relation, belonging to one of the partitions. The next action is a relational extension (level 2, going down). This results in two new "initial" attribute spaces, containing chains of 3 connected parts. Examples of such chains for one of the attribute spaces is shown in Figure 4.5(c). Note, that each point in the attribute space again only represents the value of a binary relation, but now the multiplicity of each point can be larger than 1, since several different chains can end on the same relation (Fig. 4.5(c)).

The next step is again a partitioning of an attribute space. There are now 4 candidates, which could be split. One possibility is to further refine one of the two partitions on level 1 (partition 2, going right), followed again by a relational extension of the two new partitions (level 2, going down). Alternatively one of the two "initial" attribute spaces (partition 1, going right on level 2) may be further refined, followed by a further relational extension (level 3, going down). The latter would result in chains of 4 connected parts.

This cycle of partitioning and relational extension is iterated, until a matching is found, or until no resources are left.

4.2.3 Model Parameters

There are several factors which govern the properties of the CLARET-2 algorithm and influence the classification probabilities. These factors can be divided into two groups: (1) referring to the representation of the input data, that is the graph structure, and the attributes representing the object views, and (2) referring to parameters which directly influence the performance of the algorithm.

Graph structure

In principle, there are many ways to construct a graph representing the input data in the sense of how vertices are connected by edges. This means, between which parts of the object relations are computed.

A fully connected graph, where every part is connected with every other is the most complete representation. There is an obvious drawback, that a representation with many redundancies may result, where the number of

edges can become prohibitively high. This might cause problems concerning computing power.

An adjacency graph has edges only between neighbouring parts. In the case of our objects, two spheres which share a common border in a given view also share a relation with each other. The nodes representing these parts are connected by an edge. This approach results in a smaller number of edges, thus lower computational demands. Nevertheless, such a structuring of the input data may aid the algorithm in solving the classification task.

In our case, with only four parts present, the four spheres of the objects, complexity considerations are not necessary. Therefore only these two graph structures are investigated, fully connected graphs, representing the complete available input information, and the adjacency graphs, pre-structuring the input information in a, possibly, meaningful way.[5].

Attribute values

Several attributes are computed, describing the relations between object parts. All attributes possess continuous values, there are no categorical attributes, such as "is occluding", or "is connected to". Further, most attributes are asymmetric, i.e. in general $A(i,j) \neq A(j,i)$. For a complete list of the available attributes and the details of their computation see Sec. B.1.2.

It has to be stressed, that *all* attributes are binary relations, this means they always depend on two parts (at least formally). On first sight, this seems counterintuitive, because properties of parts like size, brightness, etc. seem to be of unary nature. But a closer look reveals, why it is useful to consider them as binary. What is the meaning, if part x has a size of $s = 43.7$? This value has no meaning at all, if it can't be related to values of other parts. One way to do this is by normalising the input pattern, that is normalise the scale, position, orientation and brightness of a view. This requires some computation and may not lead to a unique solution. To normalise correctly, the knowledge of what is normalised already needs to be given a priori. Or it might be simply impossible to normalise, if complex scenes containing several objects are viewed. In any way, normalisation leads to an implicit relation of a part with all the others. The other possibility, which is more general and requires less computational effort is to make this relation to other parts

[5]It is also possible to set upper and lower bounds for the valency of the vertices in the graph, that is the number of edges originating at each vertex. This allows finer control over the size and structure of a graph and can make the problem computationally feasibly by reducing the complexity of the data. The drawback here is that the resulting structure of the graph has no "meaning", as is the case with an adjacency graph or a fully connected graph

explicit, using binary relations, as is the case in CLARET-2. Depending on the construction, this implies invariance in scale, position and orientation of the object views.

Immediately, the question arises, of whether descriptions should be restricted to binary relations. The only reason for CLARET-2 to do so is computational efficiency. Already the number of combinations, using only an attribute space with binary relations, is too large to be explored completely and optimisation and pruning is needed. Introducing more attribute spaces of higher order, which need to be handled separately, would only aggravate this situation.[6]

The attributes computed were divided into two groups, one containing mainly 2D intensity based, the other mainly 3D range based attributes.

2D Attributes These are computed using only the projection of the objects on the computer screen, the same views the human observers are presented with during learning.

There are attributes describing spatial relationships (distance, angle, border-length), relative measures of shapes of object parts (area, maximum span) and relative measures of intensity of parts (mean intensity, variance of intensity).

3D Attributes These are computed using a range image, that is, an image containing depth values, constructed from the rendered projections of the objects, as the observers see it during learning.[7]

Again there are attributes describing spatial relationships (distance, angle, border-length), relative measures of shapes of object parts (area, maximum span), but now these attributes are computed for the 3D range image, not its 2D projection. Attributes describing intensity properties can of course no longer be computed. Instead, attributes explicitly expressing depth relations (mean depth, variance of depth) are computed.

[6]There is a way around this limitation to a binary attribute space, if higher order attributes, such as an angle, are needed. See Chap. B.1.2.

[7]Note that the use of such a range image is only for convenience. The same depth information could be constructed using a projection image, from the knowledge that the objects are composed of four identical spheres and the fact that a perspective projection is used. By estimating the apparent radius of a sphere from the curvature of its boundary the depth values of its visible surface can be computed.

CLARET-2 program parameters

The following parameters governing the behaviour of CLARET-2 were varied during the simulation runs:

Length of chains By varying the maximum length chains taken from a graph can have, it can be controlled how exactly a subgraph isomorphism is approximated. In other words, how exact the structural match between two graphs is determined.

Resolution By varying the resolution or granularity, it is determined how many relations are similar to each other. The higher the resolution, the more specific is the distinction between relations.

Memory capacity By limiting the memory capacity it is determined how many groups of relations can be distinguished at one time. This factor interacts with the length of chains and the resolution. The longer the chains are and the higher the resolution, the more different groups of relations need to be stored in memory.

Probability measure It was mentioned earlier that two views are similar, if the number of similar relations is high. The probability measure is a parameter used to quantify the term "high", it describes the probability of differences occurring between views. The lower the probability measure, the higher the required similarity between two views.

4.2.4 Differences to the Original CLARET Algorithm

The original algorithm has been changed considerably, thus warranting a change of name. The complete source code now consists of over 45000 lines of C Code.

1. The procedure of matching learned to unknown parts and checking the compatibility of a new partition with existing partitions was debugged and optimised. The original program was not invariant to the exchange of learned view and unknown view. The complexity was reduced from $O(N^4)$ to $O(N^2)$, making the program much faster.

2. Originally, the quality of a match was determined by an heuristic measure, adjusted to work well by experience. This was changed to the estimation of a matching probability, based on clear assumptions of independence.

3. The task of graph matching is NP-complete, the time required for solving it depends exponentially on the number of parts. Thus it is necessary to take a "shortcut", to estimate the correct solution. Originally, this was done by ordering the learned views by their matching probability, always partitioning exclusively the most likely matching, until the limits of the resources were reached. In extreme cases this could have the effect that actually only a single learned view was investigated while the others were ignored. Nevertheless this explored view wasn't necessarily the correct match. The new algorithm also maintains a list of the matchings between the learned views and the unknown view, but in every iteration all of them are partitioned and the matching probabilities are recomputed. The most unlikely learned views, where the matching probability drops below a set limit, are removed from the list of possible matches.

4. The part of the program creating the graph structure from the input data was moved into a separate program, debugged and modified to allow a more fine grain control over the complexity, i.e. the number of edges per node, of the resulting graphs.

5. Major parts of the code were refactored and optimised for speed.

Chapter 5

Learning 3D Object Representations

5.1 Psychophysical Experiment

5.1.1 Introduction

Visual object recognition has been studied in the past in a multitude of different settings with variations in experimental paradigm, the types of stimuli used, and the tasks to be solved.

This study uses a classification task, since "there is nothing more basic than categorisation to our thought, perception, action, and speech." (Lakoff, 1987, p.5). Thus we move away from tasks such as delayed matching to sample, discrimination, etc., where it is in general not necessary for the subject to access the representations of several objects at the same time. It is sufficient to store a sample of the target (as in delayed matching to sample), or to memorise one object only, to determine whether the target stimulus is same or different. The minimal number, that forces the subjects to do "real" categorisation is classifying the views of three objects. Using only two objects, the experiment may degenerate to a discrimination task.

Studying the process of how representations of objects are learned to allow categorisation requires the use of novel objects. This also removes a confounding linguistic influence on the visual recognition, the possible naming of the objects. Learning is partitioned into repetitive learning units of alternating supervised learning with a control test.[1] The subjects learn until

[1] In the machine learning literature supervised learning means that the class label of the learning sample is known in advance (Duda and Hart, 1973). Sometimes the complete procedure of repeating units consisting of supervised learning followed by a test, is termed 'supervised learning' in psychophysical literature (Caelli et al., 1987; Rentschler et al.,

they achieve a set criterion, thus assuring that all subjects have attained the same learning status in regard to the presented object views and as measured by the used test[2]. This allows the comparison of the dynamics of learning as well as of the performance in ensuing generalisation tests.

As Vanrie et al. (2001) point out, the process of recognising objects – especially whether it is viewer centred or not – is influenced by the type of stimuli and the paradigm used. The authors hold the view that it would be of great interest to produce different behavioural response patterns without fundamental changes in variables like paradigm and stimuli. This is the motivation for preceeding the actual visual learning process with a (optional) priming phase. Leaving the actual learning process constant for all subjects, differences in performance can be pointed to the influence of the presence and type of priming knowledge. We expect that priming the subjects by allowing them to actually handle 3D-models of the objects facilitates the formation of viewpoint independent internal representations. To enhance this effect, two of the objects used were designed to be mirror-symmetric, since handedness tasks have been demonstrated to be, in general, viewpoint dependent (see citations in Vanrie et al., 2001, p.1049) . Since some previous studies maintain the existence of a strong influence of object structure on behavioural responses (Biederman and Bar, 1999, see 3.2), objects with clearly defined structural differences were chosen, which are nevertheless very similar to each other, keeping the categorisation task from becoming trivial.

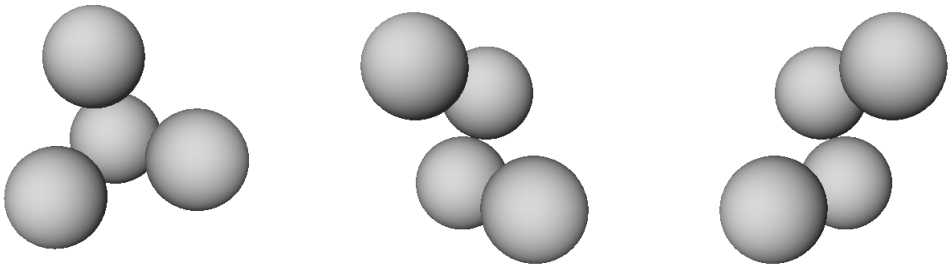
5.1.2 Methods

Subjects were 53 adults, about half of which were graduate students. All subjects reported normal or corrected-to-normal visual acuity and having no other visual disorders. Ages ranged from 25 to 45 years. The subjects were randomly assigned to 3 experimental groups, which were approximately balanced referring to age and gender.

The stimulus set consisted of three objects, constructed and displayed on the 17" screen of a SGI O2 computer using the Open Inventor software package. Each object was composed of four spheres, with three of them forming a rectangular isosceles triangle and the fourth being placed perpendicularly above the centre of one of the base spheres (see Fig. 5.1). Object 1 possesses a three-fold rotational symmetry as can be seen in Fig. 5.1(a). In the presented view the symmetry axis is perpendicular to the projection plane. A rotation in plane of 120° transforms the object into itself. Objects 2 and 3

1994; Unzicker et al., 1998)

[2]The learning status is sometimes not controlled in other studies (see e.g. Bülthoff and Edelman, 1992; Newell et al., 2001)



(a) Object 1 (b) Object 2 (c) Object 3

Figure 5.1: Objects used in the experiments on learning and recognition. Each object consisted of four spheres, with three of them forming a rectangular isosceles triangle and the fourth being placed perpendicularly above the centre of one of the base spheres. Note that object 2 and object 3 are mirror symmetric to each other.

are mirror symmetric to each other (Fig. 5.1(b), 5.1(c)). Furthermore these 2 objects posses a two-fold rotational axis, which, in the presented views, is again perpendicular to the picture plane. A rotation in plane of 180° carries object 2 into itself, and the same applies to object 3. 2D views were generated as perspective projections of the objects onto the screen plane of the computer display.

To generate the set of learning views, azimuth and elevation were sampled in 60° steps; the equatorial plane was horizontal and contained the symmetry axis of each object; the centre of the coordinate system used for sampling was situated at the centre of gravity and thus on the symmetry axis of each object; the zero views (presented in Fig. 5.1) were chosen such that the symmetry axis was located perpendicular to the picture plane and no centre of of any sphere lied exactly in the equatorial plane, to exclude occlusion artefacts; for each view an additional arbitrary rotation in plane was added, to reduce motion sequence effect; this rotation in plane could take the values 0°, 60°, 120°, 180°, which remained fixed; except for the afore mentioned arbitrary rotation in the picture plane, for every view of object 2 there exists a mirror view of object 3 and vice versa; views redundant due to rotational object symmetry were eliminated. This process is visualised in Figure 5.2. It resulted in 22 views in total (6 views for object 1, 8 views for object 2 and object 3 each, see Fig. 5.3). At the viewing distance of 1m, the images appeared under a visual angle of approximately 7°.

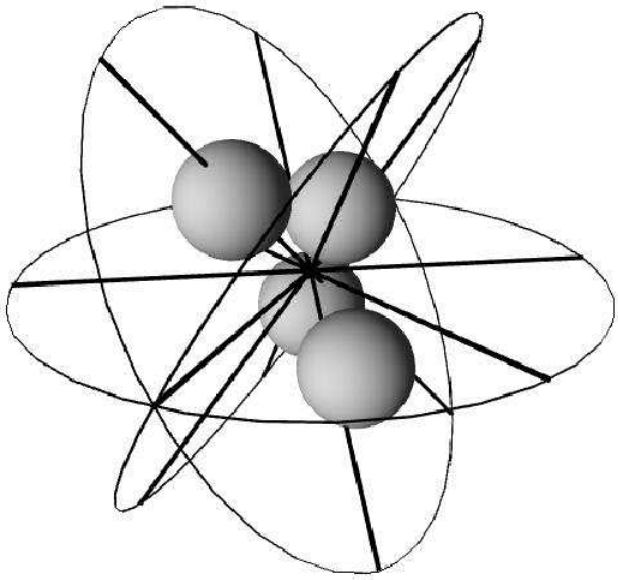


Figure 5.2: Visualisation of the 8 viewing directions on the viewing sphere. Spacing for azimuth and elevation is 60°, respectively. The coordinate system was rotated in this figure for better clarity

For haptic exploration physical object models were constructed using styrofoam balls of 6cm diameter.

See the following page for the 22 learning views in Fig. 5.3.

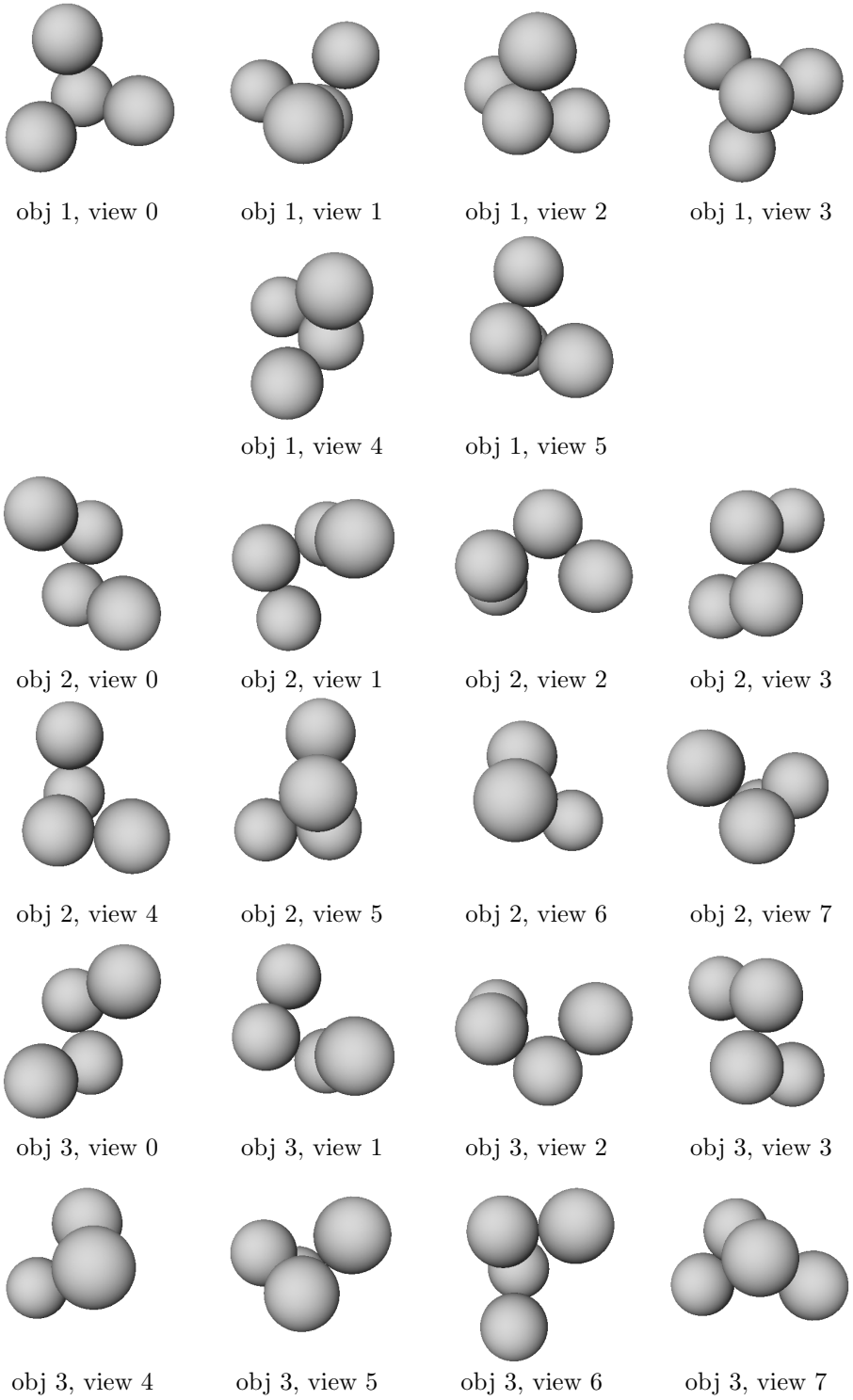


Figure 5.3: The set of 22 learning views

Object Priming

All subjects were informed about the task of learning to classify the set of 22 learning views. They were further told that the views depicted 3 objects, each constructed of 4 identical spheres, which were connected in 3 different configurations. Two of the three subject groups had the opportunity to gain

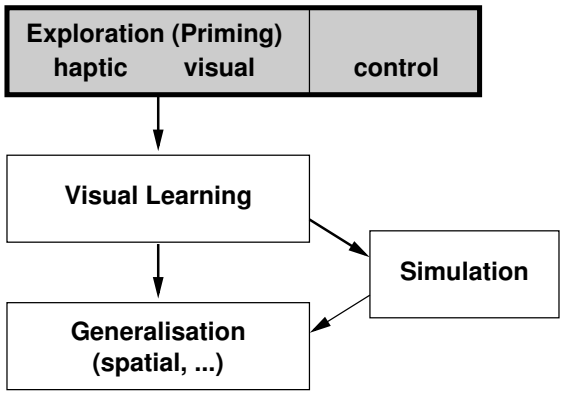


Figure 5.4: Different priming conditions allow to investigate the impact of the type and amount of prior knowledge. With stimulus input and the learning procedure kept constant, differences in recognition performance between groups can be attributed to differences in representation due to priming.

further prior knowledge by exploring the objects for about 5-10 minutes. These are the haptic priming and the visual priming conditions. The third subject group served as control condition and had no additional information or priming. Exploration, respectively instructions, were immediately followed by a procedure of supervised learning (see below).

Haptic Priming Condition Subjects were seated blindfolded at a table where they were handed styrofoam models of the three objects. With each object its number was announced. The object number was repeated later if requested. The subjects were allowed to put the objects on the table and pick them up again. They were also allowed to handle several objects at a time if they wished to do so. Exploratory hand and arm movements were in no way restricted; on the contrary, subjects were encouraged to explore the objects from all sides.

Visual Priming Condition The objects and their corresponding object numbers were simultaneously displayed on the computer screen as 2D projections ("views"). The subjects could "grasp" one of the object views at a

time by means of the computer mouse and a cursor. They could then rotate each object with three degrees of freedom and inspect it from any desired viewing direction.

Supervised Learning Procedure

The procedure of supervised learning was partitioned into "learning units" consisting of a learning and a test phase (see Fig. 5.6 for an illustration). During a learning phase, all 22 stimuli were presented sequentially in random order, each followed by their respective object number. During a test phase, all stimuli were presented again in random order. The observer was then required to assign an object number to the test views, using a computer keyboard.

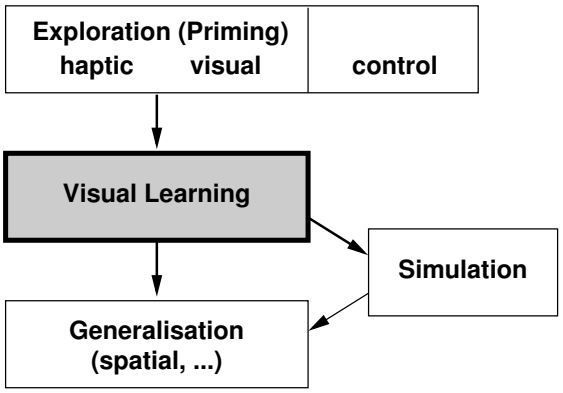
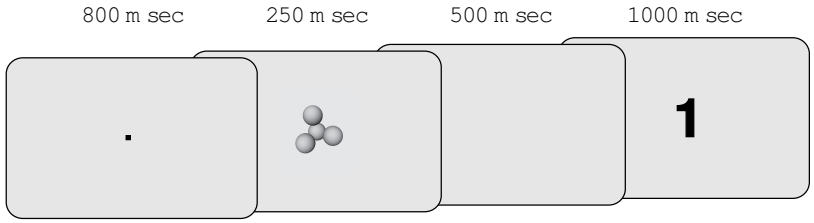


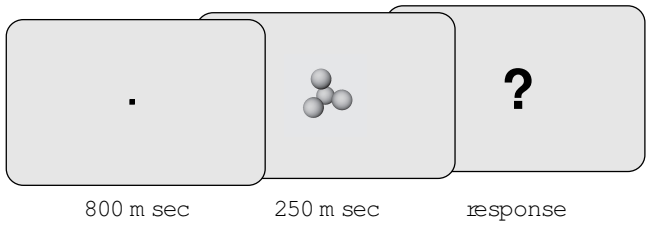
Figure 5.5: Visual Learning happens within a context of supervised learning. A presentation phase of the 22 learning views labelled with their object number is alternated repeatedly with a test phase, where subjects try to recognise the previously seen views. Learning is ended successfully when the subject reaches 90% correct answers.

The observer's answers were recorded in a "classification matrix". For each learning unit such a classification matrix was recorded, with the rows denoting the single views and the columns the 3 possible answers, object 1, 2 or 3. The recorded classification matrices for a complete learning session were cumulated, resulting in an matrix containing the relative answering frequencies per view. Examples for both types of matrices can be found in Appendix C.6.

Summary feedback of the subject's performance was given at the end of each learning unit, informing about the total and the per object percent correct answers. No other feedback was given. The stimulus display time was



(a) Learning phase



(b) Test phase

Figure 5.6: Procedure of supervised learning. During learning phase (a) presentation of a fixation cross is followed by an object view and, after a short pause, the number of the object. In this fashion all 22 learning views are presented in random order. During test phase (b) after presentation of a fixation cross and an object view, the subject is required to enter the number of the object shown. Again the 22 learning views are presented in random order. These two phases are repeated until subjects achieve the set criterion of 90% correct answers (20 of 22 views).

$250ms$. The learning units were repeated until subjects reached a criterion of 90% correct answers. This means that 20 out of the 22 views had to be classified correctly. Subjects were allowed to take a break after a learning unit was completed, if they wished to do so.

Subjects learned in sessions, which lasted about 1h – 1.5h, depending on subjective fitness and the number of breaks during learning. Subjects, which did not achieve the learning criterion within a single session, commenced learning on another day.

5.1.3 Results

From the 53 participating subjects, 2 did not reach the criterion set at 90% correct answers. One subject of the latter two did not have the time to finish the experiment for personal reasons, the other subject wished to abort the

experiment after 3 sessions containing 65 learning units. The results of these two were excluded from the data set. All other 51 subjects achieved the set criterion within a maximum of 4 sessions. As shown in Figure 5.7, learn-

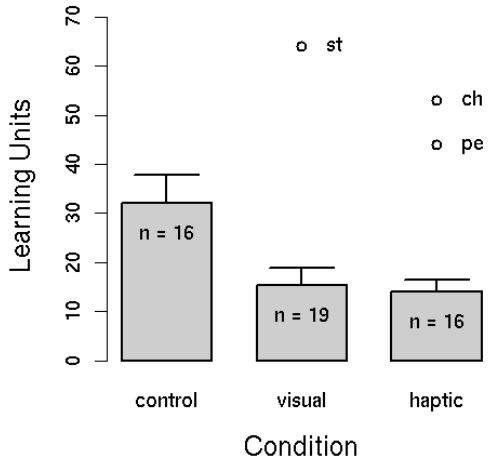


Figure 5.7: Learning time required to reach the 90% correct criterion (20% trimmed means, see Sec. C.1). Three outliers were detected. Learning times for primed subjects are significantly faster than for subjects from the control group ($p < 0.05$).

ing time, as measured by the number of learning units[3] necessary to reach the criterion, was affected by providing prior knowledge. Despite the short duration of priming, lasting about 5-10 min, learning time was drastically reduced by about 50%. Control subjects needed about 32 learning units, whereas visual subjects with about 15 units and haptic subjects with about 14 learning units were both significantly faster ($p < 0.05$). Note that haptic subjects needed even slightly less time than subjects under the condition of visual priming.

Subject's reports on learning strategies

On finishing the learning experiments, the subjects were interviewed as to the learning strategies they had employed. Mainly two strategies were reported:

Learning by Heart The subjects learned by heart which object number was associated with a given view.

[3] Analysing the learning duration not in terms of the number of units, but by measuring the absolute time required (in minutes), yielded similar results.

Mental Rotation Subjects rotated a given view to match it with an internal model of one of the three objects.

Subjects with haptic experience were most consistent in their strategy. With two exceptions, which were both outliers, all subjects reported solving the learning task mainly by mental rotation. Those views however, which were difficult to rotate (e.g. view 6 of object 2 and view 4 of object 3, see Fig. 5.3) were mostly memorised.

Subjects with visual prior knowledge were less consistent in their strategies. Most of them solved the task by mental rotation, but started using this strategy at varying points during the learning process. Some of the subjects solved the task purely by memorising.

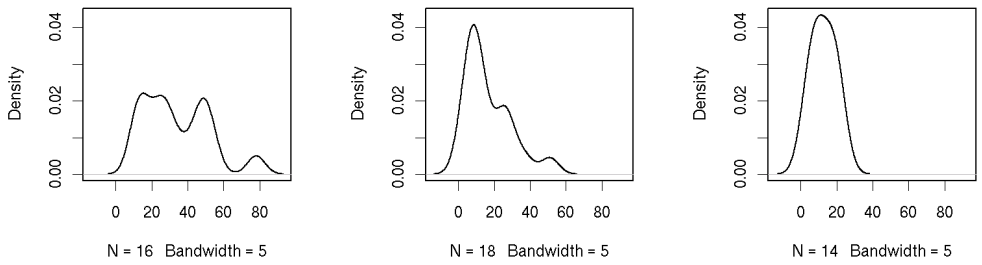
Within the visual group there was one exceptional subject which solved the task in a single learning unit. The subject used a variation of the mental rotation strategy, imagining herself at the object's position and then rotating herself. The subject was able to describe the three objects perfectly.

Control subjects with no prior experience were the most inconsistent group. Most of them learned the views by heart, but tried to employ various strategies during learning. Whereas some only memorised and stuck to this method, other subjects used various strategies to try and 'make sense' of the views. Some of the subjects were able to successfully rotate at least a subset of the views.

Figure 5.8 shows estimated density functions, essentially smoothed histograms, for the distribution of learning durations for the three experimental groups. It reveals differences between the groups which were, especially for haptic and visual subjects, hidden by looking only at the mean values. The number of learning units completed is at the abscissa. Clearly, the haptic group is the most consistent one, showing the least variance of learning duration around a peak value. The distribution for the visual group also has a clearly defined peak value, but shows more variation, especially towards longer durations. By contrast, control subjects show the most inhomogeneous distribution, with a very broad peak at around 20 learning units and another peak at about 50 units.

Outliers

Using the standard box-plot method (Chambers et al., 1983), three outliers were detected, which are also identified in Fig. 5.7. These were 2 subjects with haptic and one with visual prior knowledge. Asked about the learning strategy they had used, the subject with visual prior knowledge, subject *st*, reported learning the set of views by heart. But unlike other subjects an



(a) Control (b) Visual (c) Haptic

Figure 5.8: Estimated densities of the learning times. The three plots show the data for control, visual and haptic subjects. Outliers were removed.

additional mnemonic technique was used, associating a view with a similar image (e.g. "This view looked like the roof of a house"). That is, this subject did not associate the views with a model of the physical object, but rather with a completely unrelated real world image or object.

The next outlier, one of the subjects with haptic prior knowledge, subject *ch* reported finding the task very difficult and learning most of the views by heart. The subject was not able to draw a sketch of the objects and could only describe object 1 correctly. The other subject with haptic experience, subject *pe*, stated in the final interview not feeling confident about the task. Only during the second session did the subject actually believe that it was possible to solve the task. The subject began trying to rotate the objects and then continued with other strategies, such as learning by heart. Finally, the subject solved the task by mainly relying on mental rotation.

Per View Analysis of Cumulated Learning Matrices

The differences between groups can also be analysed on a 'per-view' basis, comparing the cumulated classification matrices, which provide relative answering frequencies. This was done by averaging the cumulated classification matrices over the subjects belonging to the respective experimental groups. A measure for the difference between two such matrices is the sum of the squared differences between the elements (see Appendix C.2). A t-test was used to determine whether one matrix was 'better' than another, i.e. whether the fraction of correct answers was higher (see Appendix C.4 for details).

The results, which are numerically summarised in Tab. 5.1, show haptic subjects being significantly better than subjects with visual prior knowl-

group 1	group 2	D	df	T	C(0.05)	P(equal)
Control	Visual	0.031	65	8.93	1.67	$p < 0.001$
Control	Haptic	0.036	65	7.40	1.67	$p < 0.001$
Visual	Haptic	0.022	65	2.01	1.67	$p < 0.05$

Table 5.1: Analysis of behavioural data for learning using a t-test. Comparison of controls with visual priming group, of controls with haptic priming group and of visual with haptic group. D denotes the difference measure between the two matrices; df the degrees of freedom (3 answers for 22 views results in 66 values being compared); T the computed t-test value; C(0.05) the critical value for a 5% level, taken from Student's t-distribution; P(equal), the probability that the matrices are from the same distribution, which is derived from Student's t-distribution using the computed value T.

edge and both performed significantly better than subjects without prior knowledge. The distance between visual and haptic matrices was about 30% smaller than the distances between each of the two and the control matrix.

5.1.4 Learning Dynamics

For every subject three matrices were sampled from the whole set of learning matrices computing a weighted mean (see C.5), centred at the begin of the learning procedure ($x_0 = 1$), after half the learning time ($x_0 = N/2$) and at the end of learning ($x_0 = N$). These sampled matrices were averaged over all subjects within the same group. With 3 groups – controls and visual and haptic priming – and 3 points in time – begin, middle and end – this resulted in a total of 9 matrices.

For a summary, displaying only the dynamics of the object means, see Fig. 5.9. The filled outer symbols show the location of the means for a perfect classification matrix, with no wrong answers. All three groups make most errors mistaking the mirror symmetric objects 2 and 3. Subjects from visual and haptic priming groups begin recognising object 1 quite well, there is little improvement over time. The most pronounced improvement can be found in discerning the objects 2 and 3. Control subjects on the other hand also begin making most mistakes between objects 2 and 3, but additionally their performance for object 1 is relatively low. They show a much higher improvement for object 1 than the other groups. In general, already at the beginning of the learning process both the groups with priming experience perform clearly better than the control group.

A more detailed graph (Fig. 5.10), showing the relative answering frequencies for each group at the three points in time, confirms these findings.

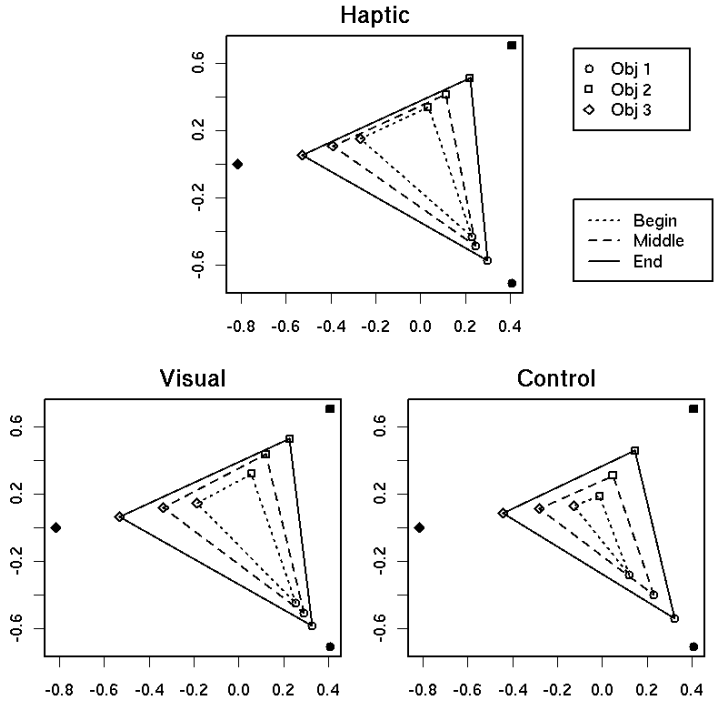
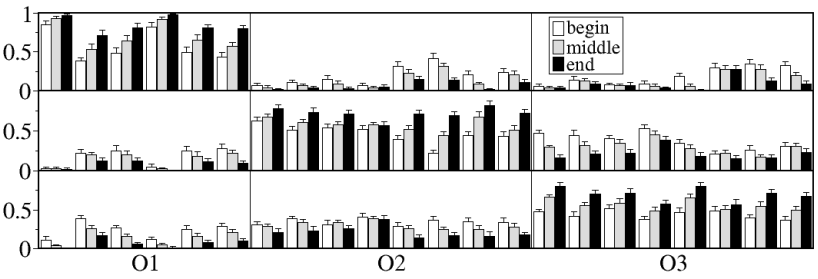


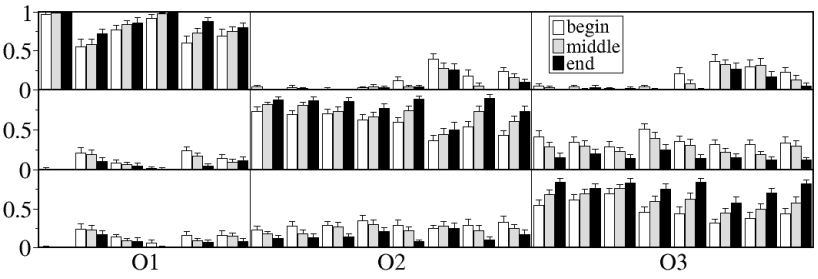
Figure 5.9: Learning dynamics, showing 20% trimmed means (see Sec. C.1) for each object, located on the $x + y + z = 1$ plane, at begin, middle and end of learning phase. Filled symbols represent the locations of perfect answers for comparison.

As with the cumulated classification matrices (see 5.1.3) a statistical t-test was applied, comparing the groups at each point in time and also comparing the matrices within each group at different times. The summary of the results in Tab. 5.2 shows, that again both haptic and visual subjects performed significantly better than control subjects without prior knowledge. Also haptic subjects were always better than visual subjects, at the beginning even significantly so. The distances decrease with time. This makes sense, since all subjects reached the criterion of 90% correct answers in the end. Also the difference between visual and haptic subjects is again much smaller, than the distance of either of them to the control group, as was the case for the cumulated matrices.

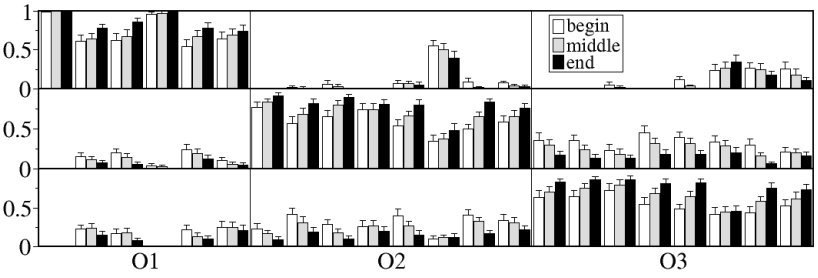
Applying a t-test within each group at successive times, shows a significant improvement at every learning stage (Tab. 5.3).



(a) Control group



(b) Visual priming group



(c) Haptic priming group

Figure 5.10: The three graphs visualise the answering matrix. They show the mean relative answering frequencies for controls and the priming groups for all 22 learning views. The top third of each graph depicts how often subjects chose 'object 1' as their answer. The middle third of each graph shows the relative frequencies of answer 'object 2' and the bottom third of 'object 3'. The graphs further discriminate between the performance at the begin of the learning phase, during the middle, and at the end of the learning phase. It can be clearly seen, that for all groups the number of correct answers (the 'diagonal' of each graph) increase over time, whereas the number of wrong answers decreases.

Control – Visual		
time	D	T
begin	0.040	7.81
middle	0.032	8.02
end	0.029	7.65

Control – Haptic		
time	D	T
begin	0.045	7.80
middle	0.037	7.15
end	0.035	6.08

Visual – Haptic		
time	D	T
begin	0.027	2.77
middle	0.026	1.73
end	0.020	1.56

Table 5.2: Differences between groups at begin, middle and end of learning phase. Distances D are squared differences between matrices. The results T of the t-test are given, with $df = 65$ and the critical value $C(0.05) = 1.67$.

Condition	time	D	T
Control	begin – middle	0.036	11.8
	middle – end	0.040	15.3
Visual	begin – middle	0.031	13.7
	middle – end	0.037	13.1
Haptic	begin – middle	0.027	8.87
	middle – end	0.033	13.2

Table 5.3: Differences between learning matrices within groups in time. D denotes the distance between matrices, T the value of the t-test, with $df = 65$ ($df = 61$ for haptic priming group) and the critical value $C(0.05) = 1.67$.

5.2 Simulation using CLARET-2

5.2.1 Introduction

CLARET-2 is an algorithm which represents objects by multiple views, but utilises structural descriptions consisting of object parts and their relations. It provides a measure of similarity between views by determining their corresponding parts and the probability of such a correspondence. CLARET-2 further allows to determine the structural complexity of simulated representations and the saliency of the attributes employed. A more thorough introduction to the CLARET-2 algorithm and a discussion of the possible reasons for using it as a tool for modelling visual object recognition are given in Sec. 4.1.

It has been pointed out by several authors that the distinction between image-based and structural description models might not be the real issue anymore in object recognition, since both can be tuned to display viewpoint-dependent or viewpoint-independent behaviour (Hummel, 2000). Furthermore psychophysical evidence suggests that the respective behaviour depends on the stimuli used and the task set (see Sec. 3). Instead questions arise as

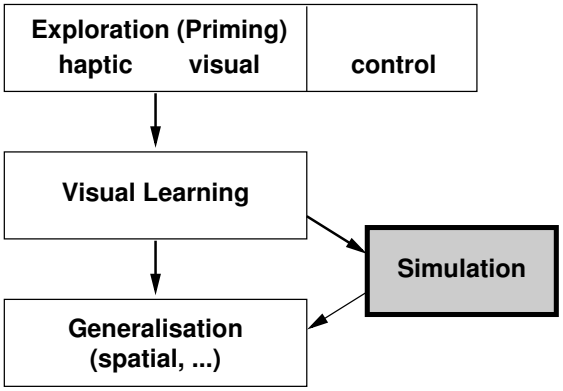


Figure 5.11: During simulation the parameters of CLARET-2, a machine learning algorithm, are tuned to achieve the same recognition results as the subjects of the priming groups and the control group. The parameters are interpreted to supply predictions for generalisation experiments.

to how to measure similarity between objects and determining the correct feature set to achieve this (Tarr and Vuong, 2002).

The construction of the stimuli used here suggests a 3D structural description as the best mode for optimal performance. Nevertheless subjects might adopt different strategies for solving the visual recognition task. An algorithm like CLARET-2, relying both on multiple views and a structural description, seems a promising approach for modelling human behaviour. In our view the specific nature of the algorithm makes it at the same time suited for answering the questions of optimal feature sets for determining similarity. The assumptions here being that (1) a representation depending on multiple views is *in principle* powerful enough, (2) the method of segmentation produces parts which are meaningful for modelling purposes, and (3) the attributes chosen to represent the parts and their relations can be mapped to the features actually employed by the human observers with no key attribute missing.

5.2.2 Methods

Modelling was based on the same set of 22 learning views (Fig. 5.3), on which the human observers were trained. Feature extraction was based on intensity images for 2D attributes and on depth images for 3D attributes[4].

[4]Depth images, where the value of each pixel denotes the distance of the corresponding surface point to the viewer, were only used as a convenient way of calculating depth attributes. The same information could have been derived from the intensity images using

It is assumed that the segmentation into parts is given by the component structure of the objects so that every sphere corresponds to a distinct part. Attributes were extracted, describing the relative size and form of the visible surfaces of each sphere, the relation in distance and angle to other spheres, and a summary description of the relative brightness (2D) and the relative distance from the viewer (3D) of each surface patch. A detailed account of the list of attributes and how they were computed is given in Appendix B.1. The information collated thereby was stored in a corresponding graph structure, resulting in 22 graphs, one for each view. Within each such graph the vertices denote the segmented parts, i.e. the single spheres[5], and the edges denote a binary relation between two parts, containing the stored attribute values. These 22 graphs constituted the set of learned views against which in turn each one was presented as an unknown view, in analogy to the testing part of the supervised learning procedure adopted for the human observers. The result of applying CLARET-2 are probabilities of classification for each of the 22 views, which can be arranged in a prediction matrix, similar to the classification matrix of an human observer. The quality of fit was determined by computing the root mean squared error (see App. C.2) between the predicted classification matrix and the observed cumulated classification matrix, using the mean matrices for each group as described in Sec. 5.1.3.

Three factors influencing CLARET-2's performance were varied in these simulation experiments.

- The attributes.
- The program parameters
- The graph structure.

Attributes

The question of investigating attributes likely to be used by human observers is handled in detail in the descriptions of the individual simulation experiments below.

Program Parameters

After preliminary investigations, the following parameters were varied within the given value sets.

---

the information given to all subjects, that the objects consisted of four identical spheres – implying equal radius.

[5]Although each node is labelled with the number of its corresponding sphere, these labels are *not* used directly to match the parts of a learned and an unknown view

Chain length	$1, 2, 3$
Resolution	$1, 2, \cdots, 10, 12, 15, 20$
Memory	$5, 10, 15, 20, 30, 40, 60, 80, 100, 120$
Probability	$0.01, 0.02, 0.05, 0.1, 0.3, 0.8$

For each graph structure and attribute set the full combinatorics of these values was applied. For a detailed explanation of the parameters see Sec. 4.2.3.

Graph Structure

For the graph structures (see Sec. 4.2.3 for more explanations and App. B.1.2 for construction details), two obvious possibilities of representing the object information within each given view were investigated. Since only four object parts (the four spheres) were present, a finer differentiation would not have made sense.

Full Fully connected graphs, where every part is connected with every other. With asymmetric relations and 4 parts (i.e. vertices) present, such a graph contains 12 relations (i.e. directed edges).

Adjacent Adjacent graphs, where only relations exist between neighbouring parts, that have a common border in the given view. Typically, such a graph contains about 6 relations.

5.2.3 Simulation Experiments

To summarise, the following assumptions are made here: (1) The segmentation into parts as suggested by the structure of the objects is meaningful for the representation formed in the human observer; (2) The attributes extracted don't miss any key features used by the human observer; (3) The graph structure based on multiple views can be mapped to the representation of the human observer, which may be based on an internal 3D-model; (4) The cumulated classification matrix, a mean over all answers given during learning, is a meaningful measure of the state of the observers representation at the end of learning, when reaching the set learning criterion.

Experiment 1: Investigate 2D vs. 3D Attributes

This first experiment had two aims. First to investigate whether a preference for either 2D or 3D attributes can be found. This was motivated by the discussion in the literature concerning the nature of the internal representation of 3D objects. The two main positions are that the representation is either 3D object centred or 2D view dependent, which should also reflect in the

kind of attributes used to represent the objects. To this end, the available attributes were grouped in 2 sets, attributes which are mainly 2D and attributes which are mainly 3D in nature. 2D attributes are computed using only the projected image of the objects, which are the same views as the human observers learned. 3D attributes are computed using a range image containing depth values of every image pixel[6].

2D Attributes	3D Attributes
distance	distance
angle	angle
area	area
span	span
border	border
mean intensity	—
variance intensity	—
—	depth
—	variance depth

For a more elaborate description see Sec. 4.2.3 and for a detailed discussion of the chosen attributes and their computation see App. B.1.2.

The second aim was to investigate the complexity of the internal representation. This depends, of course, on the parameter values. The more resources available, the more complex a representation can be. The complexity has a second aspect, though, which does not depend on the model algorithm proper, but on the structure of the input data. In our case the graph structure representing each view can be either a fully connected graph or an adjacent graph, where only neighbouring parts are connected.

CLARET-2 was run, applying the full combinatorics of the program parameters, using both fully connected graphs and adjacent graphs, with either 2D or 3D attributes.

Experiment 2: Investigate Individual Attributes

To explore the nature of internal representations in more detail, the question arises, what the minimum and the optimal number of attributes necessary for modelling are and which attributes give the best results in those cases.

Again CLARET-2 was run exploring the same parameter ranges as in experiment 1, using both a fully connected and an adjacent graph. Here, however, the number of attributes was reduced. Instead of using 7 attributes (2D or 3D), three cases were investigated.

[6]The use of a range image is for convenience only, see footnote p. 79.

Single One attribute was chosen, investigating all possibilities of selecting from the complete set of 14 attributes.

Double Two attributes were selected. All possible combinations were investigated, taking both from the full set of *either* 2D or 3D attributes.

Triple Three attributes were selected. All of them from the set of *either* 2D or 3D attributes.

To reduce the number of possible combinations, for **Double** both attributes were either from the set of 2D attributes, or from the set of 3D attributes. For **Triple** the two best attributes from **Double** remained fixed and the third was chosen from the same attribute set.

Experiment 3: Investigate Learning Dynamics

CLARET-2 was used to simulate the learning dynamics for the three experimental groups. For each group, CLARET-2 was used to predict each of the three classification matrices, observed at the begin, middle and end of the learning process. This resulted in a total of nine simulation runs. For each simulation the same conditions and parameter ranges applied as in experiment 1.

5.2.4 Results

Experiment 1

The results of simulating the learning data are listed in Table 5.4. The differences between controls and the other two groups were the most marked, where both the maximum resolution and the maximum memory capacity increased. Also to fit the control group data, 2D Attributes were sufficient, whereas the other two groups require 3D Attributes. The difference between the two priming groups visual and haptic was smaller, only the memory was increased from 20 to 30 partitions.

The analysis of the behavioural data showed a ranking of the experimental groups by their learning performance (Control < Visual < Haptic). The same ranking is found, ordering the groups by the amount of resources needed to achieve the best fit between predicted and observed classification matrices.

The results also showed that neither the probability nor the chain length parameter had a significant influence on the simulation results. Therefore these two parameters were set to the values giving the best fit, which was independent of the behavioural group, and are not mentioned anymore in

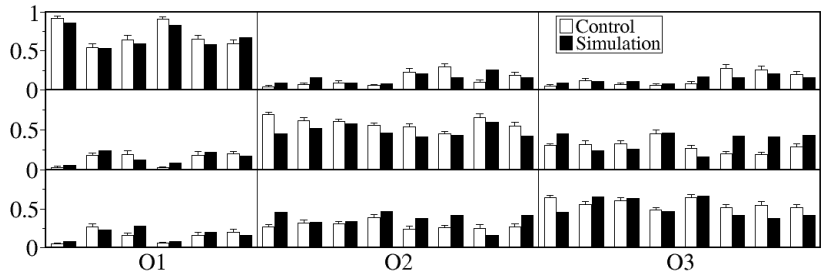
	Resolution	Memory	Attribute	Graph	Dist.
Control	5	10	2D	Adj.	0.036
Visual	9	20	3D	Adj.	0.037
Haptic	9	30	3D	Adj.	0.042

Table 5.4: Simulation of learning data. For each group the parameters, type of attributes and graph representation are given, which resulted in the best fit. The distance is the root mean square difference between the observed and the predicted classification matrices. For all groups setting the parameter chain length to 1 and probability to 0.02 yielded the best fit.

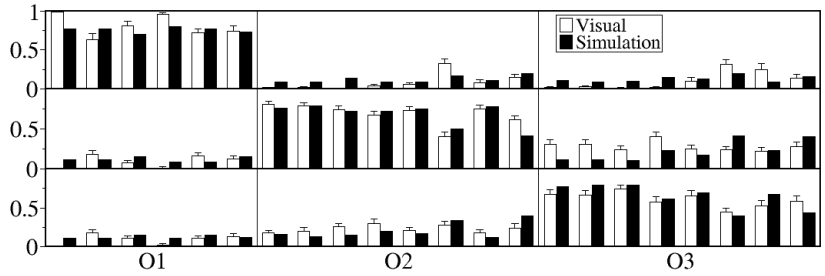
the further results. Note that a chain length of 1 was sufficient in every case to fit the behavioural data, i.e. only binary relations were required.

For all three groups the best fit was found with adjacent graphs. This held, regardless of the chosen attribute set. None of the fits, however, reached a statistically significant level.

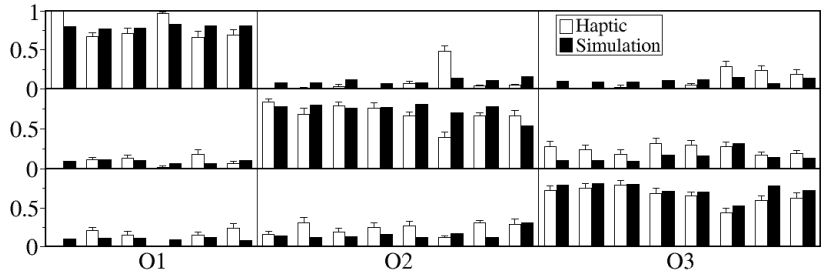
Figures 5.12 and 5.13 show plots of the data, three relative answering frequencies for each view, both as a diagram and as a scatter-plot. Inspection of the figures confirms the numerical analysis.



(a) Control group

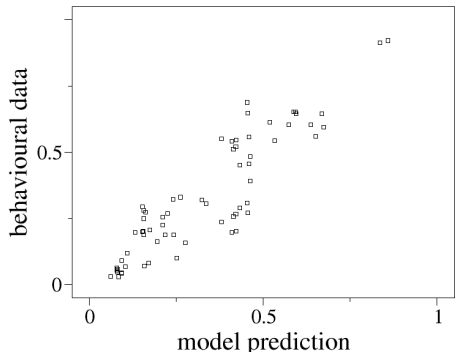


(b) Visual priming group

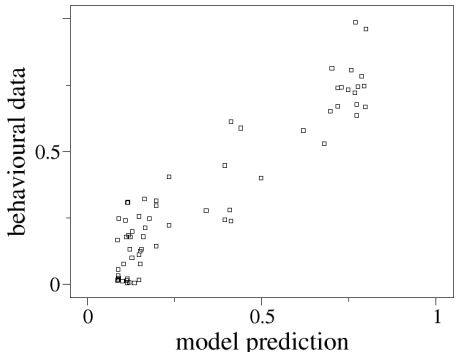


(c) Haptic priming group

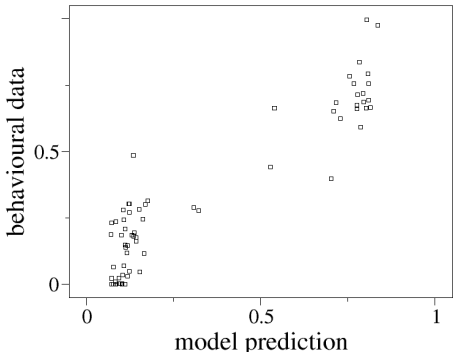
Figure 5.12: For the control and priming groups the three figures compare the observed classification matrix with the classification probabilities computed by the simulation. For each graph the ordinate lists the individual 22 learning views, grouped by object number, the abscissa measures the relative answering frequencies (behaviour) or the predicted classification probabilities (simulation). The top third of each graph depicts relative frequencies of answer 'object 1'. The middle third of each graph shows the relative frequencies of answer 'object 2' and the bottom third of 'object 3'. Priming groups show relatively the largest discrepancies between predicted and observed answering frequencies for view 1 of object 1 and view 5 of object 2.



(a) Control group



(b) Visual priming group



(c) Haptic priming group

Figure 5.13: The three figures visualise the predictive power of the simulation for the control and the priming groups. For each of the 3 possible answers for the 22 learning views the classification probability predicted by the simulation (ordinate) is charted against the observed relative answering frequencies (abscissa). For a perfect correlation all points would lie on the diagonal.

Experiment 2

The results for varying the number of available attributes are summarised in Tabs. 5.5, 5.6, and 5.7. Using two attributes minimised the error, compared to the other cases with either 1, 3 or 7 attributes.

Single Using only a single attribute already resulted in fits, which had approximately the same quality as using the full attribute sets. For control subjects the fit was even better. Independent of the subject group providing adjacent graphs with the single 3D attribute *difference in mean depth between parts* yielded the best results. Again, going from controls over visual to haptic priming group, more resources were needed.

	Resolution	Memory	Attribute	Graph	Dist.
Control	6	10	depth mean	3D-Adj.	0.031
Visual	7	20	depth mean	3D-Adj.	0.037
Haptic	8	30	depth mean	3D-Adj.	0.040

Table 5.5: Simulation of learning data using the single best attribute. For all three groups the attribute 'mean depth' gives the best results.

Double Selecting 2 attributes, where both are from the same set of either 2D or 3D attributes resulted in the optimal fit. Every subject group was now fitted with a smaller error than previously. There was no differentiation between the priming groups haptic and visual, both used the same parameters, the same graph structure and the same attributes. Fitting the control group we found to produce a stronger difference in respect to the other two groups. Here using the 2D attributes of *difference in mean intensity and maximum span between two parts* gave the best fit.

	Resolution	Memory	Attribute	Graph	Dist.
Control	10	20	intensity & span	2D-Adj.	0.030
Visual	8	15	depth mean & var	3D-Adj.	0.031
Haptic	8	15	depth mean & var	3D-Adj.	0.035

Table 5.6: Simulation of learning date using the two best attributes, with the additional constraint that both be either 2D or 3D attributes. The simulations of the observed data for the two priming groups gives the same results, 3D attributes produce the best fit. In contrast the control group is best fitted using 2D attributes.

Triple Now a third attribute was selected, leaving fixed the two best attributes found in the previous simulation. This slightly increased the errors for all three groups. All the groups are now differentiated again, in terms of selected attributes. For haptic subjects this was the additional attribute of *length of common shared border between parts* measured in 3D, for visual priming the *difference in 3D maximum span of parts*, and for controls the *differences in 2D areas between parts.*

	Resolution	Memory	Attribute	Graph	Dist.
Control	9	20	+ area	2D-Adj.	0.032
Visual	9	20	+ span	3D-Adj.	0.034
Haptic	9	15	+ border	3D-Adj.	0.038

Table 5.7: Simulation of learning data using an additional third attribute to the two found in the **double** attribute experiment. The third attribute is also the same type as the first two, 2D or 3D attribute. All three groups are best fitted using different triples of attributes.

Experiment 3

The results of applying CLARET-2 to the dynamic learning data, using the same parameter ranges as in the static case, is summarised in Tab. 5.8. For all groups both the resolution and the memory limits increased with time. The number of levels stayed constant, though. Both priming groups, haptic and visual, started off with 2D Attributes for the observed begin-matrix. They changed to 3D Attributes, as time passed. The simulation for controls started with 2D Attributes and an adjacent graph, but for the observed end-matrix the best fit was found with a fully connected graph, using 3D Attributes. Simulation errors reached their minimum at the end of the learning phase.

5.3 Discussion

The most remarkable result of this experiment is the strong effect of priming in general and specifically of haptic priming. Only five minutes of blindfolded haptic exploration of the object models improved the learning time by about 50% as measured in learning units, or by about 1 hour. Visual priming also had a strong effect compared to the controls, but was not quite as efficient as haptic priming. Especially the instant transfer from haptic priming to the visual learning task was significantly better than the transfer visual — visual. This confirms the importance of inference from prior knowledge for

	Time	Resolution	Memory	Graph	Dist.
Control	mean	5	10	2D-Adj.	0.036
	begin	3	10	2D-Adj.	0.033
	middle	4	15	2D-Adj.	0.036
	end	9	30	3D-Full	0.028
Visual	mean	9	20	3D-Adj.	0.037
	begin	5	10	2D-Adj.	0.039
	middle	9	20	3D-Adj.	0.039
	end	9	40	3D-Adj.	0.030
Haptic	mean	9	30	3D-Adj.	0.042
	begin	5	10	2D-Adj.	0.049
	middle	9	20	3D-Adj.	0.044
	end	10	30	3D-Adj.	0.035

Table 5.8: Simulation of learning dynamics. For each group the parameters are listed giving the best fit at the begin, the middle, and the end of learning. For comparison the results from Experiment 1, fitting the mean classification matrix ('mean') are repeated. The results of all three groups show a change in parameters consistent with the observed progress in learning.

visual recognition. Although the haptic subjects had never seen the objects, they were instantly better at recognising the objects and remained to be so throughout the experiment. There are several possible reasons for the advantage of the haptic modality over the visual modality as found here.

A representation formed by haptic exploration is expected to be of a spatial 3D nature[7]. This exploration experience biased most of the subjects to employ a matching strategy based on the 3D structure of the objects, i.e. mental rotation. Given that two of the objects had a mirror symmetric spatial structure this proved to be the most efficient strategy to solve the learning task. The interviews showed that 12 of the 14 subjects experiencing haptic priming chose this strategy. This uniformness was confirmed by the unimodal estimated probability density function (see Fig. 5.8). From the subjects undergoing visual priming the fraction using a mental rotation strategy early in the experiment was smaller. Again this is confirmed by the estimated probability density function, which shows secondary peaks at longer learning times.

Summarising Eleanor Rosch's work on categorisation, Lakoff (1987) ar-

[7]This doesn't necessarily imply viewpoint-independence, which would also require the integration of the sensed information over space and time. Even if the internal representation were completely viewpoint-independent the matching to a particular view may depend on the chosen view (see Sec. 3.4)

gues that motor activity plays an important role in basic-level categorisation. Basic-level categorisation is basic in several respects, among them, for instance, fast identification and shortest most commonly used names. One of the factors is the categorisation according to the type of motor interaction. Further, he cites Brent Berlin, who hypothesises that non universality of basic-level categories can develop, if experts have special training in a certain field. They might treat a more specific level as basic. Now surely, the faster and better our subjects learn, the more likely they are to treat this special categorisation task as basic, and from the above it is clear, that motor activity would play an important role in this type of training. More so, since other factors, which make a certain categorisation level basic are missing here, since we are using novel, unfamiliar objects, where naming or "world knowledge" would play little role. Looking from a neurophysiological angle Fuster (2001) argues the importance of what he calls the 'action-perception-cycle', the iterative feedback between perception and exploratory motor activity for the representation of objects. In our case this is facilitated by the *unrestricted* exploration of the objects, as opposed to experiments undertaken by (Newell et al., 2001, but see Sec. 6.2 for more details). Nevertheless, during visual priming subjects also actively explored the objects, but with a few important differences. A movement of the mouse had different effects on the objects, depending on the location of the cursor when pressing the mouse button, leading to rotations within the viewing plane or in depth. The movement itself is abstracted from the actual rotation and very much reduced in its possibilities (press and release mouse button, move left-right, forward-backward). Only a few haptic submodalities were activated involving possible movements, as well as possible sensory feedback. The mouse itself as well as its pose within the hand remained nearly constant during priming. Compare this to an active haptic exploration, using several fingers, the palm, hand and arm, both left and right, experiencing weight, inertia, surface properties, etc. These differences tie in well with the concept of an action-perception-cycle (Fuster, 2001).

The question arises, whether the haptic modality alone is already sufficient to result in optimal performance? Seven additional subjects undertook a control experiment, which was an exact replica of the haptic priming condition, with the single exception that priming was now visuohaptic, i.e. the subjects were not blindfolded. The results showed a further decrease in learning time, with visuohaptic subjects requiring about 6-7 learning units, thus learning about 50% faster than haptic subjects. Visuohaptic was therefore the optimal priming method for this learning task, as one would have ex-

pected[8].

The simulation results show a remarkably good agreement with the behavioural data (see Fig. 5.12), additionally they can be interpreted in a plausible way, regarding use of resources and preferred attributes.

In all simulation experiments the haptic priming group requires at least as many resources as the visual priming group which in turn needs at least as many resources as the control group. The relative performance of the three groups in the behavioural experiments shows exactly the same behaviour. Further, already the simulation results from the first experiment indicate that the internal object representations subjects build during learning show a qualitative difference. Whereas the control group has a 2D based representation, the visual and haptic groups build a 3D internal representation. This is confirmed by the results of the interviews.

Experiment 2 allows more precise statements. The main attributes the visual and haptic groups use, are the differences in the mean depth values and in the variance of depth values between two parts. The variance in depth can be interpreted as measuring the occlusion of a part (see Sec. B.1.2 for more details), but the important attribute is the difference in depth, since it is also used to model the representation using only a single attribute. It simply tells which part is further away from the observer's eye. By using an adjacent graph, where only neighbouring parts share relations, already the 2D — planar — structure of the object is available to the observer. Together with the depth attribute, and the variance in depth, measuring occlusion, the full 3D structure is available to model each object view.

These results allow a prediction for spatial generalisation. Both haptic and visual groups should be able to generalise to new viewpoints, whereas the controls should only do so to a lesser extent. These predictions are indeed confirmed by the results of a spatial generalisation experiment (see Sec. 6.2). The plausibility of this can already be checked using the results of learning alone. In looking at the simulation results for the haptic and the visual groups in Fig. 5.12(c) and Fig. 5.12(b), one finds one remarkable view, view 5 of object 2 (see Fig. 5.14). The simulation error is highest here, nearly twice as high as the error of the second worst view. Given that subjects from these two groups use structural 3D representations of the objects, as was suggested above, it can be expected that this view can only be discriminated from a view of object 1, if the estimation of the depth of the topmost sphere is correct. This attribute makes it possible to decide, to which of the occluding spheres in the foreground the topmost sphere is

[8]This did not hold for the spatial generalisation experiment (see Sec. 6.2) where subjects with visuohaptic priming showed no advantage over subjects with haptic priming

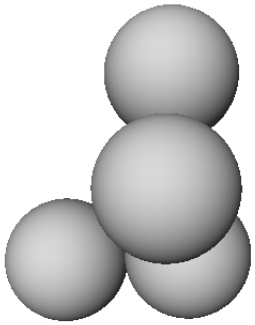


Figure 5.14: View 5 of object 2. Given the knowledge of structures of the 3 objects, this view can only be differentiated from object 1, if the depth of the top sphere is estimated correctly.

attached to, thus differentiating between a view of object 1 or object 2. While the simulation program can compute the depth precisely, and therefore shows a small probability of confusing this view with object 1, the subjects make estimations which are biased towards higher depth values. Indeed, this is one of the views, subjects tended to learn by heart. Since control subjects start without any internal model of the object structure, they don't tend to get confused with this view and simply learn it as a view of object 2. Further, even in the course of learning, the results above for controls suggested the use of 2D attributes, where depth values play no direct role.

These findings could lead to the assumption that subjects should benefit from binocular presentation, adding disparity as a further depth cue. An additional experiment with 6 subjects testing this assumption showed hardly any improvement in learning time[9]. This is in agreement with the results of Humphreys and Kahn (1992) investigating the viewpoint-dependency of response times depending on rotation in depth. The result is also not so surprising when keeping in mind the postulated importance of prior knowledge and inference for visual recognition. Since a priori the observers have no internal model of the objects they cannot benefit from the additional depth information given.

Another prediction for generalisation results from the different use of 2D and 3D attributes to reconstruct the internal representations of control subjects on the one hand and visual and haptic subjects on the other hand. If this is a valid result, the subjects should show a different ability to generalise

[9]The subjects were not tested for stereo vision, this might account for the single outlier requiring 60 learning units.

to differences in illumination and reflection properties of the objects. Since the 2D and 3D attributes are correlated, the predictions based on the types of attributes should be taken as tendencies rather than absolute demands.

It is interesting to note, that in all simulation experiments the best fit was achieved with a relational level limited to $l = 1$. This means, that binary relations, i.e. looking at pairs of spheres only, are sufficient to describe the views of the three objects. So obviously, although two objects are mirror symmetric, the complexity of the learning task is not so high to require higher level relations, where chains of three or four spheres are considered[10]. In this mode, CLARET-2 seems to have some similarity to normal decision tree learning[11], but there are some important distinctions. First, we have a "decision tree" for every view, that is being recognised, since CLARET-2 always partitions dependent on the unknown view. There is no precompiled tree structure. Second, CLARET-2 checks the compatibility between parts, while building a representation, something which is not considered in decision trees. Finally, CLARET-2 computes the probabilities not using likelihood, but by using an exponential distribution on the mismatch between the parts within every partition.

There are several possible reasons why the agreement with the observed behavioural data is not statistically significant. One reason might be that there are two distinguished views of object 1, which are recognised with hardly any errors by all subjects. Subsequently any difference between simulation and behavioural data receives a very high weighting. These views (view 0 and view 3) show the object centred on its symmetry axis. For an human observer the threefold symmetry is instantly recognisable. In simulation there is no attribute describing this symmetry explicitly, for CLARET-2 these views are as hard to classify as most other views are. A second possible reason is the fact that in the simulation the depth values are computed precisely, whereas the error in the estimation of human observers seems to depend on the specific view, as can be seen especially for view 5 of object 2. It is clear that the qualitative predictions and the quantitative behaviour of the simulation algorithm need to be tested in a separate generalisation experiment, as it is described in the following chapter.

[10]Using adjacent graphs, connecting nearest neighbours, conveys structural information to a certain degree.

[11]In decision tree learning the attribute space for the training examples is recursively split along a single attribute dimension, thus creating a tree, where every node splits in two branches. The recursion is completed, when no more splitting is possible or when in every end node (leaf) all elements belong to the same class. The class of an unknown example is determined by combining the leafs it activates in attribute space (See e.g. Breiman et al., 1984; Quinlan, 1990a).

Chapter 6

Generalisation of Learned Classification Performance

6.1 Introduction

After the supervised learning procedure all subjects from the different priming groups and the control condition have arrived at a defined learning status, where they are able to identify at least 20 of the 22 learning views correctly. Hypotheses derived from analysing the learning experiment and its simula-

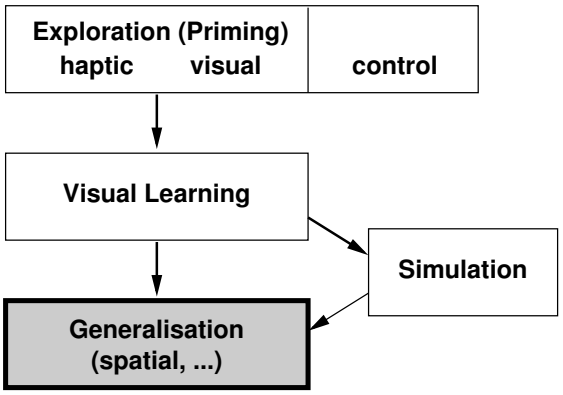


Figure 6.1: A generalisation test allows the validation of hypotheses derived from the results of supervised learning and computer simulation.

tion by means of an object recognition algorithm can now be tested by a subsequent generalisation experiment to validate any statements made. There are many possible variations to test the ability of the subjects to generalise, for instance changes in lighting properties, changes in surface properties such

as texture and reflectance, the replacement of the constituent parts by different ones, and changes of the object structure itself. The most obvious variation though is the presentation of the unaltered objects from novel directions, the test of the ability for spatial generalisation. This is the test which will be used here.

From the learning experiment one would expect unprimed control subjects to show the least ability to generalise to novel views, followed by visually primed subjects. Subjects from the haptic priming group should show the highest ability to generalise. These differences should be strongest for the chiral objects 2 and 3, where identification is facilitated most strongest by using a 3D structural representation.

6.2 Spatial Generalisation

6.2.1 Methods

Subjects

Of the 51 subjects who had reached the criterion of 90% correct in category learning, 20 participated in the generalisation test. Whenever possible, the generalisation test was conducted immediately after the conclusion of the learning task, after taking a short break. Where this was not possible because of fatigue or lack of time, the generalisation test was conducted in a separate session. In these cases, the subjects were required to confirm their learning status by solving the learning task again. All those subjects reached 90% correct answers within a maximum of two learning units.

Stimuli

The stimulus set consisted of the same three objects used for the learning task and was constructed in a similar fashion (see Sec. 5.1.2). Instead of sampling the azimuth and elevation in 60° steps, a finer sampling of 30° steps was used. The arbitrary and for each view fixed rotation in plane could take the values 0°, 30°, 60°, 90°, 120°, 150°, 180°. Again redundant views due to object symmetry were eliminated. This resulted in a total of 83 views (21 views for object 1, 31 views for objects 2 and 3 each). Of these views 19 views were from the set of learning views and already known to the subjects (known views)[1]. The rest, a total of 64 views (16 views for object 1, 24 for each of the objects 2 and 3), were unknown to the subjects (novel views). The

[1]The 'back' view with an angle of 180° was omitted

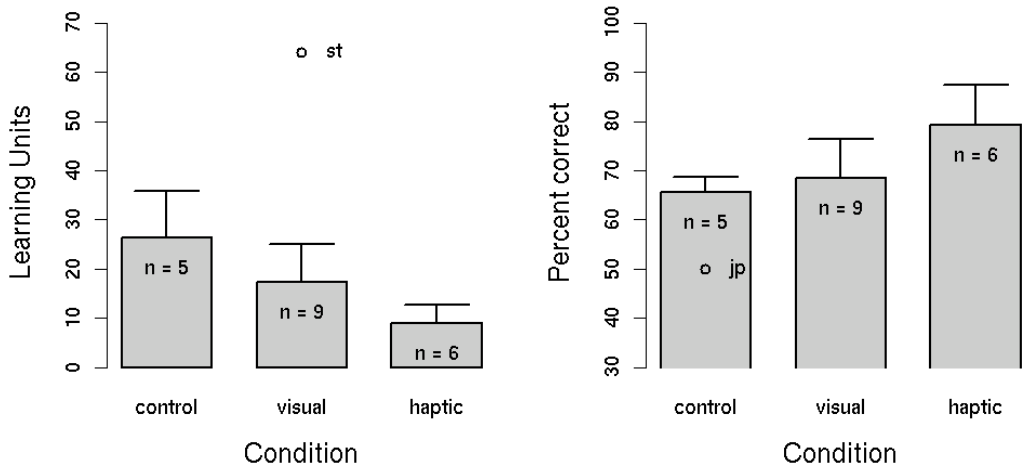
answers for known and novel views were analysed separately. The viewing distance (1m) and the presentation time (250ms) remained fixed. For the complete set of views see p. 111 of the Appendix.

The set of generalisation views, known and novel views randomly mixed, was presented once in random order. Each presentation was preceded by the display of a fixation cross and a beep signal. The observer was required to assign an object number to the test views, using a computer keyboard. After the 22nd and the 44th presentation, the subject had the opportunity for a break, if desired. No feedback was given.

6.2.2 Results

Novel views of the generalisation set

Figure 6.2, comparing the total percent correct answers with the learning times, shows differences in the ability for spatial generalisation between the three groups. The performance of all subjects was well above chance level.



(a) Learning Time

(b) Generalisation

Figure 6.2: Performance at the spatial generalisation task compared to learning times. The first subgraph shows mean number of learning units required to achieve 90% correct answers for the participating subjects. The second subgraph shows mean generalisation performance for all three objects. Both subgraphs show 20% trimmed means (see Sec. C.1).

Similar to the learning experiment, the control group showed the poorest

performance (about 66% correct), followed by the visual group (about 69% correct) and the haptic group (about 79% correct)[2]. The respective mean learning times for these subjects were 26 units for control, 17 units for visual and 9 units for haptic. The subject with the worst performance, subject *st* (about 47% correct answers) was already treated as an outlier when analysing the results of the learning section (see Sec. 5.1.3), requiring 64 units to achieve the learning criterion[3].

A qualitative difference between haptic subjects and others shows up in the comparison of classification of object 1 and the mirror-symmetric objects 2&3 respectively (Fig. 6.3). While haptic subjects showed the same

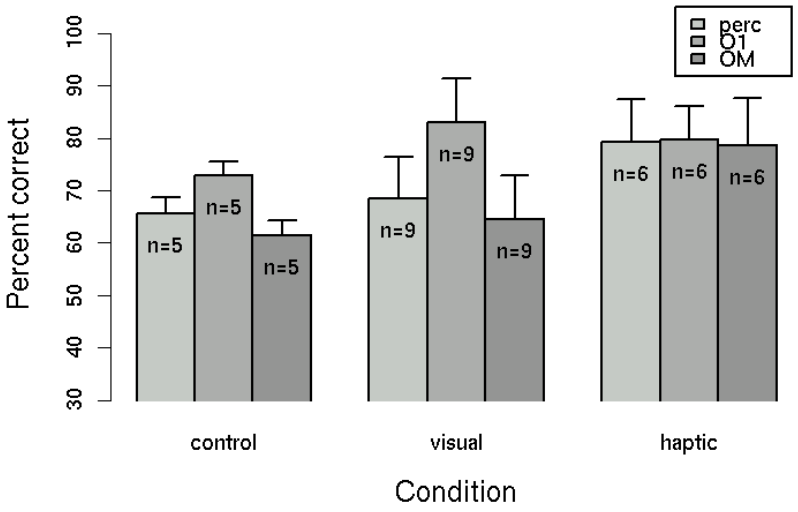


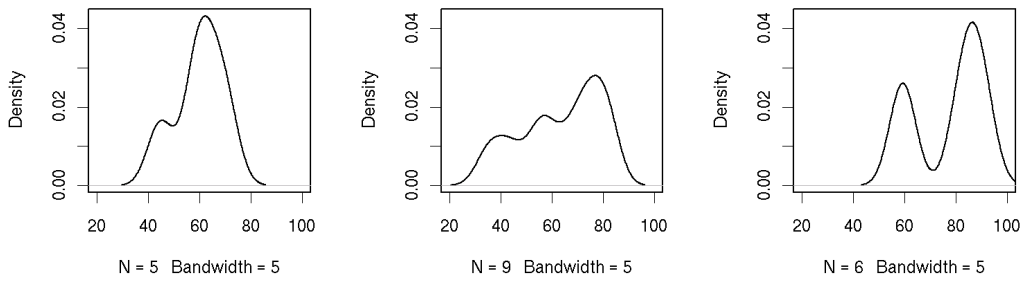
Figure 6.3: Comparison of generalisation performance for all objects (perc), for object 1 (O1), and mirror-symmetric objects 2&3 (OM). 20% trimmed means are shown (see Sec. C.1).

performance in classifying views of the three objects, control and visual subjects had more difficulties classifying the two mirror-symmetric objects.

The estimated density functions for the distribution of percent correct answers (at the abscissa) for the chiral objects 2 and 3 only are shown in Figure 6.4. Note the broad distribution for visual subjects compared with both control and haptic groups.

[2]None of the differences mentioned in the following text reach a significant level.

[3]The other two outliers mentioned there did not participate in this generalisation test.



(a) control (b) visual (c) haptic

Figure 6.4: Estimated densities of percent correct answers for objects 2 and 3 only. The three plots show the data for control, visual and haptic subjects. Outliers were removed.

Comparing Known and Unknown Views from the Generalisation Set

To assess the stability of the learning state the subjects had achieved, 19 views from the learning set (known views) were presented in addition to the 64 novel views during the generalisation tests (see Methods 6.2.1). The two types of views were presented randomly intermingled. Related classification performances are plotted in Figure 6.5. According to the percent correct responses shown there, the performance for the known 19 views hardly differs from the performance for novel views. The classification dropped to about 61% for control subjects, 72% for visual subjects, and 74% for haptic subjects. The performance thus lies clearly below the expected 90% correct answers, at which the learning procedure had been terminated.

In summary, it can be noted that the learning state is not stable to changes of task context, achieved here by adding novel views. The performance for known views drops approximately to the level of performance for novel views.

6.2.3 Correlations between Learning and Generalisation

In Figure 6.6 the correlations between generalisation performance and the results for the learning phase are compared for each individual subject, without regard for the priming group. The first graph compares the time required during the learning phase with the subjects performance during generalisa-

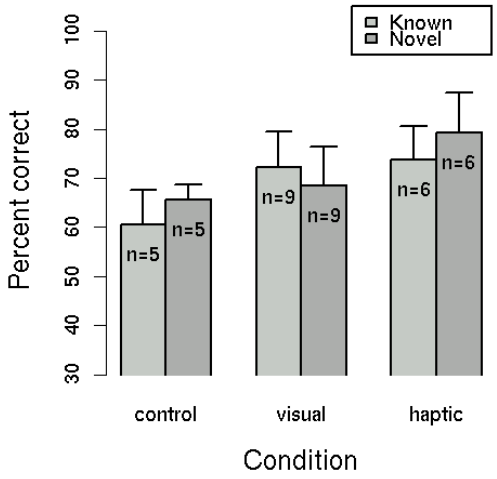


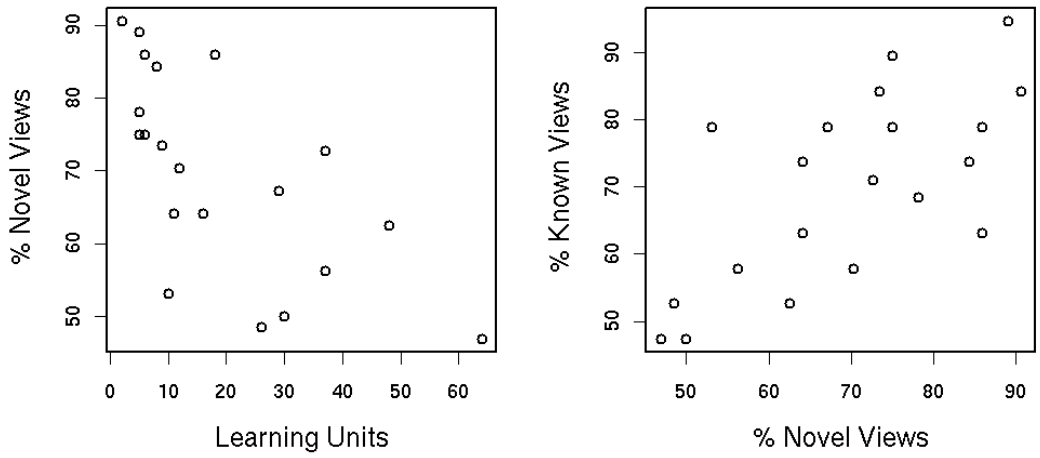
Figure 6.5: Performance during generalisation for known views is compared to the mean performance for novel views. 20% trimmed means are shown (see Sec. C.1).

tion for novel views (Figure 6.6a). The second graph relates the percent correct answers during generalisation for the 19 known views (abscissa) to the performance for the 64 novel views (ordinate, Figure 6.6b). Computing Pearson's correlation reveals a significant correlation between the duration of learning, measured by the required number of units, and the classification performance during generalisation ($R = -0.66, p < 0.005, df = 18$). Also, the performance during generalisation for known views is predicted quite well by the performance for the novel views ($R = 0.68, p < 0.001, df = 18$)[4].

Summary of Results

For testing spatial generalisation, the results of each subject correlated with the performance during the learning phase. Also, the three subject groups showed the same tendency of haptic subjects being better than subjects with visual priming and both being better than subjects without priming. Examining the generalisation performance for object 1 and the mirror-symmetric objects 2&3 showed a qualitative difference for subjects with haptic priming. These subjects performed equally well for both classes of views, whereas control subjects and subjects with visual priming show a weaker performance

[4]This mode of analysis may seem counterintuitive, but actually the deviation of the recognition performance for known views from the expected value of about 90% correct is predicted by the performance for novel views



(a) Learning time and novel views

(b) Novel and known views

Figure 6.6: The scatter plots compare the performance of each subject during generalisation test (a) for novel views with the number of units required during learning and (b) the changed mean performance for known views and novel views during generalisation. The three conditions were not separated.

for the mirror-symmetric objects 2&3. The estimated distributions of correct answers reveal a tendency towards a bimodal curve for primed subjects.

The training status did not remain stable under the altered context of spatial generalisation. For all groups the performance for the intermingled 19 known views was well below the expected minimum of 89%[5]. The performance for each group agreed with the respective performance for classifying novel views. The performance of each subject correlated with its performance for novel views and with its performance during learning.

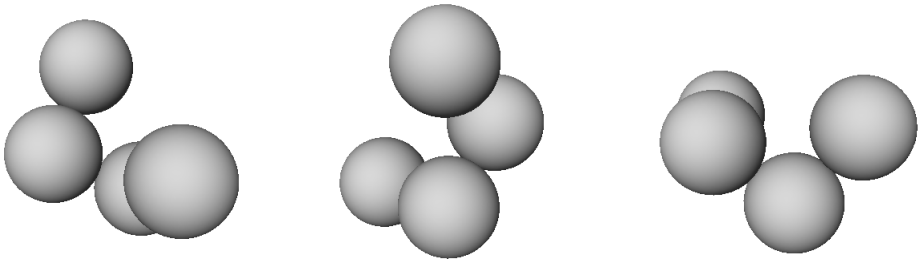
6.3 Discussion

The results from the generalisation experiment support the view that object recognition depends largely on inference. The formation of a structural representation of structure-only 3D objects depends not only on the opportunity,

[5]To reach 90% correct answers during training, the subjects had to classify 20 of the 22 views correctly. Assuming that the three views removed from the original learning set were learned correctly this leaves 17 of 19 views to be classified correctly, i.e. 89%

but also on the type of priming. It can be assumed that active exploration plays an important role there.

Structure-only objects. The correct classification of novel views of the objects depends on the representation of and matching to the 3D object structure. The objects are constructed from identical parts with no defined geometrical axes. The constituent parts of the objects, the individual spheres provide no cues to the macro-geometrical shapes of the object. In contrast, geons for instance can be described as generalised cones, i.e. a cross-section moved along an axis of elongation (Biederman, 1987). The orientation of the axes of elongation of object parts provides important cues to the object identity, which are missing here (Note that spheres are not part of the geon set proposed by Biedermann). The use of two chiral, mirror-symmetric objects among the set of three objects creates another strong bias to employ 3D structure for the classification task. To determine the handedness of objects 2 and 3 the complete 3D structure needs to be established. The match of novel views to learned views without using the internal structure depends on an external frame of reference, i.e. the positions of the spheres relative to the frame of the computer screen. The use of such an external reference frame is disrupted by adding a random rotation within the picture plane to the rotation in depth (see Figure 6.7).



(a) learned view (b) novel view (c) learned view

Figure 6.7: Three 'neighbouring' views of object 3. The rotation in the picture plane differs for all three. Even given only the two neighbouring learned views for the central novel view, with a difference of $\pm 30°$ in azimuth, demonstrates that an alignment is a highly non-trivial task, even without regarding the remaining 20 learning set views

There exist 'diagnostic' features to classify object 1 (a 'star' shape, with

the central sphere connected to three remaining spheres, vs. a 'snake' like elongation, with the four spheres lined up). These features can only be detected reliably in some of the views. No such feature exists to discriminate between the handed objects 2 and 3.

The property of the objects being "structure-only", can be further confirmed. When the views of the three objects are combined in a texture-like image, no pop out effect is visible; It is not possible to classify the views using cross-correlation; The views can't be recognised spontaneously by the subjects, even after priming. Subjects undergoing haptic priming initially perform slightly better, than subjects with visual priming.

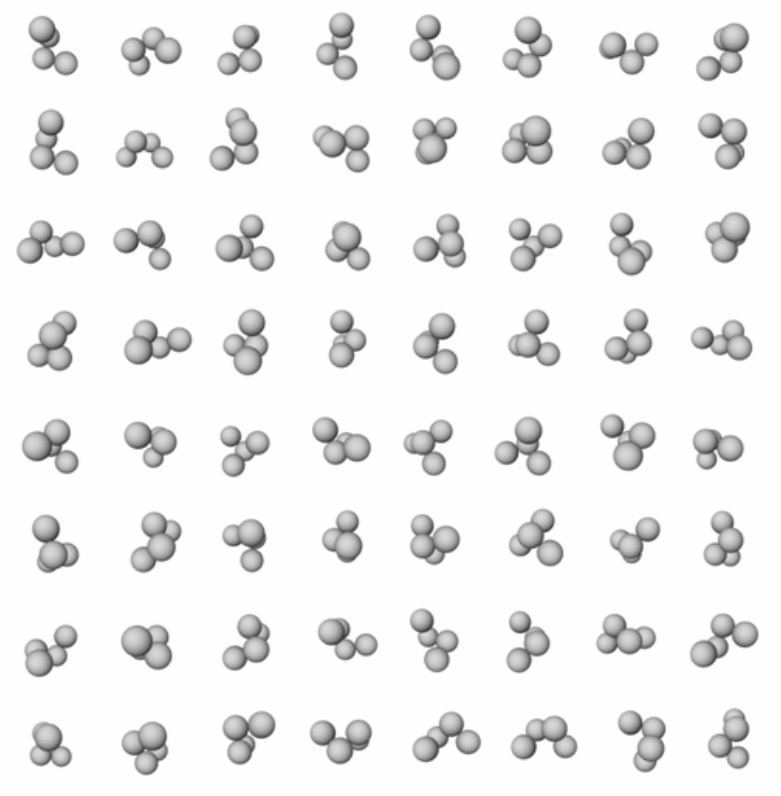


Figure 6.8: A texture like combination of views grouped by object number. The boundaries between the 'texture regions' of the three objects are not visible, i.e. there is no pop out effect.

Inference. As was discussed in Chapter 2 there exist two differing views of human vision. Vision can be seen as a purely bottom-up processing of

input stimuli based on hierarchical filtering, vision as analysis. The alternative view, vision as inference, assumes a control of visual processing in a top-down fashion, depending on prior knowledge about objects and the context they appear in. Based on such internal knowledge the probabilities for the presence of features of the outside world are computed. This is an interpretation of the neuroanatomical finding that feedforward connections, on which vision as analysis would be based, form only a small part of the overall connections in the visual cortex (e.g. Young, 2000).

The drop in performance for the 19 known views which were randomly intermingled with the novel views during the generalisation test, supports the importance of inference from prior knowledge for object recognition. For two reasons it is unlikely that this drop in performance is a memory issue: (1) some subjects weren't tested for spatial generalisation in the same session where they achieved the learning goal of 90% correct answers, but were tested on a different day. In those cases the learning procedure was repeated to assure a correct learning state. All subjects reached the set criterion within a maximum of 2 units. (2) A follow-up experiment was conducted with 7 subjects after an interval of about 1 year. First the subjects remaining ability to solve the generalisation task was tested. After that the remaining learning status was determined, using the 22 learning views. The ability of five subjects for generalisation had remained nearly unchanged after 1 year, although one of the subjects had consistently misnamed views of object 1 with 3, object 2 with 1 and object 3 with 2. Two subjects performed clearly worse than a year earlier, their performances dropped from 66% to 36% and from 78% to 45% correct answers, respectively. The learning status of the subjects was excellent, ranging from 64% to 100% correct answers for the learning test, with a 20% trimmed mean of 85%.

Therefore the drop in performance for known views can be explained in that way, that a changed context with novel views leads to a shift in prior probabilities which are used to infer an object identity from a given view. This change in prior probabilities probably happens with the first novel view which is presented. In a further experiment this hypothesis could be tested by appending another test showing only the 22 known views, with no feedback. It is to be expected, that subjects would again approach or reach 90% correct answers, as is already indicated by the follow up study.

Importance of active exploration to generate 3D-structural representation. The experiment of learning the objects already established that prior knowledge from priming increases recognition performance. The results from spatial generalisation do not only confirm this finding, but also

allow more specific statements about the influence of the type of priming. Subjects with haptic experience of the objects, handling them physically, although blindfolded, show a better ability to generalise. This increase in performance is mainly due to an increase in the ability to differentiate between the chiral, mirror-symmetric, objects 2 and 3. This finding in itself is another confirmation of the importance of inference for identifying the given objects. Although both priming groups had prior knowledge regarding the objects through manipulating them, and although both groups showed a similar learning performance, nevertheless the internal representations formed during learning appear to be of a different nature. The active haptic exploration, based on a feedback between perception and motor activity facilitates the formation of a 3D structural representation of the objects. Since the objects used here have the property of being "structure-only", as was argued above, subjects with haptic priming have an advantage in recognising novel object views, specifically views from the chiral objects 2 and 3. So not only the amount of prior information, but also its type plays an important role[6].

From the point of view of cognitive neuroscience, object representation is hierarchically organised, from the sensory to the semantic level, with reciprocal connections between sensory and motor areas providing the basis for the "perception-action-cycle" (Fuster, 2001; Fuster, 2003, chap. 4). In the dorsal as well as in the ventral stream of processing there exist areas, where visual and haptic modalities could be integrated. In the parietal cortex there seem to exist areas where information from both processing streams and from other modalities are transformed to abstract spatial representations (Landis, 2000; Milner and Goodale, 1995, sec. 4.5). In the lateral occipital complex activation by both visual and haptic object stimuli was observed, suggesting a "multimodal object-related network" (Amedi et al., 2001, p. 324). This suggests that active haptic exploration leads to coarse prior 3D models of the objects. Views of the objects are matched to prior models by spatial transformations, similar to mental rotation. During supervised learning this internal model and the extraction of information from the images are refined. This would explain that spontaneous identification at the beginning of the learning experiment as well as the generalisation to novel views for the chiral objects 2 and 3 was best after haptic priming.

As was noted for the learning experiment, the haptic effect appears to be quite strong compared to other experiments. The important fact seems to be the choice of structure-only stimuli and the opportunity for free explo-

[6]Note that also control subjects had prior knowledge about the objects of a very different type. They were verbally informed, that the objects consisted of four identical spheres, connected with each other in three different configurations.

ration. Reports from experiments where the objects were presented fixed to a table in front of the sitting subjects show a strong viewpoint dependency in the following visual recognition task (Newell et al., 2001). The subjects mainly explored the rear of the objects, and were subsequently better able to recognise views showing the rear part of the objects.

The superiority of haptic priming for the correct identification of the chiral objects 2 and 3 was already predicted by the simulation results of the learning experiment. The simulation results would have predicted a similar behaviour for subjects undergoing visual priming, which is not confirmed by the generalisation results. The motor activity during visual exploration via computer mouse does not lead to such a rich structural representation as is needed for this task. One reason for this may be that the rotation of the objects via mouse it too far abstracted from real exploratory movements. Motor activity and haptic feedback do not depend on the actual structural properties of the object. Another reason may be that the visual exploration process was not controlled. It is therefore not known to which extent the subjects explored the objects rotating them 'in-plane' or 'in-depth'. Subsequently, visually primed subjects show more variation in the recognition of the chiral objects than haptic subjects.

The importance of motor activity is illustrated by a further experiment with the same objects. There 7 subjects were primed visuohaptically, they explored the objects with open eyes, not blindfolded. Although they learned faster than haptic subjects, they showed no further improvement in their ability for spatial generalisation (about 77% correct answers).

Computer simulation The computer simulation algorithm was applied to the behavioural results using the optimal parameters found for each group modelling the learning experiment. This procedure did not lead to any conclusive results here. In view of the fact, that the performance of the subjects for the set of learned views was unstable during generalisation, this is not surprising. The simulation depends completely on the data for the learning views. If the behaviour of the subjects changes for those views, the simulation looses its basis.

A possible solution for this problem could be the extension of the CLARET-2 algorithm to incorporate continuous learning. Since the operation of CLARET-2 depends on the set of learned views, this set could be administrated by an external framework. Certain views could be selected as prototypes, depending on the history of presented views. As novel views are shown, the set of prototype views would be continually updated, depending on the feedback given, or, if this is missing, on the hypothesised object

identity. A similar framework was actually used by Pearce and Caelli (1999) to learn the recognition of handwritten characters through the predecessor algorithm CLARET.

Chapter 7

Summary

The physiological study of vision has been confined for a long time to the use of simple stimuli such as flashing lights, bars, and gratings. Only recently, there arose the question of how complex stimuli are represented in the brain, and how such representations are accessed to enable object recognition. This is not surprising given the fact that the recognition of objects from their 2D projections on the retina is, mathematically speaking, an ill-posed problem and thus not readily open to investigation. Ill-posed meaning here, there are an infinite number of objects resulting in the same 2D projection or image. This problem can only be solved by using prior knowledge of input stimuli, i.e., the external world. Another limitation of the traditional approach to visual object recognition is the use of stimulus discrimination, typically employed within the experimental paradigm of "delayed-matching-to-sample". Yet there is reason to believe that categorisation is the main function of the brain as the ability to assign manifolds of occurrences to single concepts is a conditio sine qua non for understanding. With categorisation comes the question of how to quantify similarity between stimuli, which is of fundamental importance.

It is of interest, therefore, to note that since more than a decade there are new developments in the domain of machine vision. They are motivated by the need to endow real-time computer systems with the ability to recognise complex objects in realistic, i.e., cluttered environments. The success of such attempts depends on two factors. First, algorithms for object recognition by machine employ strategies of "recognition-by-parts". That is, objects and scenes are segmented into parts, and the structure of input images is analysed by computing attributes of parts and attributes of relations between parts. The resulting object descriptions have the mathematical form of graphs. Second, recognition strategies have to rely on the comparison of input signals and class concepts, thus facing the problem of

"graph-matching". This formidable problem, mathematically speaking determining a subgraph isomorphism, is generally only tractable if heuristic shortcuts are found. Combined with probabilistic reasoning such heuristics can provide measures of similarity between stimuli. For these reasons, object recognition is heavily relying on strategies of machine learning.

The present study makes use of the principles of object recognition by machines to better understand visual object recognition by humans. To achieve this goal, the following conditions are met: Unfamiliar "structure-only" 3D objects are used that are composed of the same number of identical sphere parts. Thus, recognition using structural descriptions based on relations between parts is favoured. Recognition performance of human participants is studied within a paradigm of priming, supervised category learning, and generalisation, thus allowing inferences about the nature of internalised object representations. Computer simulations of human performance in object categorisation are enacted using a recognition-by-parts scheme, graph matching, and inductive logic programming. Thus, classification matrices are predicted that can be compared directly to the psychophysical categorisation data. Least-squares minimisation is then used to select models of internalised object representations.

The following results have been obtained: (1) Contrary to what has been claimed before, human observers are able to learn structured object representations for 3D objects that lack so-called 'geons', i.e., generalised cone components. (2) Under such conditions, recognition depends on using metric relational attributes for object description or representation. (3) Prior knowledge of 3D shape provided by active haptic exploration of physical models of test objects is needed to learn representations that preserve at least important aspects of 3D structure. (4) In the absence of prior knowledge, structural representations showing less viewpoint invariance are built. Consistent with recent reports from brain imaging and neurophysiology, these findings suggest that both structural and more view-dependent representations for object recognition exist in the human brain.

Kapitel 8

Zusammenfassung

Die physiologische Erforschung des Sehens beschränkte sich lange Zeit auf die Verwendung einfacher Reize wie z.B. blinkender Lichter, Balken, und Gittermuster. Erst in jüngerer Zeit wurde die Frage nach der Repräsentation komplexer Reize im Gehirn gestellt, und wie eine solche Repräsentation Objekterkennung ermöglicht. Dies ist nicht sehr überraschend, da die Erkennung von Objekten anhand ihrer 2D Projektion auf der Retina ein mathematisch unterbestimmtes Problem darstellt, und daher schwer zugänglich ist. Das Problem ist inkorrekt gestellt, da unendlich viele verschiedene Objekte existieren, die alle die gleiche Projektion erzeugen. Nur durch die Verwendung von Vorwissen über mögliche Eingangsreize, d.h. über die umgebende Welt, kann dieses Problem gelöst werden. Eine weitere Einschränkung, die sich im üblichen Ansatz zur Erforschung der visuellen Objekterkennung findet, ist die Verwendung von Reizdiskriminationsaufgaben, wie sie z.B. im „delayed-matching-to-sample“ zur Anwendung kommen. Es gibt jedoch Grund zu der Annahme, daß Kategorisierung eine Hauptfunktion des Gehirns ist, da die Fähigkeit der Zuordnung einer Viefalt von Erscheinungen zu einem einzigen Konzept die conditio sine qua non für Verstehen ist. Die Frage der Quantifizierung von Ähnlichkeit ist dabei von fundamentaler Bedeutung.

Daher ist es von Interesse, daß es im Bereich des rechnergestützen Bildverstehens seit über einem Jahrzent neue Entwicklungen gibt, motiviert durch die Notwendigkeit Computersysteme mit der Fähigkeit auszustatten, komplexe Objekte vor realistischem Hintergrund in Echtzeit zu erkennen. Der Erfolg dieser Bemühungen hängt von zwei Faktoren ab. Zum einen von Algorithmen zur Objekterkennung die sich auf die Zerlegung der Objekte in ihre Komponenten stützt, zum anderen vom Vergleich von Eingangssignalen mit gelernten Klassenkonzepten. Bei der Zerlegung der Objekte kann durch die Berechnung der Attribute der Teile und der Relationen zwischen den Teilen die Struktur eines Objektes beschrieben werden. Mathematisch

können solche Beschreibungen in einer Graphstruktur abgebildet werden. Der Vergleich von Eingangssignalen und Klassenkonzepten läßt sich dann als Paarung von Graphen darstellen. Eine solche Paarung, auch Teilgraph-Isomorphismus genannt, zu bestimmen ist im Allgemeinen so komplex, daß es nur mit heuristischen Näherungsalgorithmen berechenbar ist. Wenn solche Algorithmen mit wahrscheinlichkeitsbasiertem Schlußfolgern kombiniert werden, erhält man ein Maß für die Ähnlichkeit visueller Reize. Aus diesen Gründen hängt die Objekterkennung stark von den entwickelten Strategien des maschinellen Lernens ab.

Die vorliegende Arbeit benutzt bekannte Prinzipien aus dem Bereich der computergestützten Objekterkennung, um die visuelle Objekterkennung beim Menschen besser zu verstehen. Zu diesem Zweck werden folgende Voraussetzungen erfüllt: Unbekannte 3D Objekte finden Verwendung, die jeweils aus der gleichen Anzahl identischer Kugeln konstruiert werden, und sich nur durch ihre Struktur unterscheiden. Daher hängt die Erkennung dieser Objekte von der Beschreibung ihrer Struktur ab, basierend auf den Relationen zwischen den Komponenten. Die Erkennungsleistungen menschlicher Versuchpersonen werden innerhalb eines Paradigmas untersucht, das Priming, überwachtes Lernen, und Generalisierungstests vereint, und so Schlüsse über die Art der internen Repräsentation zuläßt. Computersimulationen der menschlichen Erkennungsleistungen werden durchgeführt, die die Methoden der Komponentenzerlegung, der Graphenpaarung, und der induktiven logischen Programmierung vereinten. Auf diese Art werden Klassifikationsmatrizen vorhergesagt, die direkt mit den psychophysischen Kategorisierungsleistungen verglichen werden können. Durch die Metode der kleinsten Quadrate werden die Modelle der internen Objektrepresäntation ausgewählt.

Folgende Ergebnisse wurden erzielt: (1) Im Gegensatz zu früheren Aussagen können menschliche Beobachter strukturelle Objektrepräsentationen auch für Objekte erlernen, die über keine sogenannten „geons“, d.h. verallgemeinerte Zylinder, als Komponenten verfügen. (2) Unter solchen Bedingungen hängt die Erkennung der Objekte von der Verwendung metrischer relationaler Attribute zur Beschreibung und Repräsentation der Objekte ab. (3) Durch haptische Erkundung erlangtes Vorwissen über die Form der Objekte ist nötig, um Repräsentationen zu lernen, die wenigstens die wichtigsten Aspekte der 3D Struktur enthalten. (4) Ohne Vorwissen werden strukturierte Beschreibungen erstellt, die eine größere Blickrichtungsabhängigkeit aufweisen. Übereinstimmend mit neueren Berichten, die sich auf neurophysiologischen und bildgebenden Verfahren stützen, legen die Ergebnisse nahe, daß im menschlichen Gerhirn sowohl strukturierte, als auch eher ansichtenspezifische Repräsentationen zur Objekterkennung existieren.

Appendix A

Object Views for Generalisation

To generate the set of views used during generalisation tests azimuth and elevation were sampled in 30° steps; the equatorial plane was horizontal and contained the symmetry axis of each object; the centre of the coordinate system was situated at the centre of gravity and thus on the symmetry axis of each object. The arbitrary and for each view fixed rotation in plane could take the values of 0°, 30°, 60°, 90°, 120°, 150°, and 180°. Redundant views due to object symmetry were eliminated, which resulted in a total of 83 views (21 views for object 1, 31 views for objects 2&3 each). Of these views 19 views were identical to the views from the learning set (see Fig. A.4 on the following pages). The rest, a total of 64 views were unknown to the subjects (see Fig. A.1, A.2, A.3 on the following pages).

The views were always presented in a different, randomly generated, order.

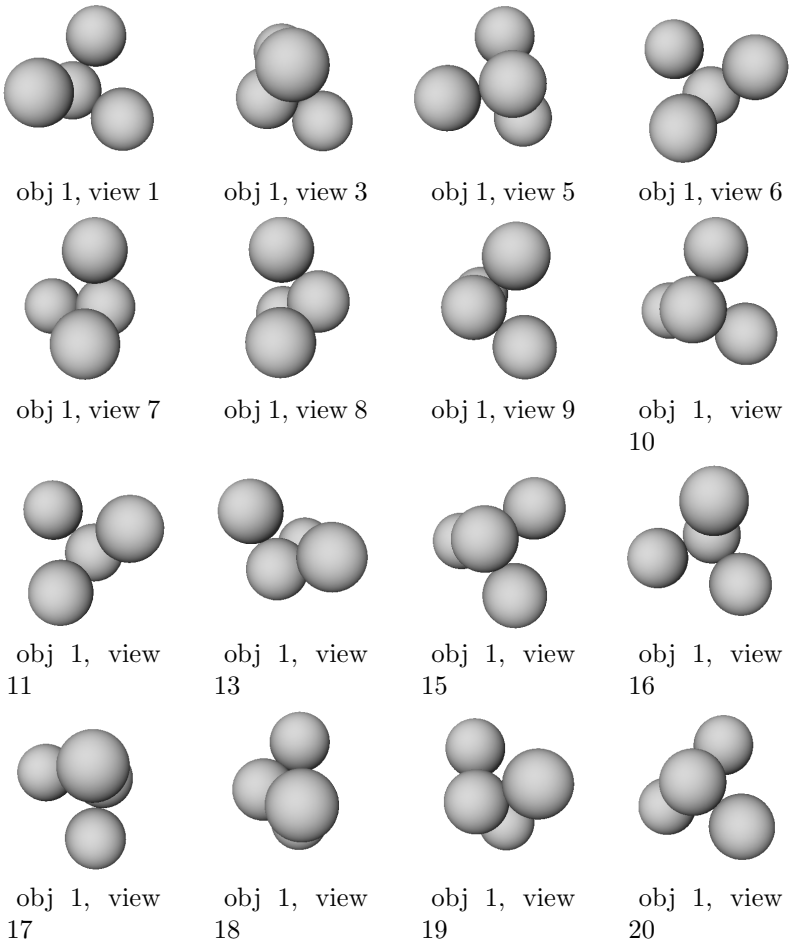


Figure A.1: Views of object 1 tested during spatial generalisation. Only novel views.

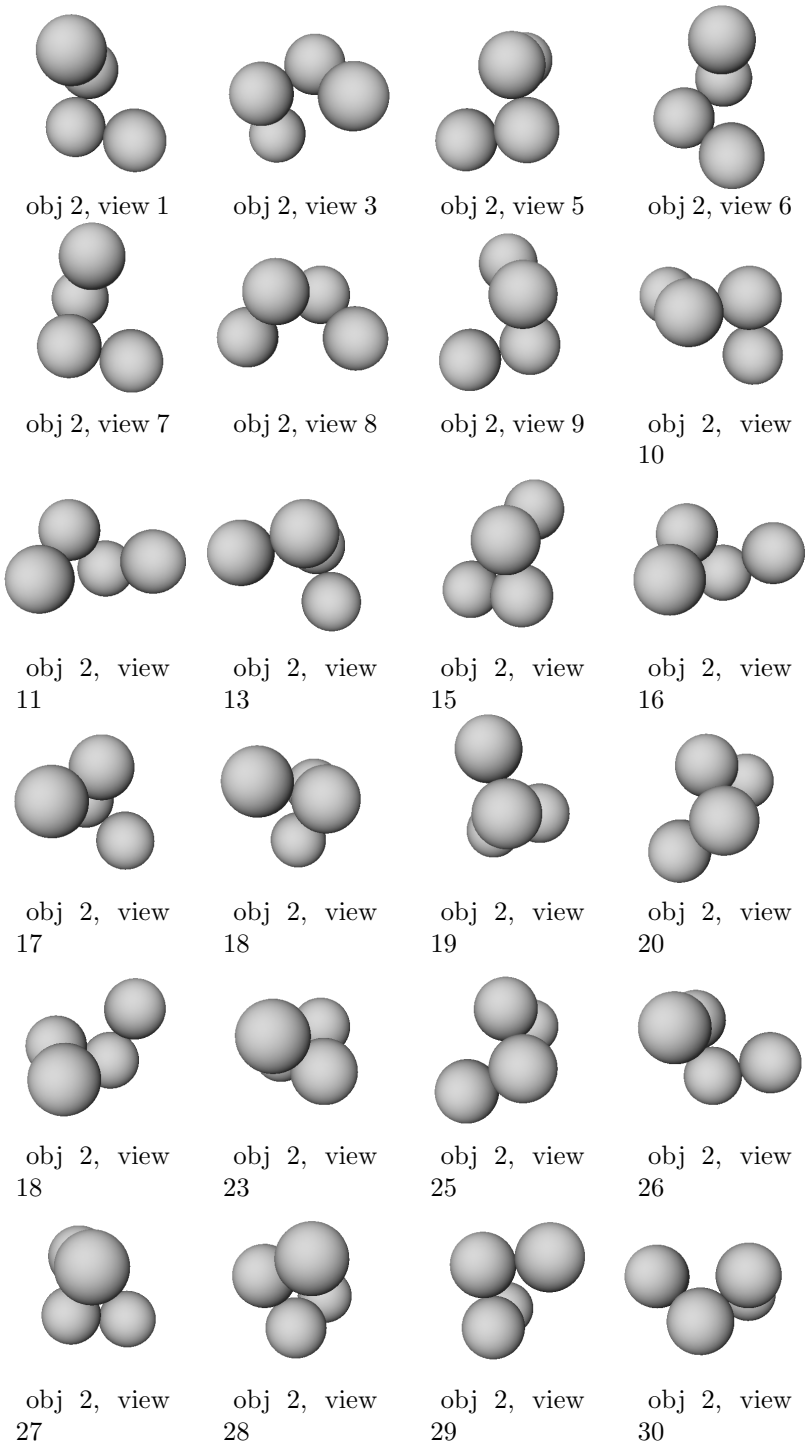


Figure A.2: Views of object 2 tested during spatial generalisation. Only novel views.

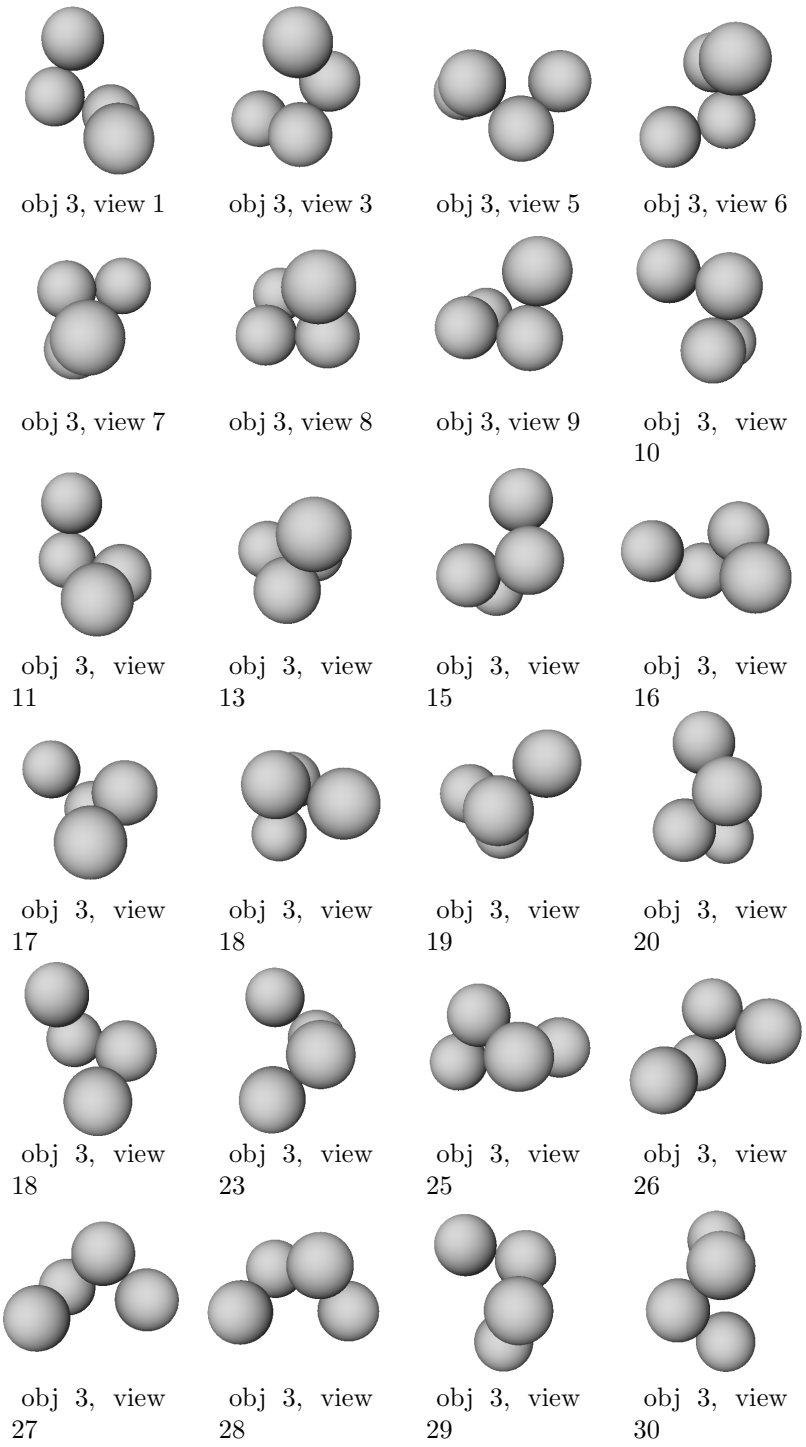


Figure A.3: Views of object 3 tested during spatial generalisation. Only novel views.

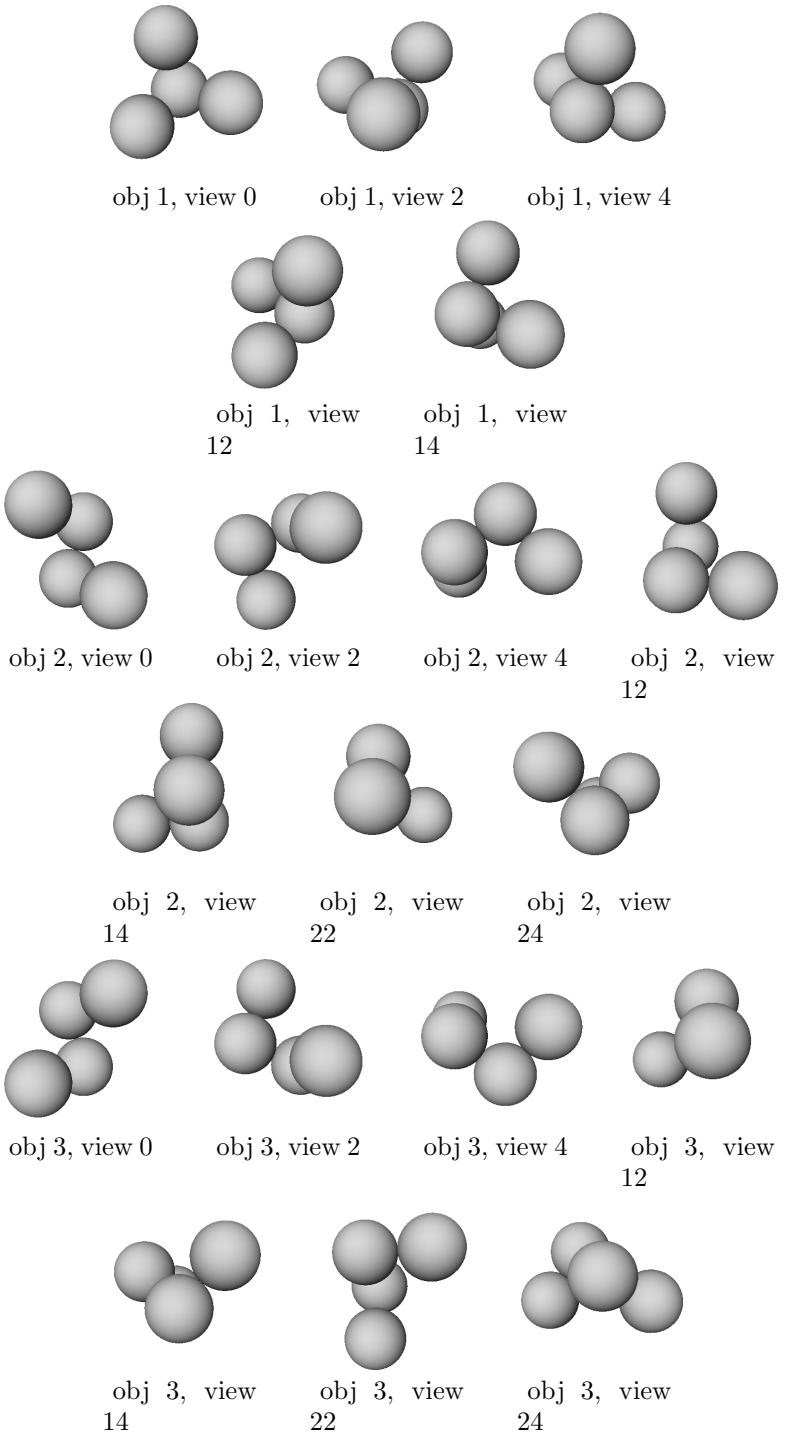


Figure A.4: Previously learned views of all three objects, which were presented randomly intermingled with novel views during the generalisation test. The numbering differs from the original set of learning views.

Appendix B

Description of the CLARET-2 algorithm

B.1 Input Representation

Learning algorithms based on recognition-by-parts have the disadvantage of relying heavily on the segmentation of their input patterns. Usually the performance depends on the type and the quality of the segmentation method used. Stability towards parameter changes and variations of the input patterns (e.g. noise, viewpoint variations) are an important issue.

In our application, we assume the segmentation of the objects already given by their structure. Since we are using compound objects, composed from several identical spheres, we identify the individual spheres with the input parts for the learning algorithm. Should the visible representation of a sphere consist of more than one contiguous region, due to occlusion, the pixels of those regions are nevertheless treated as a single set.

The images were mostly processed using self-written routines, although parts of IPRS (Image and Pattern Recognition System: Caelli et al., 1997) were used, mainly the data structures and I/O functions.

B.1.1 Intermediary Primary Attributes

In a first step the rendered views of the objects are used to compute a set of intermediary primary attributes. These attributes, which are stored in an IPRS_Featurespaceimage, are assumed to be sufficient to describe the sphere parts. This is not part of the actual CLARET-2 algorithm, but is done by two different programs, the first generating IPRS images from OpenInventor, the second using those images to extract the desired attributes.

The displayed pixels of each sphere form a region on the display, which is identical with a part. For each of these a summary description is computed called attributes of these parts. Overall 19 of these intermediary attributes are computed.

List of Intermediary Attributes

- i Label
  Every sphere of an object is uniquely identified by an integer. This information is of course not used directly in the matching algorithm, but only in the further construction of binary attributes.

- $\vec{p}(i)$ Coordinates
  The (x,y,z) coordinates of the centres of each region. In other words, the mean values of each coordinate are computed over all pixels belonging to the region.

- $v_z(i)$ Depth variance
  The variance of the z-coordinate.

- $f_3(i)$ Area in 3D
  The area of each region is computed by triangulation, summing over the area(in 3D) of all triangles constructed from neighbouring pixels.

- $s_3(i)$ Maximum span in 3D
  The length of the maximum distance between two pixels belonging to the region.

- $b_3(i, j)$ Length of shared border in 3D
  Summation over the distances within a chain of pixels, belonging to the border between two parts. Two regions can have several unconnected chains of pixels, due to occlusion, which are summed up to produce the total length. This is a binary attribute, depending on the identity of the second part. To fit such an attribute into the given format of a feature-space image a trick was used. Since we know, that the maximum number of parts is given and fixed, the lengths of shared borders with every part are enumerated. In our case, for convenience, four values are stored, each indexed by the label of the second part. The length of the shared border of a part with itself is defined to be the total length of its border.

- $l(i)$ Mean intensity value

- $v_l(i)$ Variance of intensity values
- $f_2(i)$ Area in 2D
  The area of each region, computed by triangulation, but after projection onto the camera plane.
- $s_2(i)$ Maximum span in 2D
- $b_2(i,j)$ Length of shared border in 2D

B.1.2 Construction of Graph Representations

This part of the original CLARET program, handling the creation of the relation graphs, was split off into a separate program called CGRAPH.

In a first step, the relations between all parts, vectors of binary attribute values are computed from the primary attributes.

From this fully connected structure those relations are selected for the graph, which fit the desired structure.

Binary Attributes

Using the intermediary attributes described in the previous section, the following list of binary attributes is computed. Most of the attributes are asymmetric, i.e. in general $A(i,j) \neq A(j,i)$. The attributes can be grouped, according to their dependency on depth values[1]. All values are normalised, so the input representation is scale invariant as well as invariant to the absolute level of brightness. The computed attributes form an attribute vector $\vec{A}(i,j)$, describing the relation between parts i and j.

- Geometric 3D
  - $D_3(i,j)$ Distance in 3D between two parts
    The distance is normalised by the mean distance.
    $D(i,j) = |\vec{p}(i) - \vec{p}(j)| / \frac{1}{N} \sum_{l,k>l} (|\vec{p}(l) - \vec{p}(k)|)$.
  - Angles in 3D
    Angles between two vectors $\angle(\vec{d_i}, \vec{d_j})$ are computed using the following formula:

$$\angle(\vec{d_i}, \vec{d_j}) = \arcsin\left(\frac{\vec{d_i} \times \vec{d_j}}{|\vec{d_i}||\vec{d_j}|} \frac{[\vec{d_i} \times \vec{d_j}]_z}{|[\vec{d_i} \times \vec{d_j}]_z|}\right); \tag{B.1}$$

    See sec. B.1.2 how the angle between to parts is defined.

[1] Note that of course there can be strong correlations between the groups. For instance the area of a part measured in 2D and measures as a 3D-surface are bound to be correlated.

  - $F_3(i,j)$ Relative area in 3D
    The difference of the areas of part i and part j is normalised by the area of part i:
    $F(i,j) = (f(i) - f(j))/f(i)$.
  - $S_3(i,j)$ Maximum span in 3D
    The difference of the spans of part i and part j is normalised by the span of part i:
    $S_3(i,j) = (s_3(i) - s_3(j))/s_3(i)$.
  - $B_3(i,j)$ Length of shared border in 3D
    The length of the border is normalised by the total border length of part i :
    $B_3(i,j) = b_3(i,j)/b_3(i,i)$.
  - $Z(i,j)$ Difference in depth
    The difference of the mean depths of part i and part j is normalised by the depth of part i. $z(i)$ denotes the z or depth component, i.e. the distance from the virtual camera, of the coordinate vector $p(i)$ of the centre of part i:
    $Z(i,j) = (z(i) - z(j))/z(i)$.
  - $V_z(i,j)$ Variance of depth
    The difference of the variances of part i and part j is normalised by the variance of part i:
    $V_z(i,j) = (v_z(i) - v_z(j))/v_z(i)$.

- Geometric 2D
  The following attributes are computed in analogy to the 3D geometric attributes, using the corresponding 2D primary attributes and discarding the z-coordinate values.

  - $D_2(i,j)$ Distance
  - Angle
  - $F_2(i,j)$ Area
  - $S_2(i,j)$ Span
  - $B_2(i,j)$ Border length
  - $L(i,j)$ Mean intensity
    The difference of the mean intensities of part i and part j is normalised by the mean intensity of part i:
    $L(i,j) = (L(i) - L(j))/L(i)$.

– $V_l(i,j)$ Variance of the intensity
The difference of the variances of the intensities of part i and part j is normalised by the variance of part i:
$V_l(i,j) = (v_l(i) - v_l(j))/v_l(i)$.

Geometric 3D attributes are computed using a range image providing depth values. This range image is constructed from the rendered projections of the objects, as the observers see it during learning. Note that the use of such a range image is only for convenience. The same depth information could be constructed from the knowledge that the objects are composed of four identical spheres and the fact that a perspective camera is used. By estimating the apparent radius of a sphere from the curvature of its boundary the depth values of its visible surface can be computed.

Geometric 2D attributes and intensity based attributes are computed using only the projection of the objects on the computer screen, the same views the human observers are presented with during learning.

The difference in the variance of the depth values over all visible pixels of a sphere, has the following interpretation. Since all four spheres, of which each object consists, are identical, also the variance of an unoccluded sphere is a constant. Depending on how much a sphere is occluded, the variance of depth changes. With increasing occlusion the variance first increases until half the sphere is occluded and then drops to zero, as finally only the outer rim is visible (the variance in brightness of this part should be closely correlated with this attribute).

Implementing Higher Order Attributes

Given the symmetry properties of the objects, it should not be sufficient to use binary attributes alone. Also higher order relations are necessary to distinguish between the objects. In our case the angle as a tertiary (or quaternary) attribute was implemented. Since CLARET-2 only supports binary relations, angles have to be implemented using a trick. In the same way the representation of a binary attribute within an unary framework was solved in the case of border lengths, angles can be implemented by enumerating the possible combinations, again assuming a fixed maximum number of parts.

Given the set of four parts $P = \{a, b, c, d\}$ each object consists of, the angle between two parts $A(a,b)$ is defined as the angle between the two connecting lines:

$$A(a,b) = \angle(\vec{D_{ab}}, \vec{D_{cd}}); \tag{B.2}$$

with $\vec{D}(ij) = \vec{p}(i) - \vec{p}(j)$ the connecting line between the centres of part i and part j. Should one sphere be completely occluded, the angle is defined to be zero $A(a, b) = 0$.

Graph Representations of Derived Binary Attributes

The binary attributes are organised in a graph to capture the structure of an object view. The nodes of this graph are labelled with part indices, the edges labelled with attribute vectors. The computation of this graph structure can involve two main stages, first creating a pool of relations and then optionally selecting relations from this pool to satisfy constraints regarding the minimum and maximum number of edges originating from a node (so called "valency")

For creating the pool of relations two criteria can be selected via commandline parameters:

Fully connected The initial relations connect all parts with each other, resulting in a total number of relations of $N_r = N_p(N_p - 1)$.

Connect adjacent parts Only parts which are adjacent to each other, i.e. share a border, are connected by relations.

For graphs representing objects with comparatively few parts, as in our case, the complete pool can be used to represent the object view. With complex objects having a large number of parts these initial pools of relations can be further reduced, to also reduce the computing time required by the CLARET-2 algorithm. For this end the minimum and the maximum number of edges originating from a node can be specified.

From the sorted list of relations, those are selected, where both parts have less than the desired minimum number of relations. This is the first pass of selection. After connecting all parts appropriately, in a second pass, all those relations are selected, connecting a part, which has less than the minimum number of relations. Preferably, a relation is chosen where the second part has not reached the maximum number of connections. The list of relations is constantly kept sorted first by the number of connections both parts already have and second by the distance between the parts. So if a several relations connect parts with the same number of edges, the relation with the smallest distance is selected. The result of this approach a graphs with a more regular structure, since relations between parts with smaller numbers of connections are selected first.

There is a final check making sure that the resulting graph does not dissociate into two unconnected subgraphs. In that case an additional relation creating a connection between them is selected.

B.2 Partitioning Attribute Spaces

In general a pattern is given as a graph, the vertices corresponding to labelled parts, the edges corresponding to vectors of binary attributes between the parts. The attributes are not necessarily symmetric, therefore the edges are directed.

The edges are represented as binary relations $R(P_i, P_j)$ between labelled parts P_i, P_j consisting of a vector of binary attributes $\vec{A}(i, j)$:

$$R(P_i, P_j) = \vec{A}(i, j), R(P_i, P_j) \neq R(P_j, P_i) \tag{B.3}$$

B.2.1 Attribute Space

The dimension of the attribute space equals the number of attributes used. Every relation in an input graphs is represented as a point in this space. So every pattern is represented as a set of points as large as the number of relations.

In other words, the attribute space contains all possible subgraphs of the input graphs, consisting of two nodes (parts) connected by an edge (relation).

B.2.2 Partition Representation

Partitions of the attribute space are given by a set of minimum and maximum values for each attribute dimension

$$V_k = [\vec{vl}_k, \vec{vh}_k] \tag{B.4}$$

with V_K partition k, $\vec{vl}_k, \vec{vh}_k$ the vector of the lower and the upper boundaries of the partition. In other words every partition is a hyperrectangle in attribute space. A relation $R(P_i, P_j)$ lies within a partition V_k, if its attribute vector lies within the partition as defined by its upper and lower boundaries.

B.2.3 Conditional Attribute Space

To capture more of the object's structure larger subgraphs need to be represented. This is done by relationally extending a given subgraph.

A subgraph consisting of parts P_i, P_j connected by relation $R(P_i, P_j)$ is extended by forming all possible subgraphs, consisting of three parts and two relations.

$$R(P_i, P_j), R(P_j, P_k) \tag{B.5}$$

The subgraphs $R(P_j, P_k)$ are represented as points in another attribute space, of same dimensionality as the original one. The number of points depends on the set of relations $R(P_i, P_j)$ which are extended, hence the term "conditional attribute space". To avoid a combinatorical explosion, conditional attribute spaces are only created for the relations within a partition of the higher level attribute space, not for the complete space. For every partition a separate conditional attribute space can be created. The associated metaphor is to follow chains through the graphs, matching the parts as you go from one relation to the next. Within such a conditional attribute space a single point, i.e. relation, can belong to more than one chain. For instance $R(P_m, P_n)$ could belong to $R(P_i, P_j), R(P_m, P_n)$ and $R(P_k, P_l), R(P_m, P_n)$, if both $R(P_i, P_j)$ and $R(P_k, P_l)$ are within the partition being extended.

B.2.4 Partitioning

New partitions are generated by splitting an existing partition along a selected axis or attribute. The resulting two new partitions are relationally extended by creating two new additional attribute spaces. Each new space is populated with all the relations connected to one of the relations within the parent partition.

For every partition a heuristic partition measure PM is computed. The measure is computed for each attribute dimension separately and PM of a partition is the maximum over these separate measures. So a possible candidate split is found for every existing partition, together with its respective PM.

These candidates for partitioning are kept in a list, which is sorted by PM. Partitions with the highest measure are split first. For the resulting four partitions, two partitions from the split and two from their relational extensions, the partition measure is again calculated. Then they are inserted into the sorted list of partitioning candidates.

The measure itself is supposed to approximate the minimisation of variances within partitions and maximising the variances between partitions. adding several balancing ingredients.

- Weighted distance
  This is the most important measure, taking the difference between a data point and the preceeding point, measured along the currently investigated attribute dimension. The measure is weighted by the mean

of all differences between neighbouring attribute values:

$$d = \frac{max(a_i - a_{i-i})}{\frac{\sum_j (a_j - a_{j-1})}{N}} \tag{B.6}$$

With a_i the attribute value of data point i, N the total number of data points and d the resulting weighted difference.

- Path centreing
  Centreing of chains c.
  $$c = \frac{max(n_{cl}, n_{ch})}{min(n_{cl}, n_{ch})} \tag{B.7}$$
  with n_{cl} the number of chains of all attribute values below and n_{ch} the number of chains of all attribute values higher than the current value. In a conditional attribute space each value can be the endpoint of one or more chains of relations (called chains). This term favours central splits over peripheral ones. The resulting partitions contain approximately the same numbers of chains.

- Number of splits
  The number of splits sp for the attribute dimension i in this conditional attribute space. This term favours splitting of attribute dimensions with less splits than others, thus trying to make use of all the available information.

- Command line attribute weights
  As a commandline parameter a vector of weights ap_j can be specified, for each attribute dimension j, when starting the program. This vector gives the relative probability of a split for each attribute dimension, thus favouring or disfavouring the splitting of some attribute dimensions over others. In the extreme, the probability can be set to zero for particular dimensions, thus excluding them completely from the analysis. This feature was used in the attribute-specific analysis (e.g. Sec. 5.2.3, selecting only a subset of attributes, without the need to recompute the graph representation for each case.

- Combination
  The measures described above, weighted distance, chain centreing, number of splits, and a priori split probability, computed for a specific attribute dimension a are combined into an attribute specific partition measure pm_a:
  $$pm_a = \frac{d^2 * c * ap}{(sp+1)^2} \tag{B.8}$$

The final partition measure PM is the maximum of pm_a over all attribute dimensions:

$$PM = max(pm_a) \tag{B.9}$$

B.2.5 Relational Extension

A given partition contains a set of directional relations, each connecting two parts (P_i, P_j). Relationally extending a partition means selecting all relations containing the second part. These relations form a conditional attribute space, which in turn can be further partitioned.

As part labelling is observed, relational extension selects chains through the given input graphs, connecting $(P_i, P_j), (P_j, P_k), \cdots$. The length of such chains contained within an extended partition is the relational depth of that partition.

Only acyclic chains are allowed. To insure this, all chains are stored, for every partition. This results in a memory consumption, which is exponential with the relational depth.

B.3 Matching

For every learned pattern a matching exists. The point of these matchings is to establish a mapping between the parts or nodes of the learned and the unknown pattern. A matching consists of a set of partitions, a part mapping and a probability for the match between the unknown pattern and this existing pattern. Partitions created by relational extension and splitting of existing partitions are added to a matching, if they are compatible with the partitions already contained in the matching. For every partition belonging to a matching the probability for the unknown pattern being this particular learned pattern can be computed. The total probability for a learned pattern is computed by combining the individual probabilities of all partitions. Finally, a classification probability can be computed, by combining all the total probabilities of the learned patterns, belonging to the same class.

B.3.1 Matching Algorithm

Initially a matching list is created, containing all learned patterns. For every pattern the initial matching contains the root partition, which is the whole attribute space. The part mapping is completely unspecific.

Matching is done in units. In every unit the following actions are repeated for all learned patterns, which are in the matching list.

Splitting The partition in the list of splitting candidates with the largest partition measure is split.

Compatibility check and part mapping update The children are checked, whether they are compatible with the existing partitions in the matching and the part mapping is updated.

Relational extension and partition measure update The compatible partitions are relationally extended. The partition measures for the compatible partitions and their extensions are added to the list of splitting candidates.

Probability The probability for the new matching is computed.

Pruning If the probability lies below a given threshold parameter, this learned pattern is removed from the matching list.

B.3.2 Compatibility of Rules and Part Mapping

During learning, partitions are generated as mentioned above. But not every globally existing partition is added to a particular matching between unknown and a learned pattern. First of all, the partition needs to contain relations from both the learned and the unknown pattern. Next, the candidate partition has to make a consistent statement in itself, about which part of the unknown pattern maps to which parts of the learned pattern. Finally the part mapping derived from the candidate has to be consistent with the existing part mapping, derived from the other partitions contained in the matching.

Part Mapping Information from Partition

In the beginning the mapping between parts of the learned and the unknown pattern is completely unspecific. Described as a binary connectivity matrix, all cells contain a true value. The goal is to refine this mapping to a one-to-one mapping, if possible. For this the mapping information of all partitions belonging to a matching have to be combined.

A partition contains relations belonging to the unknown and the learned pattern. Each relation connects two parts in a directed fashion, i.e.

$$rel(P_i, P_j) \neq rel(P_j, P_i) \tag{B.10}$$

for the learned pattern and

$$rel(Q_i, Q_j) \neq rel(Q_j, Q_i) \tag{B.11}$$

for the unknown pattern. Every relation in the partition provides two mappings of the parts of the learned pattern P_1 and P_2 onto the parts of the unknown pattern Q_1 and Q_2. Note that these sets don't have to have the same size. Since the relations are directed, these two mappings are to be considered separately.

A mapping of parts P_i onto parts Q_i gives three kinds of information, it constrains possible mappings, it gives evidence for a mapping and it says whether there has to exist a mapping for a certain part. How does it constrain mappings? Consider following simple example:
Let us assume having a learned pattern with 4 parts and an unknown pattern with 3 parts. The initial mapping possibilities would look like the matrix in Table B.1.

	Q_1	Q_2	Q_3
P_1	1	1	1
P_2	1	1	1
P_3	1	1	1
P_4	1	1	1

Table B.1: Initial constraint matrix

The mapping

$$M_1 : (P_2) \rightarrow (Q_2, Q_3) \tag{B.12}$$

gives us the information, that the learned part P_2 maps onto either Q_2 or Q_3. It cannot map onto the part Q_1. The subsequent constraints are shown in Table B.2(a).

	Q_1	Q_2	Q_3
P_1	1	1	1
P_2	0	1	1
P_3	1	1	1
P_4	1	1	1

(a) M_1

	Q_1	Q_2	Q_3
P_1	1	1	1
P_2	1	1	1
P_3	1	0	1
P_4	1	0	1

(b) M_2

	Q_1	Q_2	Q_3
P_1	1	1	1
P_2	0	1	1
P_3	1	0	1
P_4	1	0	1

(c) M_1 & M_2

Table B.2: Two constraint matrices and the result of their combination by a logical AND-operation

The next mapping might be

$$M_2 : (P_1, P_2) \rightarrow (Q_2) \tag{B.13}$$

with the corresponding matrix in Table B.2(b). The combination of both mappings is determined by a simple logical AND-operation between the two constraint matrices (Table B.2(c)).

We add a third mapping of the form

$$M_3 : (P_1) \rightarrow (Q_3) \tag{B.14}$$

with the corresponding constraint matrix shown in Tab. B.3(a). Now the whole riddle seems to be solved, as can be seen from the resulting total constraint matrix in Table B.3(b). We see, P_1 can only map onto Q_3 and P_2 onto Q_2, and vice versa. Nevertheless, there is still an ambiguity since Q_1 can map onto both P_3 and P_4. This is, were we need to look at the second

	Q_1	Q_2	Q_3
P_1	0	0	1
P_2	1	1	0
P_3	1	1	0
P_4	1	1	0

(a) M_3

	Q_1	Q_2	Q_3
P_1	0	0	1
P_2	0	1	0
P_3	1	0	0
P_4	1	0	0

(b) M_1 — M_3

	Q_1	Q_2	Q_3
P_1	0	1	1
P_2	0	1	1
P_3	0	0	0
P_4	0	0	0

(c) Evidences M_1 — M_3

	Q_1	Q_2	Q_3
P_1	0	0	1
P_2	0	1	0
P_3	0	0	0
P_4	0	0	0

(d) Mapping M & E

Table B.3: The result of combining M_1 — M_3. Both mappings and evidences, and the result of the logical AND-operation, are shown

information provided by a mapping, the evidences (Tab. B.3(c)). Looking at the mappings we see, that we have evidence for

- P_1 mapping onto Q_3 in M_3
- P_2 mapping onto Q_2 in M_1 and M_2

There is no evidence that either P_3 or P_4 maps onto Q_1, so these parts have to remain unmapped. The evidences can also be stored within a boolean matrix. Combination of the evidences of different mappings is now done by a logical OR-operation, adding up the evidences. The actual mapping is

found by a logical AND-operation of constraint matrix and evidence matrix (Tab. B.3(d)).

What happens, if we add another mapping M_4?

$$M_4 : (P_3) \rightarrow (Q_1, Q_2) \tag{B.15}$$

The constraint matrix will not change at all (Tab. B.4(a)). But now we have evidence, that P_3 can indeed map to Q_1 and we have a one-to-one mapping (Tab. B.3(c)), except for the part P_4, which is to be expected, since 3 unknown parts can only have a mapping to 3 learned parts, leaving one learned part unmapped.

	Q_1	Q_2	Q_3
P_1	0	0	1
P_2	0	1	0
P_3	1	0	0
P_4	1	0	0

(a) M_1 — M_4

	Q_1	Q_2	Q_3
P_1	0	1	1
P_2	0	1	1
P_3	1	1	0
P_4	0	0	0

(b) Evidences

	Q_1	Q_2	Q_3
P_1	0	0	1
P_2	0	1	0
P_3	1	0	0
P_4	0	0	0

(c) M & E

Table B.4: The result of combining M_1 — M_4. Both mappings and evidences, and the result of the logical AND-operation, are shown

Compatibility of Mappings

Let us assume, we again have the three mappings

$$M_1 : (P_2) \rightarrow (Q_2, Q_3) \tag{B.16}$$
$$M_2 : (P_1, P_2) \rightarrow (Q_2) \tag{B.17}$$
$$M_3 : (P_1) \rightarrow (Q_3) \tag{B.18}$$
$$\tag{B.19}$$

providing the constraint and evidence matrices as in Table B.3

What happens, if we now try to add a fourth mapping

$$M_4\prime : (P_3) \rightarrow (Q_2) \tag{B.20}$$

leading to the constraint matrix in Table B.5(b).

Obviously this is not compatible with the result from M_2, which states that Q_2 has to map onto either P_1 or P_2. This is also seen in the resulting – hypothetical – constraint matrix (Tab. B.5(c)).

	Q_1	Q_2	Q_3
P_1	0	0	1
P_2	0	1	0
P_3	1	0	0
P_4	1	0	0

(a) M_1 — M_3

	Q_1	Q_2	Q_3
P_1	1	0	1
P_2	1	0	1
P_3	0	1	0
P_4	1	0	1

(b) $M_4\prime$

	Q_1	Q_2	Q_3
P_1	0	0	1
P_2	0	0	0
P_3	0	0	0
P_4	1	0	0

(c) M_1 — $M_4\prime$

Table B.5: Result of adding an incompatible mapping

There is no possible mapping left for the parts P_2, P_3 and Q_2. This fact alone would not be sufficient, after all it could be justified by the data, that there is no mapping for those parts. Here is, where we need the third kind of information provided by a mapping, the existence information for a mapping. Consider for example M_1. This mapping states, that there has to exist a mapping for P_2 and there can exist mappings for Q_2, Q_3. If a combination with another mapping would lead to the row P_2 vanishing, it would be incompatible. If, on the other hand, it would lead to a vanishing column Q_2 or Q_3, it would still be compatible, since M_1, does not ascertain the existence of a mapping for those two parts at once.

Probabilities

In principle the probability $p_i(U, L_a)$, of an unknown pattern U matching a particular learned pattern L_a, gained from a single partition i , is computed using an exponential distribution.

$$p_i(U, L_a) = \frac{1}{N_i} e^{-P_0 * |r(U) - r(L_a)|} \tag{B.21}$$

$$N_i = \sum_b e^{-P_0 * |r(U) - r(L_a)|} \tag{B.22}$$

Here, $r(U)$ and $r(L_a)$ denote the number of relations contained in partition i for the unknown and the learned pattern, respectively. N_i is a normalising factor, the sum of the matching probabilities over all learned patterns, thus normalising the total probability to 1. P_0 is a weighting parameter, determined by commandline argument. It is the probability of a single relation missing. The difference in the number of relations is weighted exponentially, to get the probability of a matching. The closer the numbers match, the more similar the two patterns are and the higher the probability. It is assumed, that the event of a missing relation has a certain probability P_0 and

occurs independently of the other relations. Therefore, if $n = |r(U) - r(L_a)|$ relations are missing, these are treated as independent events, subsequently the total probability is the product of each single probability, leading to the exponential distribution.

In case the partition was created after one or more relational extensions, some modifications have to be added.

$$p_i(U, L_a) = \frac{1}{N_i} e^{-\frac{P_0 * |c(U) - c(L_a)|}{v(l)}} \tag{B.23}$$

$$N_i = \sum_b e^{-\frac{P_0 * |c(U) - c(L_a)|}{v(l)}} \tag{B.24}$$

Here $p_i(U, L_a)$ again denotes the probability gained from partition i of the unknown pattern U matching the learned pattern L_a and N_i is a normalising factor. The difference is that now the number of chains $c(U)$ and $c(L_a)$ belonging to the unknown and learned pattern is used. In a conditional, extended, attribute space, a missing (or additional) relation can give rise to a larger number of missing chains, which would have contained this relation. On average the number of chains depends on the number of relations as follows

$$c(X) = r(X) * v(l) \tag{B.25}$$

$$v(l) = \prod_{j=0}^{(l-1)} max(v - j, 1) \tag{B.26}$$

$$v = N_r / N_p \tag{B.27}$$

where $c(X)$ and $r(X)$ are the number of chains respectively relations of pattern X within a partition i. v is a global measure of the graphs, the mean valency, computed by dividing the total number of relations N_r by the total number of parts N_p. It is a measure of the density of connections. $v(l)$ is the correcting factor, depending on the mean valency of the graph v and the current level of relational extension l. Since circular paths are prohibited, the expected factor v^l is corrected for each additional level, by subtracting 1 from v. Border effects, arising from the finite sizes of the graphs are ignored.

Again assuming that each partition is an independent measurement of the similarity between unknown and learned pattern, the total probability of a matching is computed by multiplying the probabilities of all partitions belonging to this matching

$$p(U, L_a) = \prod_i p_i(U, L_a) \tag{B.28}$$

The classification probability $P(U, C_j)$, of the unknown pattern U belonging to class C_j is determined, by summing over all learned patterns belonging to this class.

$$P(U, C_j) = \frac{\sum_{a \in C_j} p(U, L_a)}{\sum_j p(U, C_j)} \tag{B.29}$$

B.3.3 Measuring Matching Quality

Part multiplicity *pm* provides a numerical value for the quality of the part correspondence found between a known and an unknown pattern. The ratio is taken between the number of parts N_t, which have to be matched, the number of total mappings so far N_m and the number of unmapped parts N_u.

$$pm = \frac{N_t}{N_m + 2 * N_u} \tag{B.30}$$

The number of parts to be matched N_m is the minimum of the number of parts of the learned and the unknown pattern. N_m captures possible ambiguities of the matching. If a part can be matched to several parts this value will be larger than the total number of parts. N_u captures the missing matchings, i.e. parts that should have a matching, but haven't. This value is adjusted by the difference between the number of current and existing parts. If there are more current than existing parts, not every current part can be matched, and vice versa. Further it is weighted to count as if it was a double matching, otherwise it wouldn't decrease the measure (Consider the example of an empty mapping matrix, with $N_m = 0$ and $N_t = N_u$).

A part multiplicity $pm = 1$ means a one-to-one mapping of all parts. As long as $pm < 1$, there is no one-to-one mapping.

As an example, consider following mapping matrix:

	Q_1	Q_2	Q_3
P_1	1	1	0
P_2	0	0	0
P_3	0	0	1
P_4	0	0	0

There are 3 existing parts and 4 current parts.

$$N_t = 3; \tag{B.31}$$

$$N_m = 3; \tag{B.32}$$

$$N_u = 2 - 1; \tag{B.33}$$

$$pm = \frac{3}{3 + 2 * 1} = \frac{3}{5} \tag{B.34}$$

Appendix C

Data and Statistical Methods

C.1 Trimmed Means

One of the most often used statistical measure is the mean, which gives information about the "central tendency" of a variable. Based on the mean and the variance of a variable more complex descriptions and tests, such as confidence intervals and significance tests are computed. Successful usage of the mean as a description of the data depends strongly on the shape of the underlying distribution, which is assumed to be normal. As Wilcox (2001) argues, the prevailing assumption that, for instance, Student's t-test is not very sensitive to the shape of the distribution, can lead to a serious misjudgement of the actual error level of the performed test. Nevertheless, if the underlying distribution *i*s normal, using means is optimal, as they converge fastest to the true result. One of the methods proposed by Wilcox (2001) to overcome this dilemma is to use *t*rimmed means. Trimmed means are a compromise between the mean and the, more robust, median. To compute a 20% trimmed mean of a sample of size 10, the 2 highest and the two lowest values would be removed from the sample. The median is then identical with a 50% trimmed mean. For a more in depth discussion of the subject and links to source code for conducting various statistical tests see Wilcox (2001).

C.2 Distance Measure of Answering Matrices

The distance between two answering matrices is computed by dividing the root squared difference between the elements by the total number of views.

$$D = \frac{1}{V}\sqrt{\sum_{i=1,j=1}^{VC} (m_a(i,j) - m_b(i,j)^2} \tag{C.1}$$

V denotes the number of views (i.e. rows) and C the number of objects (i.e. columns), and $m_{ab}(i,j)$ the element (i,j) of matrix M_{ab}. The result is divided by the number of views V to allow a better comparison between matrices obtained during learning and generalisation (which have different numbers of views).

C.3 χ^2–Analysis

A standard method to evaluate the quality of a fit is to apply a χ^2–analysis. The procedures used here were adapted from Kreyszig (1974). Computing the weighted differences of the elements of two matrices

$$d(i,j) = \sqrt{\frac{(\bar{M}_a(i,j) - \bar{M}_b(i,j))^2}{\sigma^2_{M_a}(i,j) + \sigma^2_{M_b}(i,j)}} \in N(0,1) \tag{C.2}$$

these are assumed to be from a normal distribution, with a mean $\mu = 0$ and a standard deviation of $\sigma = 1$. So the sum

$$\chi^2 = \sum_{i=1,j=1}^{VC} d^2(i,j) \tag{C.3}$$

with V the number of views and C the number of objects, follows the χ^2–distribution with $V * C$ degrees of freedom. Since not all entries of the matrices are independent, as each row is normalised to $\sum_{j=1}^{C} \bar{M}(i,j) = 1$, the degrees of freedom are adjusted accordingly to $N = V * (C-1)$.

Also data points $d(i,j)$ with $\sigma = 0$ are discarded from the sum, adjusting the degrees of freedom of the χ^2–distribution accordingly.

C.4 T–Test

To check the hypothesis, whether the subjects with haptic experience are better than the subjects with no prior knowledge, a directional test of fit

is used, the t–test. The procedures used here were adapted from Kreyszig (1974). First the differences between the elements of the two matrices are computed, weighted by the standard deviation of the elements.

$$\begin{aligned} d\,'(i,j) &= \begin{cases} \bar{M}_a(i,j) - \bar{M}_b(i,j) & , \text{ if answer(i,j) correct} \\ \bar{M}_b(i,j) - \bar{M}_a(i,j) & , \text{ if answer(i,j) wrong} \end{cases} \\ d(i,j) &= \frac{d\,'(i,j)}{\sqrt{\sigma^2_{M_a}(i,j) + \sigma^2_{M_b}(i,j)}} \in N(\mu_d, \sigma_d) \end{aligned} \tag{C.4}$$

The weighted difference values $d(i,j)$ are assumed to be independently drawn from a normal distribution with unknown standard deviation σ_d. Then a value T, derived by computing

$$\begin{aligned} T &= \sqrt{N}\frac{\bar{d} - \mu_0}{S} \\ \bar{d} &= \frac{1}{N}\sum_{i,j=1}^{N} d(i,j) \\ S &= \frac{1}{N-1}\sum_{i,j=1}^{N} (d(i,j) - \bar{d})^2. \end{aligned} \tag{C.5}$$

follows Student's t–distribution, with $N-1$ degrees of freedom.

Data points $d(i,j)$ with $\sigma = 0$ are discarded from the sum, adjusting the degrees of freedom of the t–distribution accordingly.

We can distinguish three cases,

- $\mu_d < 0$: matrix $\bar{M}_a$ is better than matrix $\bar{M}_b$,
- $\mu_d > 0$: matrix $\bar{M}_a$ is worse than matrix $\bar{M}_b$,
- $\mu_d = 0$: matrix $\bar{M}_a$ is as good as matrix $\bar{M}_b$.

Testing whether 2 datasets are equal, we test the hypothesis $\mu_d = 0$ for a level α, which is refuted, if $|T| > t_{N-1;1-\alpha/2}$ and accepted otherwise.

C.5 Time-Window for Sampling Data

During learning each subject acquired a set of answering matrices, one for each learning unit. Looking at the dynamics of learning, samples are taken at certain points in time. A sample is taken by applying a Gaussian lowpass

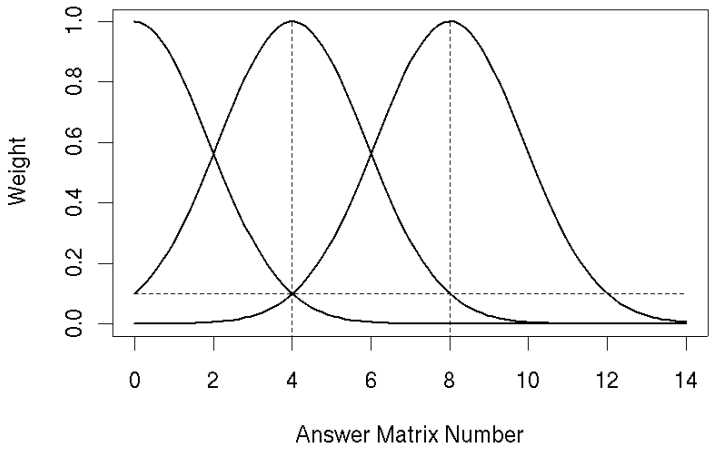


Figure C.1: Sampling from a number of answering matrices. In this example, the sampling points are at matrix #0, matrix #4, and matrix #8. The width σ of the Gaussian window is adjusted so that neighbouring samples have an overlap of 0.1.

filter, with its centre at the sampling point and normalising the rows of the resulting matrix (Fig. C.1). When looking at the overall learning performance, simply the mean over all learning matrices is taken.

So for each subject we now have one or more matrices of relative answering frequencies. For all subjects belonging to a group, the mean, the standard deviation and the standard error were computed, using the following formulas:

$$\begin{aligned} \bar{M}(i,j) &= \frac{1}{N}\sum_{s=1}^{N} m_s(i,j) \\ \sigma^2(i,j) &= \frac{1}{N-1}\sum_{s=1}^{N}(m_s(i,j) - \bar{M}(i,j))^2 \\ SE(i,j) &= \frac{1}{\sqrt{N}}\sigma(i,j) \end{aligned} \tag{C.6}$$

Here i and j are indices of the answering matrix, $m_s(i,j)$ is an element of a sampled matrix of subject s. $\bar{M}$, σ and SE denote the mean, standard deviation and standard error and N is the number of subjects included.

C.6 Example Classification Matrices

Table C.1 shows two examples for observed classification matrices. One is the matrix acquired during a single learning unit, the other is a classification matrix acquired during 26 learning units, with all the answers cumulated.

Obj, view	Answers		
	obj 1	obj 2	obj 3
obj 1, 0	1	0	0
obj 1, 1	0	1	0
obj 1, 2	1	0	0
obj 1, 3	1	0	0
obj 1, 4	0	0	1
obj 1, 5	0	1	0
obj 2, 0	0	1	0
obj 2, 1	0	1	0
obj 2, 2	0	1	0
obj 2, 3	0	1	0
obj 2, 4	0	0	1
obj 2, 5	0	1	0
obj 2, 6	0	0	1
obj 2, 7	0	1	0
obj 3, 0	0	1	0
obj 3, 1	0	1	0
obj 3, 2	0	1	0
obj 3, 3	0	1	0
obj 3, 4	0	0	1
obj 3, 5	0	1	0
obj 3, 6	0	1	0
obj 3, 7	0	0	1

(a) Matrix for a single learning unit

Obj, view	Answers		
	obj 1	obj 2	obj 3
obj 1, 0	26	0	0
obj 1, 1	17	5	4
obj 1, 2	25	0	1
obj 1, 3	25	1	0
obj 1, 4	20	4	2
obj 1, 5	21	4	1
obj 2, 0	0	21	5
obj 2, 1	0	23	3
obj 2, 2	0	19	7
obj 2, 3	0	8	18
obj 2, 4	0	14	12
obj 2, 5	6	9	11
obj 2, 6	0	19	7
obj 2, 7	1	11	14
obj 3, 0	0	14	12
obj 3, 1	0	15	11
obj 3, 2	0	14	12
obj 3, 3	0	15	11
obj 3, 4	0	2	24
obj 3, 5	1	10	15
obj 3, 6	1	11	14
obj 3, 7	2	10	14

(b) Cumulated matrix with relative answering frequencies

Table C.1: Single and cumulated classification matrices

C.7 Template Matching as Similarity Measure

The simplest method human observers could use is to recognise views of objects on a pixel level. Without any further processing, like segmentation,

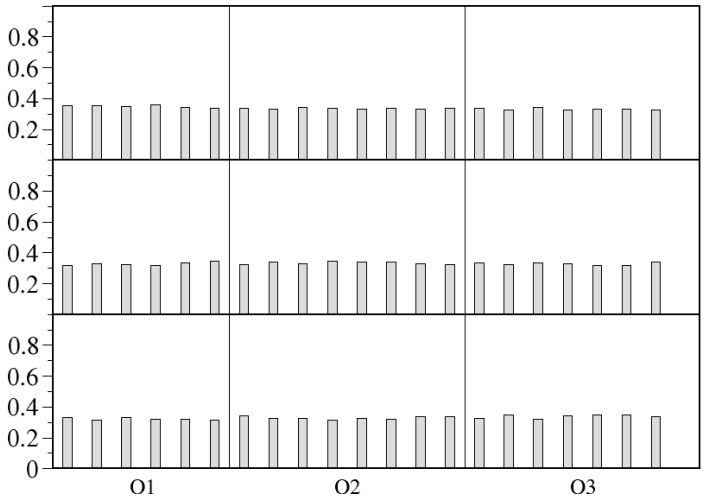


Figure C.2: Classification probabilities predicted by cross-correlation, compared to human observer data. Learning views only

edge detection, feature extraction, they could try to match a seen image to a stored template. These templates are simply the set of learning images, labelled by their object number. The easiest way to do this template matching, is by computing the cross–correlation function between images as a measure of their similarity (Duda and Hart (1973),pp. 276–284,Caelli and Rentschler (1986)).

$$C_L(m,n) = \sum_i \sum_j L(i,j) I(i-m, j-m) \tag{C.7}$$

$$C_L = \frac{\max_{m,n} C_L(m,n)}{\max_{m,n} \sum_i \sum j L(i,j) L(i-m, j-m)} \tag{C.8}$$

Here L is one of the learned object views, I is an unknown view. The similarity measure C is found by taking the maximum value of the cross–correlation function.

The classification probability is then computed by summing over the distances between the unknown view and the learning views for the respective object.

$$P(O_i) = \frac{\sum_{L, L \in O_i} C_L}{\sum_i P(O_i)} \tag{C.9}$$

In a first attempt, this was done for the learning views only. As can be seen (Fig. C.2) the results could in no way predict the actual performance

of the human observers. Similar results were found, using the generalisation views.

Some possible variations were explored, to see whether the results depended on the way the distance measure or the probabilities were formed, but with very similar results. For instance C_L can be computed by integrating over the image, not by taking the maximum. The probability could be computed by using the most similar view for every object and using the cross-correlation as a probability measure. None of these variations led to a better performance, thus eliminating the assumption, that object recognition could be done on such a simple, pixel based level.

Appendix D

Description of OpenInventor Programs

D.1 Scene Files

The inventor file describing the scene containing the objects is fairly simple. In the beginning a perspective camera is defined. The camera is positioned on the z–axis, 10 units distance from the origin, the viewing direction given by orientation is aligned with the axis. The fields position and orientation are set by the controlling program, according to the specifications in the view file. The fields nearDistance and farDistance define clipping planes for rendering. SoXtViewer automatically adjusts them to contain the complete scene.

```
#Inventor V2.0 ascii

Separator {
PerspectiveCamera {
position 0 0 10
orientation 1 0 0  0
nearDistance 5
farDistance  15
focalDistance 10
}
```

The next nodes describe rendering properties. Giving the environment an ambient intensity prevents shaded parts of the object to disappear completely. The material is defined to have a grey colour. These values were set to give a realistic impression of the objects. The complexity node, defines how many triangles are used to render an object, for instance a sphere. The chosen value of 0.7 seemed to yield a good compromise between performance and realism of the rendering.

```
Environment {
```

```
ambientIntensity  0.3
}
Material {
ambientColor   0.5 0.5 0.5
diffuseColor   0.7 0.7 0.7
}
Complexity {
    value   0.7
}
```

The preceding nodes get created with default values by the SoXtExaminerViewer, but the switch node is an essential part of the scene. The controlling program uses the switch node to select which object to render and to blank the scene otherwise. The switch node has 3 children, corresponding to the 3 objects. Each object consists of 4 spheres, which are assembled within a suitable coordinate system by very simple translations. The whole object is rotated and translated, to align the axis of symmetry with the z–axis and placing the centre of the object at the origin.

```
Switch {
whichChild 2
# object 1
Separator {
Rotation { rotation -1 1 0 -0.95531662 }
Translation {translation -.5 -.5 -.5 }
    Separator {
        Sphere {
        }
    }
    Separator {
        Transform {
            translation 2 0 0
        }
        Sphere {
        }
    }
    Separator {
        Transform {
            translation 0 0 2
        }
        Sphere {
        }
    }
    Separator {
        Transform {
            translation 0 2 0
        }
        Sphere {
        }
    }
```

```
}
# object 2
Separator {
Rotation { rotation 0 1 0 -0.78539816 }
Rotation { rotation 1 0 1 -0.261799 }
Translation {translation -0.5 0 -0.5 }
    Separator {
Translation {translation 2 -1 0 }
        Sphere {
        }
    }
    Separator {
        Transform {
            translation 0 -1 0
        }
        Sphere {
        }
    }
    Separator {
        Transform {
            translation 0 1 0
        }
        Sphere {
        }
    }
    Separator {
        Transform {
            translation 0 1 2
        }
        Sphere {
        }
    }
}
# object 3
Separator {
Rotation { rotation 0 -1 0 -0.78539816 }
Rotation { rotation -1 0 1 0.261799 }
Translation {translation 0.5 0 -0.5 }
    Separator {
Translation {translation -2 -1 0 }
        Sphere {
        }
    }
    Separator {
        Transform {
            translation 0 -1 0
        }
        Sphere {
        }
```

```
    }
    Separator {
        Transform {
            translation 0 1 0
        }
        Sphere {
        }
    }
    Separator {
        Transform {
            translation 0 1 2
        }
        Sphere {
        }
    }
}
}
}
```

Following part of the scene graph of the viewer is shown, as it is during the display. The callback node is inserted by the controlling program at the root of the scene, to catch any input from the user, as for instance responses during testing, or required input to preceed after a pause.

```
#Inventor V2.1 ascii

Separator {
    EventCallback {
    }
```

As was mentioned above, position and orientation of the camera are set by the controlling program, keeping the distance from the origin fixed. The fields nearDistance and farDistance are set by the viewer.

```
    Separator {
        PerspectiveCamera {
            position    4.33013 7.5 5
            orientation -0.741024 0.612011 0.276272  1.09428
            nearDistance        6.68296
            farDistance 12.6158
            focalDistance       10
        }
```

A light source, emitting parallel light rays, was inserted automatically by the viewer. It is positioned such, that the direction differs by about 10° to the left and upwards from the viewing direction. To express it in a casual way, the “sun is over your left shoulder”. The rotation node keeps the light aligned with the camera viewing position, and is set automatically by the viewer.

```
        Group {
            Rotation {
                rotation        -0.741024 0.612011 0.276272  1.09428
            }
            DirectionalLight {
                direction       0.2 -0.2 -0.979796
            }
            ResetTransform {
            }
        }
....
```

The rest of the file is unchanged.

D.2 Position files

A file containing a list of views – camera positions – can be passed to the controlling program. The following example shows the view positions used during the learning procedure. In the first line the number of views for each object is specified. In our case there are 6 views for object 1 and 8 views for objects 2 and 3 each. The numbers of views are delimited by "#" characters. Each view is defined in polar coordinates, giving azimuth and elevation in degrees. The third parameter specifies the rotation around the camera axis. The camera's top (0°) is aligned with the vertical y–axis, whenever possible. As can be seen azimuth and elevation are sampled in steps of 60°. The rotation of the camera around its viewing axis is chosen randomly, but also in 60° steps.

```
# 6 8 8 #

[ 0 0 120 ]
[ 0 60 60 ]
[ 0 120 120 ]
[ 0 180 60 ]
[ 60 60 60 ]
[ 60 120 120 ]

[ 0 -0 180 ]
[ 0 -60 60 ]
[ 0 -120 60 ]
[ 0 -180 180 ]
[ 60 -60 0 ]
[ 60 -120 180 ]
[ 120 -60 180 ]
[ 120 -120 60 ]

[ 0 0 0 ]
[ 0 60 120 ]
```

```
[ 0 120 120 ]
[ 0 180 0 ]
[ 60 60 0 ]
[ 60 120 120 ]
[ 120 60 0 ]
[ 120 120 120 ]
```

The next example gives the complete list of camera position used during generalisation. Since azimuth and elevation are now sampled in steps of 30° there is a total of 83 views, of which 19 views are identical to the views used during learning.

```
# 21 31 31 #

[ 0 0 120]
[ 0 30 150]
[ 0 60 60]
[ 0 90 150]
[ 0 120 120]
[ 0 150 0]
[ 30 30 90]
[ 30 60 30]
[ 30 90 0]
[ 30 120 150]
[ 30 150 120]
[ 60 30 90]
[ 60 60 60]
[ 60 90 120]
[ 60 120 120]
[ 60 150 150]
[ 90 30 30]
[ 90 60 60]
[ 90 90 90]
[ 90 120 60]
[ 90 150 120]

[ 0 -0 180]
[ 0 -30 0]
[ 0 -60 60]
[ 0 -90 60]
[ 0 -120 60]
[ 0 -150 180]
[ 30 -30 30]
[ 30 -60 0]
[ 30 -90 90]
[ 30 -120 180]
[ 30 -150 90]
[ 60 -30 120]
[ 60 -60 0]
[ 60 -90 120]
```

```
[ 60 -120 180]
[ 60 -150 180]
[ 90 -30 120]
[ 90 -60 180]
[ 90 -90 60]
[ 90 -120 120]
[ 90 -150 0]
[ 120 -30 90]
[ 120 -60 180]
[ 120 -90 0]
[ 120 -120 60]
[ 120 -150 0]
[ 150 -30 150]
[ 150 -60 180]
[ 150 -90 150]
[ 150 -120 180]
[ 150 -150 60]

[ 0 0 0]
[ 0 30 120]
[ 0 60 120]
[ 0 90 30]
[ 0 120 120]
[ 0 150 120]
[ 30 30 0]
[ 30 60 180]
[ 30 90 90]
[ 30 120 90]
[ 30 150 0]
[ 60 30 120]
[ 60 60 0]
[ 60 90 150]
[ 60 120 120]
[ 60 150 60]
[ 90 30 60]
[ 90 60 90]
[ 90 90 150]
[ 90 120 90]
[ 90 150 30]
[ 120 30 120]
[ 120 60 0]
[ 120 90 180]
[ 120 120 120]
[ 120 150 120]
[ 150 30 30]
[ 150 60 60]
[ 150 90 90]
[ 150 120 180]
[ 150 150 30]
```

Bibliography

N. Ahmed and K. R. Rao. *Orthogonal Transforms for Digital Signal Processing*. Springer, New York, Heidelberg, 1975.

A. Amedi, G. Jacobson, T. Hendler, R. Malach, and E. Zohary. Convergence of visual and tactile shape processing in the human lateral occippital cortex. *Cerebral Cortex*, 12(11):1202–1212, 2002.

A. Amedi, R. Malach, T. Hendler, S. Peled, and E. Zohary. Visuo-haptic object-related activiation in the ventral visual pathway. *Nature Neuroscience*, 4(3):324–330, 2001.

A. Amin, C. Sammut, and K. C. Sum. Learning to recognize hand-printed Chinese characters using inductive logic programming. Technical report, School of Computer Science and Engineering, University of New South Wales, Sydney, Australia, 1996.

G. Avidan-Carmel, M. Harel, T. Hendler, D. Ben-Bashat, E. Zohary, and R. Malach. Contrast sensitivity of human visual areas and its relation to object recognition. *Society for Neuroscience Abstracts*, 30:1846, 2000.

H. Barlow. Redundancy reduction revisited. *Network: Comput. Neural Syst.*, 12:241–253, 2001.

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

H. B. Barlow, R. Narasimhan, and A. Rosenfeld. Visual pattern analysis in machines and animals. *Science*, 177:567–574, 1972.

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

I. Biederman. Recognizing depth-rotated objects: A review of recent research and theory. *Spatial Vision*, 13(2,3):241–253, 2000.

I. Biederman and M. Bar. One-shot viewpoint invariance in matching novel objects. *Vision Research*, 39(17):2885–2899, 1999.

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

I. Biedermann. Human image understanding: recent research and a theory. *CVGIP*, 32:29–73, 1985.

W. F. Bischof. Learning to recognize objects. *Spatial Vision*, 13(2,3):297–304, 2000.

W. F. Bischof and T. Caelli. Learning structural descriptions of patterns: A new technique for conditional clustering and rule generation. *Pattern Recognition*, 27(5):689–697, 1994.

W. F. Bischof and T. Caelli. Scene understanding by rule evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(11):1284–1288, 1997a.

W. F. Bischof and T. Caelli. Visual learning of patterns and objects. *IEEE-SMC*, 27(6):907–917, 1997b.

M. Booth and E. Rolls. View-invariant representations of familiar objects by neurons in the inferior temporal cortex. *Cerebral Cortex*, 8(6):510–523, 1998.

B. B. Boycott and H. Wässle. The morphological types of ganglion cells of the domestic cat's retina. *Journal of Physiology*, 240(2):397–419, 1974.

I. Bratko and S. Muggleton. Applications of inductive logic programming. *Communications of the Association of Computing Machinery*, 38(11):65–70, 1995.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. CRC Press, 1984.

J. Bullier. Feedback connections and conscious vision. *Trends in Cognitive Sciences*, 5(9):369–370, 2001a.

J. Bullier. Integrated model of visual processing. *Brain Research Reviews*, 36(2,3):96–107, 2001b.

H. Bunke. Error-tolerant graph matching: A formal framework and algorithms. In A. Amin, D. Dori, P. Pudil, and H. Freeman, editors, *Advances in Pattern Recognition*, volume 1451 of *Lecture Notes in Computer Science*, pages 1–14. Springer Verlag, 1998.

H. H. Bülthoff and S. Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Natl. Acad. Sci. USA*, 89(1):60–64, 1992.

T. Caelli and W. F. Bischof. *Machine Learning and Image Interpretation.* Plenum Press, New York, 1997a.

T. Caelli and W. F. Bischof. The role of machine learning in building image interpretation systems. *International Journal of Pattern Recognition and Artificial Intelligence*, 11(1):143–168, 1997b.

T. Caelli, C. Dillon, E. Osman, and G. Krieger. The IPRS image processing and pattern recognition system. *Spatial Vision*, 11(1):107–116, 1997.

T. Caelli and A. Dreier. Variations on the evidence-based object recognition theme. *Pattern Recognition*, 26(2):733–740, 1994.

T. Caelli, M. Johnston, and T. Robison. 3D object recognition: Inspirations and lessons from biological vision. In Jain and Flynn (1993), pages 1–16.

T. Caelli, E. Osman, and G. West. 3D shape matching and inspection using geometric features and relational learning. *Computer Vision and Image Understanding*, 72(3):340–350, 1998.

T. Caelli, I. Rentschler, and W. Scheidler. Visual pattern recognition in humans. I. Evidence for the existence of adaptive filters. *Biological Cybernetics*, 57(4,5):233–240, 1987.

T. Caelli, G. West, M. Robey, and E. Osman. A relational learning method for pattern and object recognition. *Image and Vision Computing*, 17:391–401, 1999.

T. M. Caelli and I. Rentschler. Cross-correlation model for pattern acuity. *J. Opt. Soc. Am. A*, 3(11):1948–1956, 1986.

J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey. *Graphical Methods for Data Analysis.* Wadsworth & Brooks/Cole, 1983.

C. Christou and H. H. Bülthoff. View dependence in scene recognition after active learning. *Memory and Cognition*, 27(6):996–1007, 1999.

T. Dean and M. Boddy. An analysis of timedependent planning. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 49–54, St.Paul, MN, 1988. AAAI, MIT Press.

R. Desimone and L. Ungerleider. Neural mechanisms of visual processing in monkeys. In F. Boler and J. Grafman, editors, *Handbook of Neuropsychology*, volume 2, pages 267–299. Elsevier, Amsterdam, 1989.

S. J. Dickinson. Part-based modeling and qualitative recognition. In Jain and Flynn (1993).

R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

S. Duhoux, M. Gschwind, M. Vuilleumier, I. Rentschler, and S. Schwartz. Neural correlates of 3-D object learning. *Perception*, 34, 2005.

S. Edelmann and H. H. Bülthoff. Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, 32(12):2385–2400, 1992.

T. Fan, G. Medioni, and R. Nevatia. Recognizing 3-D objects using surface descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(11):1140–1156, 1989.

M. J. Farah. *Visual Agnosia*. MIT Press, Cambridge, MA, 1990.

D. J. Felleman and D. C. van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex*, 1(1):1–47, 1991.

D. H. Foster and S. J. Gilson. Recognizing novel three-dimensional objects by summing signals from parts and views. *Proc. R. Soc. Lond. B*, 269 (1503):1939–1947, 2002.

I. Fujita, K. Tanaka, M. Ito, and K. Cheng. Colums for visual features of objects in monkey inferotemporal cortex. *Nature*, 360(6402):343–346, 1992.

J. M. Fuster. The prefrontal cortex – an update: Time is of the essence. *Neuron*, 30(2):319–333, 2001.

J. M. Fuster. *Cortex and Mind*. Oxford University Press, Oxford, 2003.

I. Gauthier, P. Skudlarski, J. C. Gore, and A. W. Anderson. Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3(2):191–197, 2000.

H. J. Glock. *A Wittgenstein Dictionary*. Blackwell, Oxford, 1996.

D. M. Green and J. A. Swets. *Signal detection theory and psychophysics*. Wiley, 1966.

K. Grill-Spector, Z. Kourtzi, and N. Kanwisher. The lateral occipital complex and its role in object recognition. *Vision Research*, 41(10,11):1409–1422, 2001.

K. Grill-Spector, T. Kushnir, T. Hendler, and R. Malach. The dynamics of object selective activation is correlated to recognition in humans. *Human Brain Mapping*, 6:187–203, 2000.

W. E. L. Grimson. *Object Recognition by Computer: The Role of Geometric Constraints*. MIT Press, 1990.

O. J. Grüsser and T. Landis. Visual agnosias and other disturbances of visual perception and cognition. In J. Cronley-Dillon, editor, *Vision and Visual Dysfunction*. MacMillan, Houndsmills, Basingstoke, 1991.

S. Haykin. *Neural Networks: A Comprehensive Foundation*. MacMillan, New York, 1994.

W. G. Hayward and M. J. Tarr. Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 23(5):1511–1521, 1997.

W. G. Hayward and M. J. Tarr. Differing views on views: comments on Biederman and Bar (1999). *Vision Research*, 40(28):3895–3899, 2000.

W. G. Hayward and P. Williams. Viewpoint dependence and object discriminability. *Psychological Science*, 11(1):7–12, 2000.

D. D. Hoffman. The interpretation of visual illusions. *Scientific American*, 249(6):156–162, 1983.

D. D. Hoffman and W. A. Richards. Parts of recognition. *Cognition*, 18(1–3): 65–96, 1984.

J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison–Wesley, 1979.

D. H. Hubel. *Eye, Brain and Vision*. Scientific American Library, New York, 1995.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154, 1962.

D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.*, 195:215–243, 1968.

D. H. Hubel and T. N. Wiesel. Functional architecture of macaque monkey striate cortex. *Proc. Royal Soc. Lond.*, 198:1–59, 1977.

J. E. Hummel. Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich and A. Markman, editors, *Cognitive Dynamics: Conceptual Change in Humans and Machines*, pages 157–185. Erlbaum, Hillsdale, NJ, 2000.

J. E. Hummel and I. Biederman. Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99(3):480–517, 1992.

J. E. Hummel and B. J. Stankiewicz. An architecture for rapid, hierarchical structural description. In T. Inui and J. McClelland, editors, *Attention and Performance XVI*, pages 93–121. MIT Press, Cambridge, MA, 1996.

J. E. Hummel and B. J. Stankiewicz. Two roles for attention in shape perception: A structural description model of visual scrutiny. *Visual Cognition*, 5:49–79, 1998.

G. W. Humphreys and S. C. Kahn. Recognizing novel views of three-dimensional objects. *Canadian Journal of Psychology*, 46:170–190, 1992.

D. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, 1990.

A. Ishai, L. G. Ungerleider, A. Martin, J. L. Schouten, and J. V. Haxby. Distributed representation of objects in the human ventral visual visual pathway. *Proc. Natl. Acad. Sci. USA*, 96(16):9379–9384, 1999.

A. K. Jain and P. J. Flynn, editors. *Three-Dimensional Object Recognition Systems*. Elsevier Science Publishers B. V., 1993.

A. K. Jain and R. Hoffman. Evidence-based recognition of 3-D objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6): 783–801, 1988.

T. W. James, G. K. Humphrey, J. S. Gati, P. Servos, R. S. Menon, and M. A. Goodale. Haptic study of three-dimensional objects activates extrastriate visual areas. *Neuropsychologia*, 40(10):1706–1714, 2002.

X. Jiang, A. Münger, and H. Bunke. Synthesis of representative graphical symbols by computing generalized median graph. In *Proc. of the 3rd Int. Workshop on Graphics Recognition*, pages 187–194, Jaipur, 1999.

P. Joliceur, M. Gluck, and S. M. Kosslyn. Pictures and names: Making the connection. *Cognitive Psychology*, 16(22):243–275, 1984.

M. Jüttner, T. Caelli, and I. Rentschler. Evidence-based pattern classification: A structural approach to human perceptual learning and generalization. *Journal of Mathematical Psychology*, 41(3):244–259, 1997.

M. Jüttner, B. Langguth, and I. Rentschler. The impact of context on pattern category learning and representation. *Visual Cognition*, 11(8):921–945, 2004.

M. Jüttner, E. Osman, and I. Rentschler. Visuelles Lernen in virtuellen Realitäten. *Einsichten*, 2:22–25, 2000.

B. Kolb and I. Q. Whishaw. *Neuropsychologie*. Spektrum Akademischer Verlag, Heidelberg, 1996.

E. Kreyszig. *Statistische Methoden und ihre Anwendungen*. Vandenhoek & Ruprecht, Göttingen, 4 edition, 1974.

G. Lakoff. *Women, Fire and Dangerous Things*. The University of Chicago Press, Chicago, 1987.

T. Landis. Disruption of space perception due to cortical lesions. *Spatial Vision*, 13(2,3):179–192, 2000.

S. W. Lee, J. H. Kim, and F. C. A. Gruen. Translation- rotation- and scale invariant recognition of hand-drawn symbols in schematic diagrams. *Int. J. Pattern Recognition and Artificial Intelligence*, 4:1–15, 1990.

T. S. Lee, D. Mumford, R. Romero, and V. A. Lamme. The role of the primary visual cortex in higher level vision. *Vision Research*, 38(15/16): 2429–2454, August 1998.

T. S. Lee, C. F. Yang, R. D. Romero, and D. Mumford. Neural activity in early visual cortex reflects behavioural experience and higher-order perceptual saliency. *Nature Neuroscience*, 5(6):589–597, 2002.

W. Li, V. Piëch, and C. D. Gilbert. Perceptual learning and top-down influences in primary visual cortex. *Nature Neuroscience*, 7(6):651–657, 2004.

Z. Liu, D. Kersten, and D. C. Knill. Dissociating stimulus information from internal representation – a case study in object recognition. *Vision Research*, 39(3):603–612, 1999.

Z. Liu, D. C. Knill, and D. Kersten. Object classification for human and ideal observers. *Vision Research*, 35(4):549–569, 1995.

N. K. Logothetis. Object recognition: Holistic representations in the monkey brain. *Spatial Vision*, 13(2,3):165–178, 2000.

N. K. Logothetis and J. Pauls. Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex*, 5: 270–288, 1995.

N. K. Logothetis, J. Pauls, H. H. Bülthoff, and T. Poggio. View-dependent object recognition by monkeys. *Current Biology*, 4(5):401–413, 1994.

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

S. W. Lu, Y. Ren, and C. Y. Suen. Hierarchical attributed graph representation and recognition of handwritten Chinese characters. *Pattern Recognition*, 24(7):617–632, 1991.

D. Marr. *Vision*. W. H. Freeman, San Francisco, CA, 1982.

D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *PRSL*, 200:269–294, 1978.

B. McCane, T. Caelli, and O. de Vel. Learning and recognising 3D objects using sparse depth and intensity information. *Int. J. Pattern Recognition and Artificial Intelligence*, 11(6):909–931, 1998.

D. McNicol. *A Primer on Signal Detection Theory*. George Allen & Unwin Ltd., 1972.

B. Mel. SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, 9:977–804, 1997.

B. T. Messmer and H. Bunke. Automatic learning and recognition of graphical symbols in engineering drawings. In R. Kasturi and K. Tombre, editors, *Graphics Recognition – Methods and Applications*, volume 1072 of *Lecture Notes in Computer Science*, pages 123–134. Springer Verlag, 1996.

E. Miller. The prefrontal cortex and cognitive control. *Nat. Rev. Neurosci.*, 1(1):59–65, 2000.

E. K. Miller, C. A. Erickson, and R. Desimone. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, 16(16):5154–5167, 1996.

A. D. Milner and M. A. Goodale. *The Visual Brain in Action*. Oxford University Press, Oxford, 1995.

B. Milner. Psychological defects produced by temporal-lobe excision. *Research Publications of the Association for Research in Nervous and Mental Disease*, 38:244–257, 1958.

B. Milner. Visual recognition and recall after right temporal-lobe excision in man. *Neuropsychologia*, 6:191–209, 1968.

C. Moore and S. A. Engel. Neural response to perception of volume in the lateral occipital complex. *Neuron*, 29(1):277–286, 2001.

U. Neisser. *Cognition and Reality. Principles and Implications of Cognitive Psychology*. W. H. Freeman, San Francisco, CA, 1976.

F. N. Newell, M. O. Ernst, B. S. Tjan, and H. H. Bülthoff. Viewpoint dependence in visual and haptic object recognition. *Psychological Science*, 12(1):37–42, 2001.

N. Osaka, I. Rentschler, and I. Biederman, editors. *Object Recognition, Attention, and Action*. Springer, Tokyo, 2007.

E. Osman, M. Jüttner, and I. Rentschler. Überwachtes Objektlernen – Ein neues Paradigma zur Untersuchung des Einflusses von Vorerfahrung auf mentale Objektrepräsentationen. In H. H. Bülthoff, M. Fahle, K. R. Gegenfurtner, and H. A. Mallot, editors, *Beiträge zur 1. Tübinger Wahrnehmungskonferenz*, 1998.

E. Osman, A. R. Pearce, M. Jüttner, and I. Rentschler. Reconstructing mental object representations: A machine vision approach to human visual recognition. *Spatial Vision*, 13(2–3):277–295, 2000.

N. R. Pal and S. K. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

S. E. Palmer, E. Rosch, and P. Chase. Canonical perspective and the perception of objects. In J. Long and A. Baddely, editors, *Attention and Performance IX*, pages 135–151. Lawrence Erlbaum Associates, Hillsdale, NJ, 1981.

Y. Park. A comparison of neural net classifiers and linear tree classifiers: Their similarities and differences. *Pattern Recognition*, 27(11):1493–1503, 1994.

A. Pascual-Leone and V. Walsh. Fast back projections from the motion to the primary visual area necessary for visual awareness. *Science*, 292(5516): 510–512, 2001.

A. Pearce. *Relational Evidence Theory and Spatial Interpretation Procedures.* PhD thesis, School of Computing, Curtin University of Technology, Perth, Western Australia, 1996.

A. Pearce and T. Caelli. Interactively matching hand-drawings using induction. *Computer Vision and Image Understanding*, 73(3):391–403, March 1999.

A. Pearce, T. Caelli, and W. F. Bischof. Rulegraphs for graph matching in pattern recognition. *Pattern Recognition*, 27(9):1231–1247, 1994.

A. R. Pearce, E. Osman, M. Jüttner, and I. Rentschler. Human meets machine vision for learning to recognise objects. In *Proceedings of the International Conference of Machine Learning (ICML–99)*, volume Workshop on Machine Learning in Computer Vision, pages 1–7, Bled, 1999.

D. I. Perret, M. W. Oram, and E. Ashbridge. Evidence accumulation in cell populations responsive to faces: An account of generalisation of recognition without mental transformation. *Cognition*, 67(1–2):111–145, 1998.

A. Peters and B. R. Payne. Numerical relationsips between genicucortical afferents and pyramidal cell modules in the cat primary visual cortex. *Cerebral Cortex*, 3(1):69–78, 1993.

Z. Pizlo. Perception viewed as an inverse problem. *Vision Research*, 41(24): 3145–3161, 2001.

Z. Pizlo and A. K. Stevenson. Shape constancy from novel views. *Perception & Psychophysics*, 61(7):1299–1307, 1999.

T. Poggio and S. Edelmann. A network that learns to recognize three-dimensional objects. *Nature*, 343(6255):263–266, 1990.

T. Poggio and F. Girosi. Networks for approximation and learning. *Proc. IEEE*, 78:1481–1497, 1990.

A. Puce, T. Allison, J. C. Gore, and G. McCarthy. Face-sensitve regions in human extrastriate cortex studied by functional MRI. *Journal of Neurophysiology*, 74:1192–1199, 1995.

J. R. Quinlan. Decision trees and decisionmaking. *T-SMC*, 20(2):339–346, April 1990a.

J. R. Quinlan. Learning logical definitons from relations. *Machine Learning*, 5:1231–1247, 1990b.

R. P. N. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87, 1999.

I. Rentschler, M. Gschwind, H. Brettel, E. Osman, and T. Caelli (in press). Structural and view-specific representation for the categorization of three-dimensional objects. *Vision Research*, 2008.

I. Rentschler and M. Jüttner. Mirror-image relations in category learning. *Visual Cognition*, 15(2):211–237, 2007.

I. Rentschler, M. Jüttner, and T. Caelli. Probabilistic analysis of human supervised learning and classification. *Vision Research*, 34:669–687, 1994.

I. Rentschler, M. Jüttner, E. Osman, and T. Caelli. Multimodal representations for human 3D object recognition. In R. P. Würtz and M. Lappe, editors, *Dynamic Perception*, pages 327–332, Amsterdam, 2002. IOS Press.

I. Rentschler, M. Jüttner, E. Osman, A. Müller, and T. Caelli. Development of configural 3D object recognition. *Behavioural Brain Research*, 149:107–111, 2004.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.

I. Rock and J. DiVita. A case of viewer-centered object perception. *Cognitive Psychology*, 19, 1987.

E. Rosch, C. B. Mervis, G. W. D. Ray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

B. Russell. *History of Western Philosophy*. George Allen & Unwin, London, 2 edition, 1962.

R. N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171, 1971.

N. Sigala and N. K. Logothetis. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415:318–320, 2002.

B. J. Stankiewicz. Empirical evidence for independent dimensions in the visual representation of three-dimensional shape. *Journal of Experimental Psychology: Human Perception and Performance*, 28(4):913–932, 2002a.

B. J. Stankiewicz. Models of perceptual systems. online, 2002b.

B. J. Stankiewicz and J. E. Hummel. Metricat: A representation for basic and subordinate-level classification. In *Proceedings of the 18th Annual conference of the Cognitive Science Society*, pages 254–259, Hillsdale, NJ, 1996. Erlbaum.

Y. Sugase, S. Yamane, S. Ueno, and K. Kawano. Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, 400:869–873, 1999.

N. S. Sutherland. Outline of a theory of visual pattern recognition in animal and man. *Proc. Royal Soc. London B*, B 171:297–317, 1968.

K. Tanaka. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.*, 19:109–139, 1996.

K. Tanaka. Mechanisms of visual object recognition studied in monkeys. *Spatial Vision*, 13(2,3):147–163, 2000.

M. J. Tarr. Rotating objects to recognize them: A case study of the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review*, 2(1):55–82, 1995.

M. J. Tarr. Object recognition. In Nadel L. and Goldstone R., editors, *Encyclopedia of Cognitive Science*, pages 490–494. Nature Publishing Group/MacMillan Publishers Limited, 2002.

M. J. Tarr and H. H. Bülthoff. 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(6):1494–1505, 1995.

M. J. Tarr and S. Pinker. Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21:233–282, 1989.

M. J. Tarr and Q. C. Vuong. Visual object recognition. In H. Pashler and S. Yantis, editors, *Steven's Handbook of Experimental Psychology: Vol. 1. Sensation and Perception*, volume 1, pages 287–314. John Wiley & Sons, 2002.

B. S. Tjan and G. E. Legge. The viewpoint complexity of an object-recognition task. *Vision Research*, 38(15/16):2335–50, August 1998.

H. Tomita, M. Ohbayashi, K. Nakahara, I. Hasegawa, and Y. Miyashita. Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature*, 401:699–703, 1999.

N. F. Troje and H. H. Bülthoff. Face recognition under varying poses: The role of texture and shape. *Vision Research*, 36(12):1761–1771, 1996.

M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.

B. Tversky and K. Hemenway. Objects, parts, and categories. *J Exp Psychol Gen.*, 113(2):169–197, 1984.

S. Ullman. Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32(3):193–254, 1989.

S. Ullman. Three-dimensional recognition based on the combination of views. *Vision Research*, 67(1–2):21–44, 1998.

S. Ullman and R. Basri. Recognition by linear combination of models. *IEEE PAMI*, 13(10):992–1006, 1991.

L. G. Ungerleider and M. Mishkin. Two cortical visual systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, editors, *Analysis of Visual Behavior*, pages 549–586. MIT Press, Cambridge, MA, 1982.

A. Unzicker, M. Jüttner, and I. Rentschler. Similarity-based models of visual recognition. *Vision Research*, 38(15-16):2289–2305, 1998.

D. C. van Essen, C. H. Anderson, and D. J. Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science*, 255(5043):419–423, 1992.

J. Vanrie, B. Willems, and J. Wagemans. Multiple routes to object matching from different viewpoints: Mental rotation versus invariant features. *Perception*, 30(9):1047–1056, 2001.

R. von der Heydt, E. Peterhans, and G. Baumgartner. Illusory contours and cortical neuron responses. *Science*, 224(4654):1260–1262, 1984.

P. Vuilleumier, R. N. Henson, J. Driver, and R. J. Dolan. Mutliple levels of visual object constancy revealed by event-related fMRI of repetition priming. *Nature Neuroscience*, 5(5):491–499, 2002.

G. Wallis and H. H. Bülthoff. Learning to recognize objects. *Trends in Cognitive Sciences*, 3(1):22–31, 1999.

S. Watanabe. *Pattern Recognition; Human And Mechanical*. Wiley, New York, 1985.

R. R. Wilcox. *Fundamentals of Modern Statistical Methods*. Springer–Verlag, New York, 2001.

E. K. Wong. Model matching in robot vision by subgraph isomorphism. *Pattern Recognition*, 25(3):287–304, 1992.

M. P. Young. The architecture of visual cortex and inferential processing. *Spatial Vision*, 13(2,3):137–146, 2000.

Auswahl Publikationen

I. Rentschler, M. Gschwind, H. Brettel, E. Osman, and T. Caelli (in press). Structural and view-specific representation for the categorization of three-dimensional objects. *Vision Research*, 2008

I. Rentschler, M. Jüttner, E. Osman, A. Müller, and T. Caelli. Development of configural 3D object recognition. *Behavioural Brain Research*, 149:107–111, 2004

I. Rentschler, M. Jüttner, E. Osman, and T. Caelli. Multimodal representations for human 3D object recognition. In R. P. Würtz and M. Lappe, editors, *Dynamic Perception*, pages 327–332, Amsterdam, 2002. IOS Press

M. Jüttner, E. Osman, and I. Rentschler. Visuelles Lernen in virtuellen Realitäten. *Einsichten*, 2:22–25, 2000

E. Osman, A. R. Pearce, M. Jüttner, and I. Rentschler. Reconstructing mental object representations: A machine vision approach to human visual recognition. *Spatial Vision*, 13(2–3):277–295, 2000

T. Caelli, G. West, M. Robey, and E. Osman. A relational learning method for pattern and object recognition. *Image and Vision Computing*, 17:391–401, 1999

A. R. Pearce, E. Osman, M. Jüttner, and I. Rentschler. Human meets machine vision for learning to recognise objects. In *Proceedings of the International Conference of Machine Learning (ICML–99)*, volume Workshop on Machine Learning in Computer Vision, pages 1–7, Bled, 1999

T. Caelli, E. Osman, and G. West. 3D shape matching and inspection using geometric features and relational learning. *Computer Vision and Image Understanding*, 72(3):340–350, 1998

E. Osman, M. Jüttner, and I. Rentschler. Überwachtes Objektlernen – Ein neues Paradigma zur Untersuchung des Einflusses von Vorerfahrung auf mentale Objektrepräsentationen. In H. H. Bülthoff, M. Fahle, K. R. Gegenfurtner, and H. A. Mallot, editors, *Beiträge zur 1. Tübinger Wahrnehmungskonferenz*, 1998

Lebenslauf

Name	:	Erol Osman
geboren	:	26.4.1963 in München
Staatsang.	:	deutsch, britisch
Familienstand	:	ledig

1969 – 1983	Grundschule, Gymnasium, Abitur in München
1983 – 1985	Zivildienst
1985 – 1987	5 Semester Magisterstudiengang Germanistische Linguistik an der Ludwig–Maximilians–Universität München
1987 – 1995	15 Semester Diplomstudiengang Physik an der Technischen Universität München
1994 – 1995	Diplomarbeit am Institiut für Medizinische Psychologie
1995	Abschluß des Studiums mit Diplom, Prädikat “sehr gut bestanden”
1995 – 1996	Tätigkeit als Research Associate für Prof. Caelli am Dept. of Computer Science, Curtin University of Technology, Perth
1996	Beginn der Promotion am Institut für Medizinische Psychologie
1996 – 1999	Stipendiat, bzw. Mitglied, des Graduiertenkollegs “Sensorische Interaktion in biologischen und technischen Systemen” der Ludwig-Maximilians-Universität München
1997 – 2002	Wissenschaftlicher Mitarbeiter am Institut für Medizinische Psychologie, Ludwig-Maximilians-Universität München
2002 – 2007	Freiberufliche Softwareentwicklung
2007 –	Softwareentwickler bei CLIPPERnet GmbH