in: Recent Research Developments in Pattern Recognition, Vol. 1,
pp. 179-197, (Part-I), Transworld Research Network, 2000.

Theories of Three-Dimensional Object Perception
A Survey

Gabriele Peters

Institut für Neuroinformatik
Ruhr-Universität Bochum
Universitätsstr. 150
D-44780 Bochum, Germany
Gabi.Peters@neuroinformatik.ruhr-uni-bochum.de

ABSTRACT

In this review current theories of the visual perception of three-dimensional form are introduced. Starting with a brief overview of low-level visual processes, which contribute to the recognition of 3D objects, such as the perception of structure from shading or texture, this article mainly concentrates on advanced theories, which try to explain object recognition from a more elaborate point of view. The differences between object- and viewer-centered representations are treated, as well as the nature of characteristic views. In addition, the recognition of novel views by view interpolation and the importance of temporal context while acquiring an object representation are also dealt with. The described theories are assessed with respect to their biological plausibility. Evidence from psychological studies is given, which either supports or contradicts the different computational models. Besides this, recent results from physiological studies reflect the hierarchical processing of 3D information in the primate brain.

INTRODUCTION

In the study of human visual perception, as well as in technical applications the problem of recognizing the three-dimensional structure of the environment has become a major topic in recent years. The fact that perceiving systems, including our own brains, are able to create a three-dimensional notion of the environment, although the source of information is two-dimensional [1] and even without utilizing binocular disparity, is an inspiring field of investigation throughout the life sciences.

This survey introduces recent findings about the nature of representations of three-dimensional objects. The processes of acquiring these are treated, as well as object recognition and pose estimation. [2]

[1] the images on the retinae

[2] Three-dimensional object perception is a big subject. I have concentrated this review on the *visual* processing of 3D information. But, without doubt, investigations should not be restricted to vision. Many studies combine diverse stimulus variables, particularly across modalities. For example, sound can alter the perceived 3D path of ambigous 2D motion [1]. Even more important for the acquisition of object representations seems to

The history of exploring the mechanisms underlying the establishment of representations of 3D structure from 2D features started with analyzing low-level visual processes such as edge detection. In section "Low-Level Visual Processes", which is based on [2], findings about these low-level processes from a biological as well as a technical point of view are reviewed. For most of these theories are limited either by their biological plausibility or by their technical applicability (often caused by narrowly defined constraints) the majority of activities now has shifted the focus to high-level principles of cortical processing. Section "High-Level Theories of 3D Object Perception" gives an overview of these theories, most of which were derived from computer simulations, artificial systems and technical applications. The results from psychological and physiological research, which either support or contradict high-level models, are reviewed in section "Behavioural and Physiological Evidence for High-Level Theories".

LOW-LEVEL VISUAL PROCESSES

Different types of optical information are used in an early stage of visual processing to infer the 3D shape of an object from its 2D features, like the shading of the object, the texture of its surfaces, its contours or binocular disparity. Almost all of these sources have in common that the information they provide may fundamentally differ depending on, e.g., the position and type of the source of illumination or the relative position of observer and object. That means, that the patterns of shading or texture are not invariant over changes in the viewing position and direction of illumination. Many mathematical models have been proposed, which provide very precise predictions about the spatial layout of the environment, but they suffer from implausible, implicit assumptions about biological principles. For example, there is evidence from psychological experiments that humans perform poorly on estimating the precise distance between two points in the environment, whereas they are able to assess very well which point is closer to them [3]. More generally, the human perception of Euclidean metric structures is very limited, even though the performance is good on recognizing and interacting with objects in the environment.

Nevertheless some important lessons about the processing of 3D structure have been learned from these low-level models, which are introduced now in more detail.

Structure from Shading

Artists have long known that it is possible to give an impression of the 3D shape of an object by the gradation of its surface luminance, i.e., by its shading. But it still is not quite clear how our visual system processes this information. The first attempts to calculate the orientation of a surface from its intensity were based on restrictions on the environment, like the smoothness of the surface, uniform and known reflectance properties, and uniform and known illumination fields [4]. Subsequent analyses relaxed the restrictions and gained surface orientation from the gradient of the image intensity. The magnitude and direction of the gradient can be used to estimate slant and tilt, respectively. Almost all shape-from-shading methods require the prior knowledge about the surface reflectance properties or the illumination pattern and thus are not biologically plausible. Conditions that cause major difficulties include indirect illumination, shadows, trans-

be the interaction between vision and motor control, like when grasping an object.

parency, and specular highlights, with which humans can easily cope [5]. But some plausible neuronal mechanisms have also been proposed, e.g., for computing relative surface depth [6] and for obtaining isointensity contours at different spatial scales [7].

Structure from Texture

The size of the elements of a textured surface and their spacings both decrease with increasing distance from the observer. Moreover, the elements are compressed if the angle between observer and surface decreases. Both effects can contribute to recover the 3D structure of surfaces from texture. This was first discussed in [8]. Most approaches for recovering shape from texture assume homogenously distributed texture elements, but humans seem to have few difficulties in perceiving shape from texture with nonhomogenously distributed elements [9].

Structure from Contour

The occluding contour of an object is constituted by surface points with a surface normal perpendicular to the viewing direction. It contains the silhouette and internal contours. If we move away from an occluding contour on a surface, depth decreases monotonically until a local depth minimum is reached. Thus contours provide a source of information about the 3D structure of an object [10], and there is evidence from psychophysical experiments that humans make use of the regions near occluding contours to perceive 3D shape [3].

Structure from Binocular Disparity

Binocular disparity denotes the fact that two slightly different images of the viewed scene are projected to both retinae. Since it has been thought of as a very powerful source to recover 3D shape information many mathematical models have been proposed, which promised to provide metric surface properties. But, while at close viewing distances (less than 2 m) the disparities are supposed to be proportional to the viewing distance, the same does not hold true for larger distances. Incorrect scaling between depth and disparity has been discovered [11]. Within small objects at a distance above 3 m our stereopsis is unable to resolve depth differences.

Structure from Multiple Sources

The relevance of different aspects of the 3D structure of the environment (like shading, texture, and so on) may vary with the task to be performed. Most researchers agree that multiple sources of information are combined, but it is not yet clear how different cues are combined. Cue fusion seems to be a highly flexible process depending on the task demands, specific stimulus configurations, and the amount of noise for each cue.

One can distinguish between *weak fusion* models, where each source of information is first interpreted independently and combined only afterwards, e.g., by linear combination of multiple sources [12], and *strong fusion* models, which modulate non-linear interactions between different cues during the computation of the 3D structure, as in Bayesian models [13], and which are more likely to be realized in humans.

Structure from Motion

In contrast to sources of information like shape and texture, which are contained in one single, static image, motion provides information about 3D structure by a sequence of images, derived either from a movement of the observer or from a movement of the object. Thus motion constitutes a different quality of information compared to the previously described sources. Much effort has been made to de-

termine how many views or points or derivatives of flow fields are necessary to uniquely reconstruct an objetc's 3D structure, but there is (again) evidence that human object recognition does not utilize such precise, mathematical analysis. For example, it was shown [14] that three views of four non-coplanar points are the minimum needed to reconstruct the 3D structure of an object under orthographic projection, whereas humans are able to recover the affine 3D structure just from two views of a motion sequence [15].

Motion is a complex feature and it is of different quality than the other sources of information described in this section. It is not at all a "low-level" feature, although often mentioned in this context. For this reason detailed information about it is given in "Temporal Context" of section "High-Level Theories of 3D Object Perception".

HIGH-LEVEL THEORIES OF 3D OBJECT PERCEPTION

The above mentioned weaknesses underline the necessity to assume higher-level functions underlying the perception of objects in our three-dimensional world. In this section I summarize advanced theories and models, which contribute to explain the perception of 3D objects. Their plausibility and limitations with respect to biology are discussed in the section "Behavioural and Physiological Evidence for High-Level Theories".

Existing theories about the nature of 3D object representations can be classified according to many different aspects of vision. If the coordinate system the representation is based on is of interest, a possible classification is achieved by the distinction between *object-centered* and *viewer-centered* representations. A division into *volume-based* and *view-based* representations reflects the different nature of features which constitute a representation. The context in which different models and theories are discussed always depends on the aspects of vision to be explained. Table 1 shows a classification scheme valid for this review. [3]

Volume-Based versus View-Based Representations

Volume-Based Representations: Most of the objects in the visual world can be divided into one or more volumetric parts, thus it should be possible to represent them by these constituent parts and their spatial relations. Representations based on this principle are called *volume-based* or *model-based*. Marr and Nishihara [16] were the first to propose recognition by reconstruction. According to their model the visual input is totally reconstructed and matched to a three-dimensional representation in memory.

Other approaches can be subdivided into *parametric* representations, which need a large number of quantitatively defined primitives, and *vocabulary-based* representations, which get by on only a few, but qualitatively defined parts, which constitute the object representation. An example for vocabulary-based representations was provided by Biederman [17]. In his model an object is represented by *geons* (which are volumetric primitives like cylinders or rectangular solids) and the spatially invariant relations among them. Each object is composed of several of these parts, thus thousands of objects can be represented by combinations of only a few (about 10) elementary, volumetric primitives. Our visual system then is sup-

[3] The terms *viewer-centered* and *object-centered* representations are mostly used synonymously to the expressions *2D*- and *3D*-representations, respectively. I will follow this terminology, although some authors refer to *volume-based*, object-centered representations only as *3D*.

	object-centered = 3D	viewer-centered = 2D
volume-based	Marr and Nishihara [16] Biederman [17] Marr [18]	–
view-based	Lowe [19] Ullman [20]	Koenderink [21] Poggio and Edelman [22] Bülthoff and Edelman [23] Ullman and Basri [24]

Table1: Classification Scheme For Object Representations

posed to recognize an object by decomposing it and comparing its parts to stored templates. If a sufficient number of geons is identified the object is recognized. According to this model, viewpoint invariance can thus be derived from a single view of the object.

An example for parametric representations are the *generalized cylinders* introduced by Marr [18]. Generalized cylinders are constructed mathematically by sweeping a two-dimensional cross section along an axis. A large number of primitives can thus be generated by slight variations in, e.g., the size of the cross section.

View-Based Representations: Many computational models have been proposed, e.g., by von der Malsburg et. al. [25], which show that 2D object views can be combined into the equivalent of a 3D object representation.

The simplest view-based description of an object is a densly sampled collection of views of it, which are treated independently. The addition of new views would not increase the complexity of the description, but only increase the size of the search space. Even for a such a simple view-based representation the visual system would need the ability to transform views to different viewing angles inside narrow ranges, otherwise an infinite number of views would have to be stored.

A solution can be a representation of an object in the form of a (smaller) collection of relevant views only and the spatial relations among them. Recognition of intermediate views with such a representation could be achieved, e.g., by interpolation or extrapolation of stored views (see subsection “How to Recognize Unfamiliar Views”). The relations among stored views preserve the spatial information, which is lost in simpler view-based approaches. A description of 3D objects by *aspect graphs*, proposed by Koenderink and van Doorn [26, 21], is one example for such an advanced view-based representation. The vertices of an aspect graph are constituted by views which can be imagined as special points on a transparent viewing sphere with the object in its center. These stored views represent distinct *aspects*, and the relations between them are expressed by *events*. Events occur, whenever changes in the viewpoint lead to qualitative changes in the appearance of the object. Aspect graphs have been applied, e.g., by Seibert and Waxman. In [27] they describe the learning of representations for 3D objects from arbitrary sequences of the rotating object. Based on an edge and corner detection, they cluster the views into different aspects.

Each object is represented then by a "transition matrix", which contains a kind of probability for the transition from one aspect to another. Utilizing these transition matrices their system is able to recognize objects, but pose estimation or the generation of intermediate (non-experienced) views is not possible.

Another view-based approach, which even allows pose estimation, is proposed by Murase and Nayar [28]. They represent objects by a manifold in eigenspace. An input image of an object to be recognized is projected to the eigenspace of the learned objects. The object is recognized based on the manifold it lies on. The exact position of the projection on the manifold determines the object's pose.

Object-Centered versus Viewer-Centered Representations

Object-Centered Representations: *Object-centered* representations are characterized by a description of the parts of the object relative to an object-centered coordinate system. For example, Marr specified the object's parts relatively to the object's main axis [18]. A single description of the object is valid for all possible viewpoints, i.e., the description of the object is independent of the position of the observer. Biedermans geon approach, reported previously, belongs to the object-centered representations, too.

But an object-centered representation is not necessarily volume-based. The relations between object parts can also be expressed in two-dimensional terms like relations between lines and corners, as proposed by Lowe [19] with his *viewpoint consistency constraint*. He projects each 3D model stored in memory to a hypothesized viewpoint and matches the resulting projected locations of the 2D features to the input image. A very similar approach by Ullman [20] is known as *recognition by alignment*.

Such single-description, object-centered representations seem to be very economical, but they require the ability of the visual system to boundlessly transform across views.

Viewer-Centered Representations: If the parts of an object are described relative to a coordinate system based on the observer, the representation is called *viewer-centered*. In this case the description of the object depends on the viewing angle. The previously mentioned aspect graph approach by Koenderink [21] is an example for this category. Other approaches, e.g., by Poggio and Edelman [22], Bülthoff and Edelman [23], and Ullman and Basri [24] are described in the subsection "How to Recognize Unfamiliar Views".

Object-centered representations seem to be appropriate for recognition tasks, because the recognition of an object should not depend on the the viewing angle of the observer. The visual guidance of interactions with objects, however, requires the observer as frame of reference. That means, for this task a viewer-centered, and thus view-specific representation will be more suitable.

Characteristic Views

Across the different models for 3D object perception, the notion of *characteristic views* is a prominent topic. Derived from experimental research, characteristic views can be defined as views which are easier to recognize than other views of the same object. From a technical point of view, a characteristic view is a view which is useful for matching an object. [4]

Still open questions regarding characteristic views are the number of views necessary for

[4] A term which should be mentioned in this context is *canonical view*, which is defined in [29] as view "humans find easiest to recognize and regard as most typical". Thus, the expression *canonical view* is mostly used synonymously to *characteristic view*.

different visual tasks and the manner in which they are defined. Many explainations have been suggested, why some views are easy to recognize and others are not. Different features have been considered for this classification. One aspect for defining characteristic views might be the constellation of visible corners and edges. For example, Gray [30] clustered the viewing sphere of line drawings of geometrically facetted objects into regions of similar views based on corners and edges. He obtained nine clusters, where each view of a cluster shared at least 17 features with every other view of the same cluster, thus he obtained nine characteristic views, which represent the whole object.

In another approach Koenderink and van Doorn [31] applied a mathematical extension of singularity theory to smooth objects. In this context one type of singularities are contours that separate visible from occluded surfaces (see section "Low-Level Visual Processes"). The shapes of these contours are classified in general. A classification of contour shapes of a special object results in clusters of vantage points with unchanged singularities. At unstable vantage points the set of singularities changes and the observer experiences an event. Because this approach leads to a huge number of characteristic views, it is questionable if it can account for human object perception.

Although characteristic views are more strongly connected with viewer-centered concepts object-centered models also have to explain the phenomenon of characteristic views. Approaches utilizing the object's principal axis can define characteristic views by the projected length of this axis [16]. Views with a foreshortened main axis are supposed to be more difficult to recognize.

Another idea is the concept of *salient* or *non-accidental* features used by, e.g., Lowe [19] or Biederman [17]. In this scheme some parts of an object are of particular salience and their visibility facilitates recognition. Accordingly, characteristic views are not derived from a general procedure, rather they highly depend on the specific object.

How to Recognize Unfamiliar Views

If we start from the assumption that characteristic views play a dominant role in object recognition then the question arises how non-characteristic views, i.e., views which have not been experienced before and are not stored in the representation, can be recognized. One of the most supported idea is the *interpolation* of unfamiliar views. Novel views can be generalized from stored views by a view approximation as described, e.g., by Poggio and Edelman [22]. According to this theory humans and other primates can achieve viewpoint-invariant recognition of objects by a system that interpolates between a small number of stored sample views as shown by Bülthoff and Edelman [23]. Unfamiliar views lying between stored views on the same rotation axis should be recognized more easily, than those which are somewhere else on the viewing sphere. In addition, recognition should deteriorate with an increasing distance of the novel view from a stored view.

Another theory about the recognition of unfamiliar views, which also belongs to the view-based approaches, is the recognition by *linear combinations* of views. Ullman and Basri [24] showed mathematically that, under orthographic projection, the 2D-coordinates of an object point for a special view can be expressed as a linear combination of the object point coordinates in a limited set of other viewpoints, provided the correspondence between points in all views is known. The number of required views depends on the complexity of the object and the allowed 3D transforma-

tions. At most six images would need to be stored to allow the reconstruction of an object from any viewpoint. In contrast to the interpolation model, this model predicts an equally high recognition rate for unfamiliar views lying in the space spanned by the stored views, independent of the distance beween novel and stored views. Among others, Beymer and Poggio used a linear combination approach to apply prior knowledge of an object class (faces) to generate virtual views for face recognition [32].

The *alignment* of 3D models, already mentioned in the subsection "Object-Centered versus Viewer-Centered Representations" is a third theory for generalizing from familiar to unfamiliar views. It is strongly connected to the notion of *mental rotation* described by Shepard and Cooper [33], which belongs to the class of object-centered approaches. During recognition by alignment each stored model undergoes an aligning transformation after which it is compared to the input image. A visual system, which utilizes aligned 3D models, can recognize perfectly, as long as all features used for the transformation are visible.

Temporal Context

Up to now I have focused this review on representations of three-dimensional objects and the recognition process utilizing *static* views, only. But many studies have proved the importance of the temporal context during observation of a *sequence* of a moving object, when either the object rotates or the observer moves (see "Structure from Motion" in the section "Low-Level Visual Processes"). Additional information is provided to the visual system by a sequence. The establishment of a representation is facilitated by tracking object features, thus providing correspondences in neighbouring views. Corresponding points are given by the temporal context and need not be derived from static sources of information, which is often misleading. It was shown, e.g., in [34] that better correspondences are derived from the continuity of successive views than from (the same) disconnected, static views.

Different Kinds of Representations For Different Tasks

The utility of a representation has to be assessed with respect to the task, e.g., recognition, pose estimation, or interaction. Different visual tasks may require different types of representations. On the one hand, one can imagine that volume-based representations are especially useful for visual guidance of interactions with objects, like grasping them. For just recognizing an object it seems to be not necessary to utilize the spatial relations of its constituent parts. For recognition a view-based representation may be sufficient. On the other hand, if object-centered versus viewer-centered representations are compared, then it would be advantageous for interactions with objects, if the frame of reference for the object is already the same as for the interacting individuum, i.e., if the object is represented by a viewer-centered description.

Another pair of terms, not mentioned so far, is given by representations which facilitate *viewpoint invariant* performance during recognizing an object and those which lead to a *viewpoint specific* (or *viewpoint dependent*) behavior. Viewpoint invariance of an object representation can be discussed in two terms. First, in terms of error rates, i.e., how often would a visual system misclassify a perceived object depending on the view it is exposed to. Second, viewpoint invariance can be refered to the time needed to recognize an object. [5] In gen-

[5] Object-centered, volume-based representations are refered to as viewpoint invariant in terms of error rates

eral, recognition requires viewpoint-invariant performance. Imagine an animal which cannot distinguish between a poisonous and a nutricious plant independently of the view point. Viewpoint-specific behaviour can be advantageous if pose-estimation or interaction are demanded.

Another interesting point about learning object representations is described in [35]. If the orientation of an object affects the recognition time, then practice greatly diminishes this effect. However, this effect of practice does not transfer to new objects. Thus, what is learned is specific to a set of stimuli and is not a general-purpose procedure. This does not support a concept of a "frame-independent" representation, as proposed, e.g., in [36].

BEHAVIOURAL AND PHYSIOLOGICAL EVIDENCE FOR HIGH-LEVEL THEORIES

In this section I describe results of studies with humans and monkeys which either support or contradict the theories of 3D object recognition I introduced in the last section. The first subsections mainly summarize psychological (i.e., behavioural) studies, whereas physiological results, mainly derived from single neuron recordings from monkeys, are reported in the last subsection "Hierarchies in Higher-Order Visual Processes in the Brain".

Evidence for Volume-Based Representations

Two reasons, derived from psychological experiments quoted in [37], support the plausibility of the geon theory.

a) First, viewpoint invariance for familiar objects was confirmed by assessing naming latencies, as long as the same geons were visible in the training and in the test images of line drawings of objects. This was independent of the viewpoint-specific appearence of single geons.

b) Second, subjects were able to distinguish immediately between unfamiliar objects, if their constituting geons were distinguishable. These results could not have been achieved, if pure view-based coding is assumed.

c) A third argument contradicts a pure view-based representation. As described, e.g., in [38], humans can memorize the shape of a pattern even if they are not able to recall its left-right orientation. In a view-based representation the left-right orientation would be stored intrinsically with the shape.

Evidence for View-Based Representations

There are some arguments that support the model of an advanced view-based description of 3D objects by our visual system. (Simple view-based representations in the form of a collection of independent views are unlikely to be realized in the human brain. Otherwise it would be difficult to explain humans' ability to recognize novel views of familiar objects [37].)

a) If a set of object views is presented to humans and their reaction time and their error rates during recognizing unfamiliar views of the same object are measured, both, reaction time and error rates increase with increasing angular distance between the learned (i.e., stored) and the unfamiliar view [39]. The decrease in recognition speed for unfamiliar views was already reported in [40] and [41]. For monkeys the ability to generalize from training views was also found to be worse as the rotation angle increases [42, 43]. These results strongly support a view-based model.

b) This angle effect on the performance de-

and response time. Viewer-centered, view-based representations are defined as view specific, whereas object-centered, view-based representations predict a viewpoint invariant performance in terms of error rates, but not in terms of response times.

clines, if intermediate views are experienced and stored [44]. This would not be expected either, if viewpoint invariance from a single view is assumed, as done by geon theory, for example.

c) Similar results have been derived from recordings of single neurons in the inferior temporal cortex (IT) of monkeys by Logothetis et. al. [45]. They found populations of IT neurons, that responded selectively to only some views of a previously unfamiliar object. The response declined gradually as the object was rotated away from this preferred view.

d) The object-centered representations proposed by Biederman [17] predict well the view-independent recognition of *familiar* objects. However, it is difficult to distinguish whether the recognition can be put down to a three-dimensional description or to the fact that the system has already been exposed to a sufficient number of two-dimensional views. In addition, there are some unanswered question in this approach. How is view invariance achieved for the volumetric parts themselves, and how is invariance achieved for objects that cannot be decomposed further?

Evidence for Characteristic Views

a) The fact that some views of an object are better suitable for recognition than others was confimed early by observations of patients with right parietal cortex lesions [46]. They were poorer than control subjects in recognizing objects from "unusual" views, whereas "usual" views were not affected.

b) In experiments carried out by Edelman and Bülthoff [39, 47] naming was fastest if a stimulus was in a characteristic view. And these views were established even if in the training phase each view of an unfamiliar object appeared with equal frequency.

c) Regarding the nature of characteristic views there are hints that a non-foreshortened principal axis of the object is not as important for recognition as the visibility of salient features. In a study by Palmer et. al. [29] subjects had to choose characteristic views for several objects, and the views they selected often had a viewing angle of about 45° to the principal axis of the object. That means, that the principal axis was foreshortened considerably.

d) Koenderink and van Doorn's singularity theory [31] is also unlikely to account for the nature of characteristic views. It predicts views to be characteristic which show only slight changes of their contours under slight rotations. But, as Perrett and Harries [48] found out, humans prefer views with the principal axis of the object either parallel or perpendicular to the line of sight - and these views provide strong variations of their contours under slight rotations. Views with the principal axis of the object either parallel or perpendicular to the line of sight are often "plan views", equivalent to those drawn by an architect to represent an object. The benefit of such views is, according to [49], the absence of perspective distortions in the third dimension.

e) An ever recurring question concerns the distribution of characteristic views, i.e., their number and distances. [6] Experiments with monkeys were made by Logothetis et. al. [43] which showed that familiarization with a "limited number" of views of a novel object can provide viewpoint-independent recognition. Three views of a wire-like object, 120° apart, often were sufficient for recognizing any view resulting from rotations around the same axis. For the entire viewing sphere about 10 views were sufficient to achieve view-

[6] The studies reported in this section do not provide arguments for characteristic views only, but also strongly favour viewer-centered approaches.

independent performance. But the same study claimes that the number of required viewpoints may depend on the object class. It may reach a minimum for a novel object of a familiar class, e.g., for a new individual face one view only may be sufficient. The unability of monkeys to recognize objects rotated by more than approximately 40° from a *single* familiar view is also reported.

f) A similar result for human object recognition was published earlier by Rock and DiVita [50]. Subjects became very poor in recognizing wire-like objects for view distancies larger than approximately 30°. They could not even imagine, how the objects would look when rotated further.

Evidence for View Interpolation

a) Bülthoff and Edelman [23] made psychophysical experiments to compare the three theories about recognition of unfamiliar views, I described in section "High-Level Theories of 3D Object Perception": nonlinear view interpolation, linear combination of views, and alignment of 3D models.

Subjects were shown two training views of a computer generated 3D wire-like object, which were 75° apart. [7] In the test phase a novel view was presented, which was either on the same rotation axis *between* the training views, or on the same rotation axis *beyond* them, or on an axis *orthogonal* to the training axis. The error rates during recognition mostly fit the predictions of the interpolation model, i.e., the error rates were lowest for the *between* condition, medium for the *beyond* condition and highest for the *orthogonal* condition. This contradicts the linear combination model, which predicts the same good performance for the *between* and *beyond* condition and poor performance for the *orthogonal* condition. The experiment also contradicts alignment models, which predict uniformly good performance for all three test conditions.

b) Monkeys were trained with two views of a computer-rendered wire or spheroidal novel object, which were far apart, e.g., 0° and 120°. They could recognize all test views inside this interval, whereas the extrapolation along either the same or an orthogonal axis was limited as Logothetis et. al. found out [42]. This also supports the interpolation model.

c) The linear combination of views proposed by Ullman and Basri [24] is of theoretical interest, but its validity for human object perception seems to be limited. First, it is applicable to line drawings only, second, the viewing angles must be known for the calculations.

Evidence for the Importance of Temporal Context

Context in general (not necessarily temporal) can improve the recognition of novel views. This was shown by Christou et. al. [51]. However, temporal context seems to be of special importance for the establishment of object representations. This is supported by a series of psychophysical experiments.

a) Niemann et. al. [52] report on experiments with parts of statues of human figures on a turntable. The eye movements of subjects watching the rotating objects were recorded. They found that the eye movements were often directed to the same details from different vantage points. This also supports the relevance of tracking of local features.

b) Another argument is furnished by Harman and Humphrey [53]. They claim that different object representations are generated, depend-

[7] In another experiment subjects were also shown two training views 75° apart, oscillating 15° around a fixed axis, so they experienced small sequences. The results of these experiments were similar to the ones I describe now.

ing on the presentation of either regular or random sequences of views of the object. When a sequence of rotations is encoded, the associated temporal context may lead to the construction of a linked, higher-order system of representations for a given object. Without temporal context, a single representation of each object rotation may be constructed.

c) If an object in a 3D scene is rotated, its perceived depth increases, as described by Sauer et. al. [54]. Subjects had to jugde the shape of objects, and they perceived more acute angles with an increased rotation of the object.

d) If 3D objects are represented by a collection of stored views, then this collection is structured in the sense that views belonging together because of their successive appearance are more closely associated with each other in the representation. This is claimed by Edelman and Weinshall [55] as well as by Perrett et. al. [56].

e) That this result is not simply due to shared structural information, is suggested by Kellman's [57] research with infants. He found out that they have the ability to perceive the three-dimensional form of an object if only information about continuous optical transformations given by motion is available. They are not able to apprehend the overall form of an object from static views, even if they are multiple or sequential. This holds true even for eight months old infants. Adults, however, are able to perceive 3D form from static views of objects. The recognition from static views seems to lean on extrapolations to the whole form based on simplicity or symmetry considerations, which may be products of learning, whereas the other mechanism is innate or develops early.
(Sixteen-week-old human infants are able to distinguish optical displacements given by their own motion from displacements given by moving objects, and they use only the latter to perceive the unity of partly occluded objects [58].)

f) Spelke claims that one of the Gestalt principle developed earliest in infants is motion [59]. Their perception does not seem to attend to nonaccidental geometric relations in visual arrays. Rather they divide their visual input into units that move as connected wholes and separately from another. This also supports the great significance of motion for object perception.

g) Also some physiological reasons support the importance of temporal context for three-dimensional object perception. Miyashita [60] trained monkeys to match complex fractal patterns, which were presented successively in a fixed series of 100 items. After training cells in the anterior temporal cortex were found to show selectivity for a small number of patterns which had been presented successively. This gives evidence for learning based on temporal associations rather than on pattern overlap.

Evidence for the Coexistence of Different Systems

There is evidence for the coexistence of separate representation systems in the human brain for identification (recognition) of objects on the one hand, and for visual guidance of interactions with objects, on the other hand.

The difference between patients which suffer from agnosia (unability to recognize an object's identity) and those with apraxia (failing to interact with objects) can make this clear. In [61] an agnostic patient is described who cannot recognize objects, but nonetheless she is able to interact with them, guiding her hand in the appropriate shape for grasping. There are also patients who suffer from apraxia without agnosia.

Hierarchies in Higher-Order Visual Processes in the Brain

Two major pathways project from the primary visual cortex to higher parts of the primate brain. The dorsal pathway, which encodes the spatial layout of the environment, projects to the parietal cortex, which is known to control motor action. The ventral projections mediate object form and project to the temporal cortex. Here especially cells in the superior temporal sulcus (STS) are concerned with object recognition. A fusion of this "What"- and "Where"-information is realized by direct axonal projections between the parietal and temporal cortex by indirect connections via the hippocampus, where place cells can be found, which form cognitive maps.

Cells in STS of macaques have been found which respond selectively to faces, hands, and other classes of biologically significant objects. The majority of these cells exhibits a viewer-centered response pattern, i.e., some of them respond selectively to face or profile views of heads, as described by Perrett et. al. [62], although at the same time they generalize across image position, size, orientation in the image plane, color, and lighting conditions [63]. There are more cells which are optimally tuned to characteristic views (like full face or profile) than to other views [64] [8]. The tuning covers views between 45° and 70° until the response is reduced to half of its maximum. Maybe STS combines inputs from earlier viewer-centered describtions in the inferior temporal cortex (IT), which are size and orientation [9] specific, to size and orientation tolerant cells. Tanaka et. al. [65] suggest, that the coding of faces, hands and arbitrary objects share an early stage of analysis in IT.

There are also cells in STS which exhibit object-centered coding, i.e., they respond equally to all views of an object. This was shown, e.g., by Perrett et. al. [64] for the coding of heads and by Booth and Rolls [66] for the coding of small plastic objects [66] in macaque brains. In [64] they found cells selective to all views of one individual's head, but unresponsive to all views of a different individual. On the other hand, in [66] neurons are desribed which were responsive to all views of one, as well as, more objects.

These results suggest that 3D object recognition in the primate brain operates in a hierarchical fashion with increasing levels of abstraction. Starting with size-, orientation- and viewpoint-specific representations, to size- and orientation-invariant, but viewpoint-specific codings, up to viewpoint-invariant representations.

SUMMARY AND CONCLUSIONS

Most neuroscientists agree on the mutual stimulation between brain research and neural computation. Biological realism is crucial for a powerful artificial object recognition system, if it claims to be comprehensive. Thus the review of computational models has to consider the realization of *principles* of cortical processing, rather than the exact mapping of biological or computational details.

Taking this into account, I come to some main conclusions regarding the nature of three-dimensional object perception:

[8] Interestingly, the same views seem to be important physiologically and psychologically. Also the relative importance of views is comparable. Face and profile views appear more important than half-profile views, and all of these front views are more important than rear views of a head, both in behavioural and physiological studies, as reported in [49].

[9] orientation in the image plane

- Most of the strongest arguments favour a view-based approach to the perception of 3D form.
- Objects seem to possess characteristic views, which facilitate recognition, compared to other views.
- It is likely that novel views can be recognized by a kind of interpolation between previously experienced views.
- For learning a new object representation the experience of temporal context during object (or observer) motion is of special importance.
- There is strong evidence that cortical processing of 3D object form operates in a hierarchical way from viewer-centered to object-centered representations.

In my view, neither the extreme of a pure viewer-centered nor the extreme of a pure object-centered way of representing 3D form is realistic. Rather than this, different principles of processing may be realized, the use of which depends on the task to perform. Typically viewer-centered information may be stored, which is utilized in everyday situations and processed in basic computations. But in some circumstances object-centered information may also be available. Kosslyn [35] suggests that it may correspond to "routines", that can follow to locate distinguishing parts of an object if it is in any orientation. Such routines would require effort to be acquired, stored, and executed. This is consistent with the finding that object-centered information could be encoded only with additional effort, when subjects knew it would be useful for a later task. In addition, Perrett et. al. [67] report on longer latencies for object-centered responses in IT than for view-specific responses.

Thus, the hierarchical cortical processing of object form could not only reflect an evolution of the frame of reference from retinotopic to egocentric to allocentric [68], but could also represent a scheme, which uses more elaborate computations with increasing degree of difficulty of the task to perform.

ACKNOWLEDGEMENT

I thank ONR, grant No. N 00014-98-1-0242, for support of this work. In addition, I want to express my thanks to Prof. Christoph von der Malsburg who made this work possible, and to Dr. Rolf Würtz, Achim Schäfer and Pervez Mirza for proof-reading it.

REFERENCES

[1] L. Boucher, R. Sekuler, A. Talwalkar, and A. B. Sekuler. Motion Perception is Influenced by Accompanying Sound: 2- & 3-Dimensional Motion. In *Investigative Ophthalmology & Visual Science, ARVO Annual Meeting*, volume 39(4), page 1094, Fort Lauderdale, Florida, USA, May 10–15, 1998.

[2] J. S. Tittle and J. T. Todd. Perception of Three-Dimensional Structure. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 715–718. MIT Press, 1995.

[3] J. T. Todd and F. D. Reichel. Ordinal Structure in the Visual Perception and Cognition of Smoothly Curved Surfaces. *Psychol. Rev.*, 96:643–657, 1989.

[4] B. K. P. Horn and M. J. Brooks. *Shape from Shading*. MIT Press, Cambridge, MA, 1989.

[5] E. Mingolla and J. T. Todd. Perception of Solid Shape from Shading. *Biol. Cybern.*, 53:137–151, 1986.

[6] A. P. Pentland. A Possible Neural Mechanism for Computing Shape from Shading. *Neural Computation*, 1:208–217, 1989.

[7] S. Grossberg and E. Mingolla. Neural Dynamics of Surface Perception: Boundary-Webs, Illuminants, and Shape from Shading. *Commun. Vis. Graph. Image Proc.*, 37:116–165, 1987.

[8] J. J. Gibson. *The Perception of the Visual World*. Houghton Mifflin, Boston, 1950.

[9] J. T. Todd and R. A. Akerstrom. Perception of Three-Dimensional Form from Patterns of Optical Texture. *Percept. & Psychophys.*, 13:242–255, 1987.

[10] J. J. Koenderink and A. J. van Doorn. The Shape of Smooth Objects and the Way Contours End. *Perception*, 11:129–137, 1982.

[11] E. B. Johnston. Systematic Distortions of Shape from Stereopsis. *Vis. Res.*, 31:1351–1360, 1991.

[12] L. Maloney and M. Landy. A Statistical Framework for Robust Fusion of Depth Information. In W. A. Perlman, editor, *Visual Communication and Image Processing IV, Proc. Soc. Photo-Opt. Instrum. Eng.*, volume 1199, pages 1154–1163, 1989.

[13] H. H. Bülthoff and A. L. Yuille. Bayesian Models for Seeing Shapes and Depth. *Comments Theor. Biol.*, 2:283–314, 1991.

[14] S. Ullman. *The Interpretation of Visual Motion*. MIT Press, Cambridge, MA, 1979.

[15] J. T. Todd and P. Bressan. The Perception of Three-Dimensional Affine Structure from Minimal Apparent Motion Sequences. *Percept. & Psychophys.*, 48:419–430, 1990.

[16] D. Marr and H. K. Nishihara. Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes. In *Proceedings of the Royal Society of London, B(200)*, pages 269–294, 1978.

[17] I. Biederman. Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94:115–147, 1987.

[18] D. Marr. *Vision*. Freeman, San Francisco, MA, 1982.

[19] D. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Boston, MA, 1985.

[20] S. Ullman. Aligning Pictorial Descriptions: An Approach to Object Recognition. *Cognition*, 32:193–254, 1989.

[21] J. J. Koenderink. *Solid Shape*. MIT Press, Cambridge, MA, 1990.

[22] T. Poggio and S. Edelman. A Network that Learns to Recognize Three-Dimensional Objects. *Nature*, 343:263–266, 1990.

[23] H. H. Bülthoff and S. Edelman. Psychophysical Support for a Two-Dimensional View Interpolation Theory of Object Recognition. In *Proceedings of the National Academy of Science of the United States of America*, volume 89, pages 60–64, 1992.

[24] S. Ullman and R. Basri. Recognition by Linear Combinations of Models. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 13(10):992–1006, 1991.

[25] C. von der Malsburg, K. Reiser, G. Peters, J. Wieghardt, and K. Okada. 3D Object Representation by 2D Views. In *Proceedings of the 6th ATR Symposium on Face and Object Recognition*, pages 11–12, Kyoto, Japan, July 1999.

[26] J. J. Koenderink and A. J. van Doorn. The Internal Representation of Solid Shape with Respect to Vision. *Biological Cybernetics*, 32:211–216, 1979.

[27] M. Seibert and A. M. Waxman. Adaptive 3-D Object Recognition from Multiple Views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):107–124, 1992.

[28] H. Murase and S. Nayar. Visual Learning and Recognition of 3-D Objects from Appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[29] S. E. Palmer, E. Rosch, and P. Chase. Canonical Perspective and the Perception of Objects. In I. Long and A. Baddeley, editors, *Attention and Performance IX*, pages 135–151. Erlbaum, Hillsdale, N.J., 1981.

[30] M. Gray. Recognition Planning from Solid Models. In *Proceedings of the Alvey Computer Vision and Image Interpretation Meeting, Bristol*, pages 41–43, Sheffield, England, September 1986. Sheffield University Press.

[31] J. J. Koenderink and A. J. van Doorn. The Singularities of the Visual Mapping. *Biological Cybernetics*, 24:51–59, 1976.

[32] D. Beymer and T. Poggio. Face Recognition from One Example View. Technical Report CBCL Paper 121/AI Memo 1536, Massachusetts Institute of Technology, Cambridge, MA, September 1995.

[33] R. N. Shepard and L. A. Cooper. *Mental Images and their Transformations*. MIT Press, Cambridge, MA, 1982.

[34] G. Peters, B. Zitova, and C. von der Malsburg. Two Methods for Comparing Different Views of the Same Object. In T. Pridmore and D. Elliman, editors, *Proceedings of the 10th British Machine Vision Conference (BMVC'99)*, pages 493–502, Nottingham, UK, September 13-16, 1999.

[35] S. M. Kosslyn. *Image and Brain*. MIT Press, Cambridge, MA, 1996.

[36] M. C. Corballis. Recognition of Disoriented Shapes. *Psychological Review*, 95:115–123, 1988.

[37] 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(6):1162–1182, 1993.

[38] R. E. Frederickson and J. C. Bartlett. Cognitive Impenetrability of Memory for Orientation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 13:269–277, 1987.

[39] S. Edelman 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.

[40] M. J. Tarr and S. Pinker. When does Human Object Recognition use a Viewer-Centered Reference Frame? *Psychological Science*, 1:253–256, 1990.

[41] D. J. Bartram. The Role of Visual and Semantic Codes in Object Naming. *Cognitive Psychology*, 6:325–356, 1974.

[42] N. K. Logothetis, J. Pauls, H. H. Bülthoff, and Poggio T. Evidence for Recognition based on Interpolation among 2D Views of Objects in Monkeys. *Invest. Ophthalmol. Vis. Sci. Suppl.*, 34:1132, 1992.

[43] N. K. Logothetis, J. Pauls, H. H. Bülthoff, and Poggio T. View-Dependent Object Recognition by Monkeys. *Current Biology*, 4:401–414, 1994.

[44] M. J. Tarr. *Orientation Dependence in Three-Dimensional Object Recognition*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1993.

[45] N. K. Logothetis, J. Pauls, and Poggio T. Shape Representation in the Inferior Temporal Cortex of Monkeys. *Current Biology*, 5(5):552–563, 1995.

[46] E. K. Warrington and A. M. Taylor. The Contribution of the Right Parietal Lobe to Object Recognition. *Cortex*, 9:152–164, 1973.

[47] H. H. Bülthoff, S. Edelman, and M. J. Tarr. How are Three-Dimensional Objects Represented in the Brain? Technical Report No. 5, Max-Planck-Institut für biologische Kybernetik, Tübingen, Germany, 1994.

[48] D. I. Perrett and M. H. Harries. Characteristic Views and the Visual Inspection of Simple Faceted and Smooth Objects: "Tetraheder and Potatoes". *Perception*, 17:703–720, 1988.

[49] D. I. Perrett, M. W. Oram, J. K. Hietanen, and P. J. Benson. Issues of Representation in Object Vision. In M. J. Farah and G. Ratcliff, editors, *The Neuropsychology of High-Level Vision - Collected Tutorial Essays*, pages 33–61, Hillsdale, New Jersey, 1994. Lawrence Erlbaum Associates.

[50] I. Rock and J. DiVita. A Case of Viewer-Centered Object Perception. *Cogitive Psychology*, 19:280–293, 1987.

[51] C. G. Christou, B. S. Tjan, and H. H. Bülthoff. Old Paperclips, New Context. In *Investigative Ophthalmology & Visual Science, ARVO Annual Meeting*, volume 39(4), page 853, Fort Lauderdale, Florida, USA, May 10–15, 1998.

[52] T. Niemann, M. Lappe, and K.-P. Hoffmann. Visual Inspection of Three-Dimensional Objects by Human Observers. *Perception*, 25:1027–1042, 1996.

[53] K. L. Harman and G. K. Humphrey. Encoding "Regular" and "Random" Sequences of Views of Novel 3D Objects Rotating in Depth. In *Investigative Ophthalmology & Visual Science, ARVO Annual Meeting*, volume 39(4), page 856, Fort Lauderdale, Florida, USA, May 10–15, 1998.

[54] C. W. Sauer, M. L. Braunstein, and G. J. Andersen. Effects of Rotation on Perceived Object Shape in Motion Parallax Scenes. In *Investigative Ophthalmology & Visual Science, ARVO Annual Meeting*, volume 39(4), page 855, Fort Lauderdale, Florida, USA, May 10–15, 1998.

[55] S. Edelman and D. Weinshall. A Self-organizing Multiple-View Representation of 3D Objects. *Biological Cybernetics*, 64(12):209–219, 1991.

[56] D. I. Perrett, A. J. Mistlin, and A. J. Chitty. Visual Neurons Responsive to Faces. *Trends in Neurosciences*, 10:358–364, 1989.

[57] P. J. Kellman. Perception of Three-Dimensional Form in Infancy. *Perception and Psychophysics*, 36:353–358, 1984.

[58] P. J. Kellman, H. Gleitman, and E. S. Spelke. Object and Observer Motion in the Perception of Objects by Infants. *Journal of Experimental Psychology: Human Perception and Performance*, 13(4):586–593, 1987.

[59] E. S. Spelke. Principles of Object Perception. *Cognitive Science*, 14:29–56, 1990.

[60] Y. Miyashita. Associative Representation of Visual Long Term Memory in the Neurons of the Primate Temporal Cortex. In E. Iwai and M. Mishkin, editors, *Vision, Memory and the Temporal Lobe*, pages 75–87. Elsevier, New York, 1990.

[61] A. D. Milner, D. I. Perrett, R. Johnston, P. J. Benson, T. R. Jordan, D. W. Heeley, D. Bettucci, F. Mortara, F. Mutani, E. Terazzi, and D. L. W. Davidson. Perception and Action in "Visual Form Agnosia". *Brain*, 114:405–428, 1991.

[62] D. I. Perrett, P. A. J. Smith, D. D. Potter, A. J. Mistlin, A. S. Head, A. D. Milner, and M. A. Jeeves. Visual Cells in the Temporal Cortex Sensitive to Face View and Gaze Direction. In *Proceedings of the Royal Society of London, B(3)*, pages 293–317, 1985.

[63] J. K. Hietanen, D. I. Perrett, M. W. Oram, P. J. Benson, and W. H. Dittrich. The Effects of Lighting Conditions on the Responses of Cells Selective for Face Views in the Macaque Temporal Cortex. *Experimental Brain Research*, 89:157–171, 1992.

[64] D. I. Perrett, M. W. Oram, M. H. Harries, Bevan R., J. K. Hietanen, P. J. Benson, and S. Thomas. Viewer-Centred and Object-Centred Coding of Heads in the Macaque Temporal Cortex. *Experimental Brain Research*, 86:159–173, 1991.

[65] K. Tanaka, H.-A. Saito, Y. Fukada, and M Moriya. Coding Visual Images of Objects in the Inferotemporal Cortex of the Macaque Monkey. *Journal of Neurophysiology*, 66:170–189, 1991.

[66] M. C. A. Booth and E. T. Rolls. View-Invariant Representations of Familiar Objects by Neurons in the Inferior Temporal Visual Cortex. *Cerebral Cortex*, 8(6):510–523, 1998.

[67] D. I. Perrett, J. K. Hietanen, M. W. Oram, and P. J. Benson. Organization and Functions of Cells Responsive to Faces in the Temporal Cortex. In *Philosophical Transactions of the Royal Society of London*, volume 335, pages 23–30, 1992.

[68] A. M. Waxman, M. Seibert, and Bachelder I. A. Visual Processing of Object Form and Environment Layout. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 715–718. MIT Press, 1995.