PERGAMON

Vision Research 39 (1999) 603–612

Vision
Research

Dissociating stimulus information from internal representation—a case study in object recognition

Zili Liu [a,*], Daniel Kersten [b], David C. Knill [c]

[a] *NEC Research Institute, 4 Independence Way, Princeton, NJ 08540, USA*
[b] *Department of Psychology, University of Minnesota, MA, USA*
[c] *Department of Psychology, University of Pennsylvania, PA, USA*

Received 14 August 1997; received in revised form 17 October 1997; accepted 18 May 1998

Abstract

Human object recognition is a function of both internal memory representation(s) and stimulus input information. The role of the latter has been so far largely overlooked, and the nature of the representation is often directly equated with recognition performance. We quantify stimulus information for three classes of objects in order of decreasing object complexity: unconnected balls, balls connected with lines, and balls connected with cylinders. In an object discrimination task, subjects' performance improved with the decreasing object complexity. We show that input information also increases with decreasing object complexity. Therefore, the results could potentially be accounted for either by differences in the object representations learned for each class of objects, or by the increased information about the three-dimensional (3D) structure inherent in images of the less complex objects, or by both. We demonstrate that, when image information is taken into account, by computing efficiencies relative to a set of ideal observers, subjects were more efficient in recognizing the less complex objects. This suggests that differences in subjects' performance for different object classes is at least partly a function of the internal representations learned for the different object classes. We stress that this conclusion cannot be achieved without the quantitative analysis of stimulus input information. © 1998 Elsevier Science Ltd. All rights reserved.

Keywords: Object recognition; Representation; Ideal observer; Image information; Object complexity

1. Introduction

If a few points are drawn at the joints of an invisible human figure, the configuration of the dots as a human figure cannot be recognized (Johanssen, 1973). If we connect the points with appropriate lines, obviously the human figure can be recognized, even though the relative lengths of the body parts in three-dimension (3D) are indeterminate. Now if we replace the lines with cylinders so that shading details on a cylinder surface are visible, we will then have a much better idea about the structure of this human figure in 3D.

Such an observation may not be considered surprising since we have available, increasingly more information (or less ambiguity) about the human figure from the images. Moreover, we may prefer a stick human figure to a bunch of dots as a 'natural' human figure representation. In fact, both factors, additional image information and preferred intrinsic object representation, could account for the above observation. The question is, to what extent are the perceptual differences a function of these two factors?

This question is of critical importance because of the following, seemingly obvious, problem in high-level vision. An experimental effect in human subjects' performance is often attributed to a functional difference in the brain, although the difference in the stimuli may well have accounted for the effect. For example, Farah, Rochlin & Klein (1994) have shown that discriminating objects that are defined by simple closed curves depend on whether the objects are made of thin wires wrapped around the curve or surfaces interpolated within the curve. Their subjects generalized from familiar to novel object views better for the 'surface' (or 'potato chip') objects than for the wire objects. However, one cannot

[*] Corresponding author. Fax +1 609 9512481; e-mail: zliu@research.nj.nec.com.

PII: S0042-6989(98)00167-9

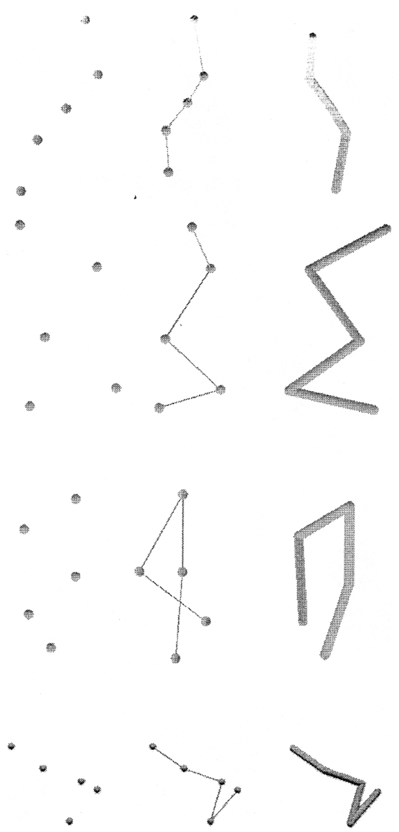

Fig. 1. Three classes of stimulus objects. Left, balls; middle, tinker toys; and right, wires. The top three rows illustrate the objects actually used in the experiment. In each row, the 3D positions of the balls are identical, although the three views in each row are from different angles. The bottom row shows an example of three objects from the same viewpoint.

conclude from these data that wire objects are represented in memory as strongly viewpoint-dependent structures, whereas 'potato chips' are represented as less viewpoint-dependent surfaces, because the nature of the information provided by the images of the wires and 'potato chips' is markedly different[1]. Consequently, without defining and quantifying stimulus information, it is impossible to access the characteristics of the internal object representations.

In this paper, we will demonstrate the interplay between image information and object representations in a discrimination task. We will first define image information and characterize the extent to which the representations of these objects differ, after taking into account differences in the information provided by images of different objects. We use three classes of simplest objects: a set of unconnected balls, the same balls connected by thin lines, and the same balls connected by cylinders (Fig. 1). Informally, we refer to these objects as balls, tinker toys, and wires. We will measure object discrimination in terms of generalization from familiar to novel viewpoints, and show that the balls are harder to discriminate than the tinker toys, which in turn are harder than the wires. A priori, this result could be accounted for either by differences in the representations learned for each object class, or by the increased information about the 3D structure inherent in images of less complex objects, or by both. We will quantify image information for these objects, and show that after the image information has been taken into account, human observers are increasingly more efficient in discriminating the balls, the tinker toys, and the wires. This strongly suggests that the internal representations of these objects are different, perhaps qualitatively. We conclude by discussing the putatively different representations of these objects in the brain.

2. Image information and the ideal observers

We have introduced three classes of objects in decreasing order of object complexity. The 3D configuration of these objects (in each row in Fig. 1), as defined by the relative positions of the balls, is the same. Yet the tinker toys and wires have the reduced object complexity by being connected, in that the balls are chained in an ordered sequence. The wire objects reduced the object complexity even further by having surfaces, the shading of which provides information about the 3D structure of the objects. This means that the information provided by images of the three classes of objects is qualitatively different, with images of wire objects providing the most information, followed by the tinker toys and the balls. We can quantify the differences in the information relevant to a particular experimental task by simulating ideal observers for the different classes of objects. Ideal observers are statistically optimal estimators of some unknown stimulus variable, given the information provided in a stimulus. In an object categorization task, for example, the ideal observer would be the estimator that categorizes objects with the lowest error rate theoretically possible, given the available stimulus information. For a general introduction to the use of an ideal observer in object recognition, see Liu, Knill & Kersten (1995) and Tjan, Braje, Legge & Kersten (1995).

The reliability of the information in images of the balls, tinker toy, and wire objects is clearly different. Some are inherently more ambiguous than others. One way to characterize the ambiguity is to calculate the number of bits of added information needed to determine which of an infinite set of object models matches the image. Images of collections of balls are clearly the most ambiguous. No 3D information is provided about

[1] See Liu (1996) for a similar critique regarding the prolonged debate on the viewpoint-dependence of internal representations in object recognition (Tarr & Bülthoff, 1993; Biederman & Gerhardstein, 1995).

the relative depth of the balls (under orthographic projection and assuming no occlusions). Moreover, the correspondence between balls in the image and any internal model of a collection of balls is ambiguous. A simple calculation shows that for an object composed of n balls, the number of bits needed to correctly determine its 3D structure is $np + (\log_2(n!) - 1)$ bits (assuming p bits are needed to specify the depth of each ball, and that the (x, y) coordinates of each ball are known accurately). The first term reflects the depth ambiguity, the second the matching ambiguity (up to a head-tail ambiguity). Images of tinker toy objects disambiguate the correspondence between balls in the image and balls in an internal object representation but provide no 3D information; thus, the ambiguity is np bits. Wire objects similarly disambiguate the correspondence between an image and an internal model, but also, with surface shading, theoretically provide enough information to completely reconstruct the 3D structure of an object, leaving 0 bits of ambiguity.

We accordingly define object complexity as the additional number of bits needed to specify an object shape relative to a comparison object.

In a realistic experimental setting, the ambiguities described above will not be enough to constrain performance, since subjects typically have a finite (and small) number of possible choices. In order to effectively measure differences in the information content of images derived from different object classes, we must add uncertainty to stimuli and use a task for which we can derive exact ideal observers. The performance of these ideal observers serves as operational measures of the uncertainty in stimuli derived from different object classes for a specific experimental task. Clearly, the ideal observer for the wire objects will perform better than the ideal observer for the balls objects. However, simulations of the ideal observers on the same stimuli and experimental task used to measure human performance are necessary to compute the relative informativeness of the two classes of stimuli.

In the current work, we use a task in which subjects must decide which of two distorted versions of an object is more similar to the learned object. In the remainder of this section we will define the task in more detail and describe the ideal observers for each of the three object classes we used.

We had subjects learn a randomly generated object by viewing it from a number of viewpoints in an initial learning phase. We then made distorted versions of the learned object by adding Gaussian positional perturbations to the (x, y, z) coordinates of each ball of the object, using a fixed standard deviation for the distortions, σ_t (for the tinker toy and wire objects, the lengths of the lines and cylinders were adjusted accordingly). Images of these objects generated using orthographic projection served as target stimuli. In an experimental trial, subjects were shown a target stimulus alongside a distractor, created by adding larger Gaussian positional distortions to the learned object σ_d ($\sigma_d > \sigma_t$) and viewed from the same viewpoint as the target stimulus. They were asked to judge which of the two stimuli was more similar to the learned object, measured as the squared distance between corresponding balls of the input object and the object model. Subjects' performance was measured as the amount of distractor noise σH/d needed to correctly judge, 75% of the time, the target stimulus as more similar to the learned object.

In order to quantify the information content of stimuli for this task, we simulated ideal observers (derived independently for each of the three object classes) on the same task and measured the threshold levels of distractor noise σI/d needed for the ideal observer to achieve the same level of performance (75% correct). We have proved in Liu, Knill & Kersten (1995) that the threshold performance of a human observer $\sigma H/d$ relative to that of the ideal σI/d in the form of:

$$E = \frac{(\sigma_d^I)^2 - \sigma_t^2}{(\sigma_d^I)^2 - \sigma_t^2} \tag{1}$$

is exactly the statistical efficiency. Namely, the effective image information used by a human observer relative to that by the ideal.

In what follows, we will describe the conceptual derivation of the ideal observers for the three object classes (Appendix A). The ideal observers consider all possible viewing positions and give rise to the best guess as to the image that is more similar to the learned object. Theoretically, there are sufficient views to recover the exact light source direction and 3D structure from the learning views given specific prior assumptions (Appendix A). We assume that the ideals know the constant Gaussian variance σ_t^2, and that the center of mass of every object is positioned at the origin of the coordinate system (see Werman & Weinshall (1995) for a proof that aligning the center of mass of each object is the optimal strategy in dealing with translations). It computes the conditional probability P(image/object) for both the target and distractor images and chooses the one with the larger value as being more similar to the learned object.

For the balls, the ideal observer has available the (x, y) coordinates of each ball. From each viewing angle, it projects the 3D object model onto a 2D image, computes the squared Euclidean distance D^2 between the projected image and the input image, and converts it into the probability measure through the Gaussian function $(2\pi\sigma_t^2)^{-n/2} \exp(-D^2/2\sigma_t^2)$. Since the correspondence between the balls in the two images is unknown, all possible combinations must be considered for a match against the internal model. The ideal observer then integrates over these probability measures throughout the viewing sphere, and the probability measure is obtained for this input image.

For the tinker toys, the ideal observer has the additional information about the ordering of the sequence of the balls except that it does not know which end is the head and which the tail. Therefore, only these two correspondence possibilities will be considered in the overall probability computation.

For the wires, we will show in the Appendix A that the 3D structure of the input object can be precisely reconstructed in principle. The ideal observer therefore has available the (x, y, z) coordinates of the balls from the input. It does not know the orientation of the input 3D object relative to the learned 3D model, nor which ball is the head and which the tail, so all possible relative orientations and the head-tail uncertainty will be considered.

Given the decreasing complexity of these object classes, it is expected that the ideals' discrimination performance will be improving. What is not obvious is the relative performance of humans versus the ideals, or the statistical efficiency, for these objects. In the next section, we will first present data to confirm our expectation that the objects are increasingly easier to discriminate for both human and ideal observers. We will then show that the pattern of efficiency results in the same order. Consequently, the differential performance can be accounted for in part by the different image information, but not completely. The remaining difference has to be due to internal processing of the stimuli, such as would be caused by differences in the internal representations of the different classes of objects.

3. Experiment

3.1. Stimuli

Fig. 1 shows the three classes of objects used in the experiment. The diameters of the cylinders and the balls were the same, the balls were therefore not clearly visible for the wire objects. The diameter of each ball was 0.20 cm. The prototype objects were generated by positioning a sequence of balls in space with random relative orientations but with a fixed distance between neighbors (2.54 cm). The diameter of the thin lines was 0.0254 cm. The stimuli were rendered with matte Lambertian reflectance and a point light source at infinity, with a 0° tilt and 63.44° slant.

The images were rendered on a Stellar ST2000 computer, using the Doré 3D graphics package. The viewing distance was 60 cm, and thus, the diameter of a ball was 0.19° in visual angle.

3.2. Method

Each prototype object was tested in a block. The order of the test was counter balanced across the subjects. Within each block, the procedure was as follows.

3.2.1. Learning

The prototype object was rotated around the x-axis (horizontal in the screen plane) in six steps, 60°/step, and then around the y-axis (vertical) in six steps, resulting in 11 views.

3.2.2. Practice

The subject was presented with two images shown side by side. One was a learning view of the prototype (chosen randomly from the 11 views), the other a view of a distractor object. The distractor was generated either by adding positional distortions to the vertices of the prototype object or by randomly generating a new object of the same class. The subject decided which of the two objects was the prototype by pressing a response key, with feedback. The subject needed 20 correct responses in a row to pass this stage, up to 100 trials maximum.

3.2.3. Test

Two objects were presented side by side. Both were generated from the same prototype object, and rendered from the same viewing angle—randomly selected from a uniform distribution on the viewing sphere, plus a random rotation in the image plane. Hence, the viewing angle was almost always novel. The target was generated by adding Gaussian positional distortions to the vertices of the prototype, with a 0.254 cm standard deviation. The distractor was generated similarly, except that its standard deviation was greater. We used a staircase procedure to find the standard deviation of the distractor that kept the subject 75% correct (Watson & Pelli, 1983). The number of trials per object per subject was 100. No restrictions were imposed on the range of the distractor standard deviation otherwise. No feedback was given either.

Two naive subjects with no psychophysics experience and the three authors participated.

3.3. Results

3.3.1. Thresholds

Fig. 2 shows the thresholds of distractor noise at 75% correct for each of the three object classes, and for both human and ideal observers. Since it is important to obtain accurate estimates of the threshold performance for the ideals, we used six objects in each object class (as opposed to three for humans)[2] and 2000 trials per object (as opposed to 100 for humans)[3]. Since the 3D

[2] More objects were used simply to get rid of any idiosyncrasies of individual objects. So that an ideal observer's performance characterizes its object class.

[3] More trials were used simply to obtain more reliable estimates.

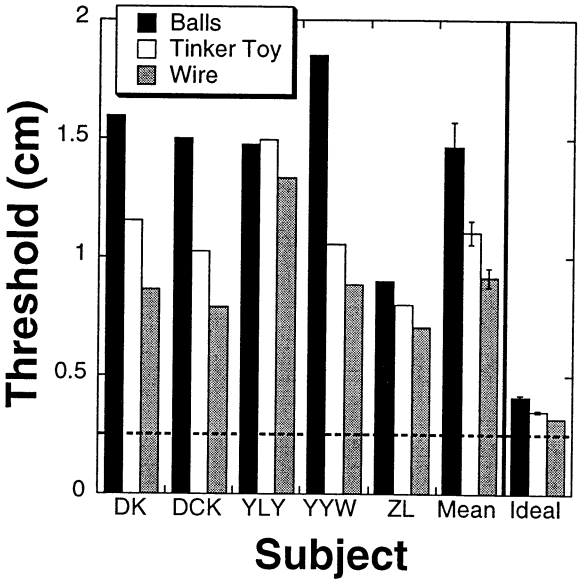

Fig. 2. Individual subject's performance, their averages, and the averages of the ideal observers' performance, for each of the three object classes: balls, tinker toys, and wires. The error bars represent standard errors. The error bars of the ideal performance were small, indicating accurate estimates of the thresholds. (For the wires, the ideal's average threshold is 0.32 cm, very close to the theoretical estimate of 0.31 cm when the 3D pose of the input object is assumed known) (Liu, Knill & Kersten, 1995).

positions of the balls were identical across the object classes in each row in Fig. 1, the objects were matched across object class. With this matching, we conducted a Friedman rank-order test for the human data (Hays, 1988). The threshold difference between the three object classes was statistically significant ($X_r^2 = 63.07$, $d.f. = 2$, $P < 0.001$). A further comparison between the wires and the tinker toys using the Wilcoxon test for matched pairs yielded a significant difference ($T = 11$, $N = 15$, $P < 0.003$), suggesting that it was easier to identify the wires than the tinker toys. A similar comparison between the tinker toys and the balls also yielded a significant difference ($T = 31$, $N = 15$, $P < 0.05$), suggesting that it was easier to identify the tinker toys than the balls[4].

Since the 3D positions of the balls for each of the six objects used by the ideals were also the same across the object classes, we conducted a matched item analysis for the ideals' thresholds. The difference between the ideals' thresholds in the three object classes was significant (0.41, 0.35, and 0.32 cm; $F(2, 10) = 67.09$, $P < 0.001$). A further analysis showed that the difference between the balls and tinker toys was significant ($F(1, 5) = 54.51$, $P < 0.001$), and so was the difference between the tinker toys and wires ($F(1, 5) = 37.59$, $P < 0.002$).

3.3.2. Efficiencies

We calculated statistical efficiencies according to Eq. (1) for each subject and each object class. An overall analysis showed that the efficiencies between the three object classes were significantly different (6.15, 10.23, and 15.56%; $F(2, 8) = 17.94$, $P < 0.001$). A further comparison showed that the efficiencies between the balls and tinker toys were significantly different ($F(1, 4) = 11.92$, $P < 0.026$), so were those between the tinker toys and wires ($F(1, 4) = 20.91$, $P < 0.01$). Fig. 3 shows the average efficiencies for the three object classes. These results indicate that, when the difference in image information has been taken into account, human subjects were better at discriminating the wire objects than the tinker toys, which in turn were better than the balls.

4. Discussions

4.1. Understanding the differences in efficiency

Subjects' thresholds for correctly labeling test objects decreased significantly as object complexity decreased. Some of this decrease can be explained by the objective differences in the information provided by stimuli for each object class; however, subjects' pattern of efficiency across the different object classes suggests that stimulus information cannot account for all of the difference. For example, even accounting for the large differences in information provided by images of the balls objects and the wire objects, subjects remained more than twice as good (in terms of efficiency) for the wire objects. In order to explain the differences in

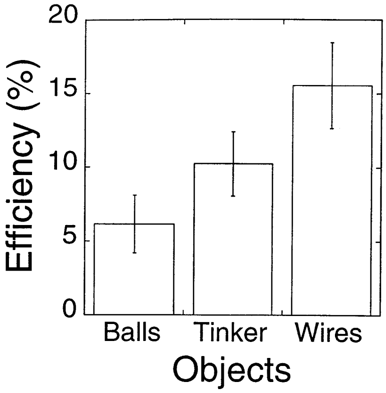

Fig. 3. Statistical efficiency of the human subjects relative to the ideal observers. The efficiencies between the different object classes are significantly different, implying that the internal representations for these objects are also different.

[4] The performance of Subject ZL, one of the authors, was substantially better than that of YLY, a naive subject with no previous psychophysics experience. Although we could not find a reliable learning effect, we believe that such a difference may be due to learning, as ZL programmed the experiment and was extremely familiar with the object classes.

efficiency across object classes, we must look into differences in the visual processing and representation of stimuli within each class. Towards this end, we will look at the possible sources of inefficiency in performing the task. This will allow us to conclude that the empirically measured differences in efficiency reflect differences in the higher-level representations of the different classes of objects or in the process that matches stimuli in the testing phase to stored representations.

Inefficiency arises from a number of possible sources within an observer. We will classify them broadly into three types: (1) encoding of the input objects on a test trial; (2) initial learning of an internal representation; and (3) matching the inputs on each test trial against the learned representation. The first of these is a generalization of what we have previously referred to as the information content of stimuli, which takes into account the possibility that early visual processing may be more or less efficient for the different classes of objects. To the extent that we can discount effects of (1) as an explanation of the results, we can use the results to draw inferences about higher-level representational and matching processes.

4.1.1. Encoding test stimuli

A common source of inefficiency across all object classes is that humans are less accurate than the ideals in representing the effective vertex positions of the stimulus objects. In addition, the degree to which humans use available input information to encode a stimulus object depends on the pertinent object class. For a balls object, the ideal observer encodes the (x, y) positions of the balls, and considers all possible orderings. Humans may consider only a small subset of the $5! = 120$ total possible orderings. For a tinker toy object, the ideal observer encodes the (x, y) positions of the balls and the head-tail sequence possibilities. Since there are only two sequence possibilities and five balls, the humans may well be able to represent the two sequences. Therefore, the encoding inefficiency for a tinker toy may be largely in encoding the effective (x, y) positions of the balls. For a wire object, the information is sufficient for the ideal observer to encode the (x, y, z) positions of the balls. Humans may not be able to extract much 3D information from shading on the cylinder surfaces. Consequently, the coding inefficiency for a wire is greater than for a tinker toy. We remark that that statistical efficiency for the wires is still higher than for the tinker toys strengthens our claim—the internal representation and matching process are more efficient for the wires than for the tinker toys.

4.1.2. Internal representations

Since there are sufficient views for each object and there is a training phase with feedback, the ideals presumably can reconstruct the (x, y, z) coordinates of the vertices for all objects. Humans, however, may construct different representations for different object classes. For a balls object, since correspondence between one view to the next is difficult and virtually no 3D information is available from each view, they may store the training views as 2D templates as the internal representation. For a tinker toy object, humans may represent the prominent object parts—the positions of balls, while the lines connecting the balls serve to reduce the correspondence ambiguity. The internal representation of a tinker toy may be its 3D structure. For a wire object, humans may represent in 3D the cylinders as the prominent object parts. From each view, the length and shading pattern of each cylinder suggests a clearer orientation in 3D of the cylinder than the line in a tinker toy does. The resulting representation may be a more precise 3D structure. In sum, the less complex an object is, the more efficient it may be represented internally.

4.1.3. Matching

Even if humans could represent the three object classes equally precisely, they would still not be equally efficient in matching the internal representations to an input stimulus. First of all, unlike an ideal observer, humans are unable to integrate over all correspondence possibilities. They will be most inefficient for a balls object because the ambiguity is the greatest. Second, humans are unlikely to be able to ‘rotate’ the internal representation in depth in large angles. Perhaps it is easier to mentally ‘rotate’ a wire object than a balls object. Therefore, when everything else is equal, the matching may be the least efficient for the balls, and the most for the wires.

4.2. The interplay between information and representation

So far we have focused on image information and internal representations, and argued for their importance in interpreting object recognition performance. However, we believe that an even more important determinant that constrains both image information and object representations is object structural complexity[5].

The structural organization of objects plays a major role in object representation and recognition, both because it constrains the ‘language’ with which object categories are represented and because it is a major factor in determining the image information that is

[5] This does not imply that object complexity is the only constraint. For example, stimulus contrast will surely influence input information and therefore the resulting performance. We thank the anonymous reviewer for pointing this out.

useful for learning and recognizing objects from both familiar and novel views. We will consider these two issues in turn.

4.2.1. Constraints on the internal representation

The qualitative differences between the three object classes used in the experiment suggest different types of representations for each object class. The natural representation for the balls may be a specification of the positions of the balls. For the tinker toys, their balls remain the salient 'parts', but the neighboring relations specified by the connecting lines suggest a more compact representation in terms of relative positions of pairs of neighbors (indeed, they may constrain the representation to be of this type). For the wire objects, the cylinders connecting the balls are the salient parts, and their balls, which now appear as vertices of the objects, serve as joints between the parts. A natural representation of such objects may be in terms of the lengths of the cylinder segments and the 3D angles between neighboring segments.

The learning and the use of different representational schemes for internal models of the balls, tinker toy, and wire objects would have had, without the ideal observer analysis, explained the improvement in subjects' performance with decreased object complexity. It is reasonable to assume that the visual system is designed to represent spatial relationships between connected parts of objects, rather than between disjoint parts. Furthermore, the system may well be optimally designed to represent the relationship between contiguous parts of objects. This would explain the progressive improvement in performance from objects that match neither of these constraints (the balls) to objects that match one of them (the tinker toys have well specified part relations, but their arguably salient parts, the balls, are not contiguous) to objects that match both of them (the wires). The complexity of the metric representation needed to specify and match the balls objects is certainly greater than that needed for the tinker toy and wire objects, which by itself would explain at least one aspect of the results.

4.2.2. Image information available for object learning and recognition

The second factor that affects the relationship between object structure and object recognition is the information provided by views of an object for learning object models and recognizing individual views of objects. The different classes of objects in the experiment admitted different amounts of image information about the same 3D configuration. Clearly, images of the last two classes of objects provide more information for recognition than those of the balls, since they disambiguate the matching of the parts between object views and any putative object models. Furthermore, a wire object provides significantly more information about its 3D shape than the other objects. This includes the visually apparent shading on the cylinder segments, which varies with its orientation; salient occlusion information, which disambiguates depth ordering at intersections in the image between cylinder segments; and the shape of the creases formed at the joints between cylinder segments, which provides a strong clue to relative segment orientation in 3D. The improvement in discrimination performance could therefore have, again, without the ideal observer analysis, resulted entirely from differences in the information content of the different classes of stimuli.

The two different factors that could have played a role in determining performance in this experiment (representation and image information) are in reality strongly interdependent. The nature of object representations is inextricably tied to the nature of the information provided about the objects. For example, the 3D information provided in the images of the wire objects specifies the relative orientations of the segments, not the relative 3D positions of the vertices. Thus, it makes sense for the visual system to represent segment orientations and lengths rather than vertex positions in any learned model of such objects. The tinker toy objects, however, are different in that very little 3D information was available from any single view of these objects in the experiment. In more natural viewing conditions, with stereo and motion information available, the most reliable 3D information would directly specify the spatial layout of the balls rather than the connecting lines, suggesting a different class of representation for the tinker toy objects than for the wire objects. The same argument is even stronger when applied to the balls. Thus, the available image information strongly constrains the class of representation, which the visual system can use for storage and retrieval of object information. Since object structure has similar effects both on what images of objects provide information about and on what is represented, we can expect the predictions of a purely information-based explanation and a purely representation-based explanation to be, more often than not, consistent with one another.

This highlights, on one hand, the need to jointly consider the effects of object structure on information and representation when studying object recognition based on an analysis of the object class complexities and how these complexities affect image informativeness for that structure. On the other hand, this also highlights the difficulty we are facing in interpreting human object recognition performance before we can objectively characterize stimulus information.

"Why should internal representation and information content be decomposed into two different contributers? It might be, with equal plausibility, that more information result in a richer representation that, in turn, will allow better discrimination performance"[6]. We believe that there is no conflict between distinguishing the contributions from the two and the interplay between them. Precisely because we can quantify the contribution from the input information, we can turn the speculation above into empirical evidence. In other words, only after analyzing the information content, can we claim that more information results in a richer representation, and that the better discrimination is therefore due to both the richer information and representation.

Other researchers have also studied the relation between information and representation in object recognition. Tjan & Legge (1998), for example, quantified stimulus information by adding luminance noise to an image, with an ideal observer that represents this image in pixel values. They demonstrated that the degree to which one object can be distinguished from another depends strongly on the object set. They therefore argued that human performance in distinguishing these objects should be determined in part by the stimulus information. Our contribution in this paper is that we could reasonably assume a form of representation (where the information is) far beyond a raw image, and explicitly analyze the contribution of various nameable sources of information such as correspondence, connectivity, occlusion, shading, and the geometry of creases formed by two connecting cylinders.

In this respect, we believe that the methodology introduced in this paper is unique and essential to the analysis of human object representation.

Acknowledgements

We thank John Oliensis, Sebastien Roy, the anonymous reviewer, and Ellen Hildreth (the editor) for many helpful suggestions. We thank Bosco Tjan in particular for his many insightful discussions.

Appendix A

We describe the details of the ideal observer's calculation, the probability that an image is from a 3D object model, for each object class. We assume that in all cases the (x, y) coordinates of the balls are known. We assume also that no correspondence between the balls in an image and in the model is known for the balls objects, and that correspondence is known for the tinker toy and wire objects up to a head-tail ambiguity.

We first show that the 3D structure of a wire object can be completely recovered from its image under the following assumptions:

1. The surface albedo of the cylinders is an unknown constant.
2. The shading is Lambertian.
3. Image projection is orthographic.
4. The 3D object model in every object class can be reconstructed in the learning phase. Specifically, since an object is rotated around the x-axis, each ball's y-coordinate is unchanged, and therefore the correspondence between one image and the next is easy to establish. This is true for rotation around the y-axis as well. According to Ullman's structure-from-motion theorem (Ullman, 1979), four points and three views are required to reconstruct the 3D structure of the object (with a depth reversion uncertainty). Since each rotational step is 60° with a known rotational direction, the depth reversion can be disambiguated. Since each object has 11 views available as opposed to the minimally required four, and since there is a practice phase with feedback, it is reasonable to assume that for the ideal observer the uncertainty associated with reconstructing the (x, y, z) coordinates of an object is negligible. As we are comparing between the three object classes, specifying, the (x, y, z) coordinates to the same precision does not introduce any bias.
5. The direction of the point light at infinity can be recovered in the learning phase and is therefore known. Since the 3D model of an object is known in the learning phase, then the pose of the object in any image in the learning phase is also known. From the shading patterns on the cylinder surfaces, the illumination direction can be then derived. The uncertainty about such direction is assumed negligible as there are altogether 11 images per object, and there are altogether three objects and four cylinders per object (the same illumination direction is in theory also recoverable from the shading patterns on the balls from the remaining six objects).
6. The cross section of a cylinder is known to be circular.
7. The 3D orientation of each cylinder in a wire object is recoverable in the test phase. The goal here is to determine the cylinder's orientation that has two degrees of freedom. We start by picking an arbitrary straight line on the cylinder surface along its axis. As the illumination direction is known, the surface normal associated with the line is constrained on a set of concentric circles in the two-dimensional orientation space, centered around the illumination direction. Each circle corresponds to one specific surface albedo value. Since the surface normal is perpendicular to the cylinder's orientation vector,

[6] We thank the anonymous reviewer for raising this point.

this vector is also constrained to a set of concentric circles. As the cylinder has to project to the same given image, its degrees of freedom is reduced to one. Therefore, the vector is now constrained to a line in the orientation space. Similarly, we can pick another straight line on the cylinder surface that has a different shading value. This will constrain the cylinder's orientation vector to another line in the orientation space. The two lines in the orientation space must intersect. Because the cross section of the cylinder is known to be circular and the image projection is orthographic, the locations on the image plane of the two straight lines on the cylinder surface suffice to specify the angular separation of their surface normals. Therefore, the relative orientation of the two lines in the cylinder's orientation space can be precisely specified, so is their intersection. Therefore the orientation of the cylinder is completely specified.

Since the 3D pose of an object from the input image relative to the 3D pose of the 3D object model is unknown, all viewing possibilities will be considered with an equal likelihood. Computationally, 32785 points on the viewing sphere are chosen whose distribution on the spherical surface is maximally uniform (the convex hull enclosed by these points has the largest possible volume (S. Roy, personal communications)). From each viewing direction, a specific pose of the 3D object model is taken, and the object is projected into a 2D image for the balls and tinker toy objects. The center of mass of the model and that of the input image are aligned to yield the closest match in translation (Werman & Weinshall, 1995). The model is then rotated around the viewing direction to find the closest match to the input image. It is easy to find the angle θ that minimizes the Euclidean distance between the model vector in the image plane

$$(x_{\mathrm{M}}^1, y_{\mathrm{M}}^1, x_{\mathrm{M}}^2, y_{\mathrm{M}}^2, \ldots), \tag{2}$$

and the input image vector

$$(x_{\mathrm{I}}^1, y_{\mathrm{I}}^1, x_{\mathrm{I}}^2, y_{\mathrm{I}}^2, \ldots). \tag{3}$$

The squared distance $D^2(\theta)$ between the two vectors as a function of their relative rotational angle θ is:

$$\begin{aligned} &D^2(\theta) \\ &= \textstyle\sum_k (x_{\mathrm{I}}^k - x_{\mathrm{M}}^k \cos\theta - y_{\mathrm{M}}^k \sin\theta)^2 \\ &\quad + \textstyle\sum_k (y_{\mathrm{I}}^k + x_{\mathrm{M}}^k \sin\theta - y_{\mathrm{M}}^k \cos\theta)^2. \end{aligned} \tag{4}$$

Let

$$\frac{\mathrm{d}}{\mathrm{d}\theta} D^2(\theta) = 0, \tag{5}$$

we have

$$\tan\theta = \frac{\sum_k (x_{\mathrm{I}}^k y_{\mathrm{M}}^k - y_{\mathrm{I}}^k x_{\mathrm{M}}^k)}{\sum_k (x_{\mathrm{I}}^k x_{\mathrm{M}}^k + y_{\mathrm{I}}^k y_{\mathrm{M}}^k)}. \tag{6}$$

Effectively, we have used a rotation invariant measure for 2D image comparisons for the balls and tinker toy objects. We use this rotation invariant calculation to approximate the true Bayesian method that integrates θ over $[0, 2\pi)$ because of the prohibitive simulation time for the balls objects[7]. For comparison purpose, we use this rotation invariant measure for the tinker toys as well.

For an image of a wire object, the z-coordinates are retained after image projection. So the D^2 (θ) has the contribution from the z-coordinates as well. In this case, we integrate θ over in $[0, 2\pi)$ in the image plane. This gives rise to a slightly better performance for the ideal than the rotation invariant approximation does. Accordingly, the corresponding efficiency for the wires is also slightly lower. The fact that the efficiency is still significantly higher than the other object classes attests that humans are indeed more efficient in representing the wires than the tinker toys and the balls.

[7] In fact, aligning the center of mass of two images rather than integrating the (x, y) over the entire image plane, which the Bayesian method requires, is also an approximation.

References

Biederman, I., & Gerhardstein, P. C. (1995). Viewpoint dependent mechanisms in visual object recognition: a critical analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1506–1514.

Farah, M. J., Rochlin, R., & Klein, K. L. (1994). Orientation invariance and geometric primitives in shape recognition. *Cognitive Science*, *18*, 325–344.

Hays, W. L. (1988). *Statistics*. New York: Holt, Rinehart, and Winston.

Johanssen, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, *14*, 201–211.

Liu, Z., Knill, D. C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research*, *35*, 549–568.

Liu, Z. (1996). Viewpoint-dependency in object representation and recognition. *Spatial Vision*, *9*, 491–521 Special Issue on Perceptual Learning and Adaptation in Man and Machine.

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

Tjan, B. S., Braje, W. L., Legge, G. E., & Kersten, D. (1995). Human efficiency for recognizing 3D objects in luminance noise. *Vision Research*, *35*, 3053–3070.

Tjan B.S. & Legge G.E. (1998). The viewpoint complexity of an object-recognition task. *Vision Research*, *38*, 2335–2350 (Special issue: models of recognition).

Ullman, S. (1979). The Interpretation of structure from motion. *Proceedings of Royal Society of London*, *B*, *203*, 405–426.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception and Psychophysics*, *33*, 113–120.

Werman, M., & Weinshall, D. (1995). Similarity and affine invariant distances between 2D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*, 810–814.