Imperfect invariance to object translation in the discrimination of complex shapes

Marcus Dill
Center for Biological and Computational Learning
MIT E25-201, Cambridge, MA 02142, USA

Shimon Edelman[*]
Department of Psychology, 232 Uris Hall
Cornell University, Ithaca, NY 14853-7601, USA

June 1999; revised July 2000; second revision November 2000

Abstract

The positional specificity of short-term visual memory for a variety of 3D shapes was investigated in a series of *same-different* discrimination experiments, using computer-rendered stimuli displayed either at the same or at different locations in the visual field. For animal-like shapes, we found complete translation invariance, regardless of the inter-stimulus similarity, and irrespective of direction and size of the displacement (experiments 1 and 2). Invariance to translation was obtained also with animal-like stimuli that had been "scrambled" by randomizing the relative locations of their parts (experiment 3). The invariance broke down when the stimuli were made to differ in their composition, but not in the shapes of the corresponding parts (experiments 4 and 5). We interpret this pattern of findings in the context of several current theories of recognition, focusing in particular on the issue of the representation of object structure.

[*] To whom correspondence should be addressed. E-mail: se37@cornell.edu

1 Introduction

Any visual system that aims to attain object constancy must address Höffding's problem: how to treat equivalently two images in which the same object appears in different locations in the field of view (Neisser, 1967).[1] The common phrasing of this problem suggests that the human visual system is expected to achieve constancy or invariance in the face of translation as a matter of routine (even when normalization of the stimulus location through eye movements is discounted). In practice, the issues surrounding translation invariance turn out to be more complicated. In the present study we explore the conditions under which translation invariance in same/different shape discrimination starts to break down.

1.1 Prior results

Our work has been motivated by reports of non-invariant processing of shapes, which qualify the notion of constancy under translation. Early evidence of an effect of translation on shape perception can be found in (Wallach and Austin-Adams, 1954). In that study, briefly shown 2D shapes primed the subjects' perception of ambiguous figures displayed at the same location as the prime, but not the perception of figures shown at an analogous location in a different quadrant. A similar confinement of the facilitation effect to a quadrant has been found in the subliminal priming study of (Bar and Biederman, 1998).

Data from a number of other experiments, mostly involving 2D patterns, also indicate that the recognition of novel complex stimuli is not completely invariant under translation (Foster and Kahn, 1985; Nazir and O'Regan, 1990; Dill and Fahle, 1997a; Dill and Fahle, 1997b). If, for example, subjects have to determine whether two sequentially flashed random-dot clouds are *same* or *different*, decisions are faster and more frequently correct when both stimuli are presented at the same rather than at different locations in the visual field (Foster and Kahn, 1985; Cave et al., 1994; Dill and Fahle, 1997a). This *displacement effect* has been shown to be gradual (i.e. larger displacements produce poorer performance), and to be specific for *same* trials. Control experiments ruled out explanations in terms of afterimages, eye movements, and shifts of spatial attention (Dill and Fahle, 1997a). Similarly, Larsen and Bundesen (1998) found that the d' measure of

[1] This problem is named after the 19th century psychologist who claimed that the need to achieve object constancy, and, in particular, equivalence across translation, is a stumbling block for associationism.

performance decreased monotonically with spatial separation, but only if the patterns differed both by a translation and a rotation.

While *same-different* matching involves only short-term memory in the range of a few seconds, Nazir and O'Regan (1990) also found positional specificity in learning experiments that lasted at least several minutes. They trained subjects to discriminate a complex dot pattern from a number of distractors. Training was restricted to a single location in the parafoveal field of view. Having reached a criterion of 95% correct responses, subjects were tested at three different locations: the training position, the center of the fovea, and the symmetric location in the opposite visual hemisphere. Discrimination accuracy dropped significantly for the two transfer locations, while at the control location the learned discrimination was not different from the training criterion. Dill and Fahle (1997b) have isolated two components of the learning effect in this case. Immediately after the first few trials, subjects recognize patterns at a level clearly above chance. From this rapidly reached level, performance increases in a much slower learning process, until the accuracy criterion is reached. This learning process can last up to several hundred trials. Dill and Fahle showed that accuracy at transfer locations is at about the same level as the performance at the beginning of the slower learning process. This suggests that the fast component — immediate recognition — is translation-invariant, while the slower process involving perceptual learning is much more specific to the location of the training.

1.2 The role of novelty

The basic requirement imposed on the stimuli in psychophysical studies of invariance is novelty: if the stimuli are familiar, the subjects are likely to have been exposed to their transformed versions prior to the experiments. Because of this constraint, stimuli both in *same-different* matching and in learning studies tend to be somewhat unnatural and difficult to process. With more familiar patterns, the performance may well be insensitive to retinal translation. Biederman and Cooper (1991) tested subjects with line drawings of familiar objects and asked them to name the object. Repeated presentation reduced the naming latency in a manner independent of the relative location in the visual field of the priming and the test presentations. Part of the priming effect, however, may have been non-visual: Biederman and Cooper found a reduction of the naming latency also if a different instance of the same object class was presented (e.g. a flying bird instead of a perched one). As pointed out by Jolicoeur and Humphrey (1998), the visual part of the priming effect may be too small to detect an influence of position, size or other transformations.

Stimulus novelty as such does not automatically lead to a breakdown of translation invariance. Novel 3D shapes of the "paper-clip" variety, for which significant deterioration of recognition with rotation in depth has been previously reported (Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992), yielded no effects of position (Bricolo and Bülthoff, 1992; Bricolo and Bülthoff, 1993; Bricolo, 1996). Interestingly, reducing the paper-clips to 3D "clouds" of points by omitting the limbs that connect the vertices brought back the translation effect, in line with the results of (Nazir and O'Regan, 1990) and of (Dill and Fahle, 1997b).

In view of all these findings, which range from complete invariance to pronounced effects of translation, much additional work is needed to characterize the conditions under which invariance is to be expected, and to understand the processes that support it. We set out, therefore, to investigate the issue of translation invariance (or the lack thereof), using as stimuli tightly controlled animal-like shapes (cf. Figure 1) that were previously found effective in studying the influence of rotation in depth on object recognition (Edelman, 1995a). Here, we report experimental results indicating (i) that translation invariance in the discrimination of complex 3D objects can be imperfect, and (ii) that the imperfection manifests itself when structure, rather than local information, has to be discriminated.

2 Experiment 1: Discrimination of animal-like objects

The first experiment tested positional specificity of *same-different* discrimination among six animal-like shapes (shown in the left column in Figure 1). An earlier study with this class of objects had yielded highly significant effects of orientation in depth on discrimination rate (Edelman, 1995a). Our goal was to determine whether translation has a similar effect on performance.

2.1 Methods

Subjects. 10 observers participated in experiment 1. Except for the first author, they were undergraduate or graduate students from the Massachusetts Institute of Technology, who either volunteered or were payed for their participation in one-hour sessions. All observers had normal

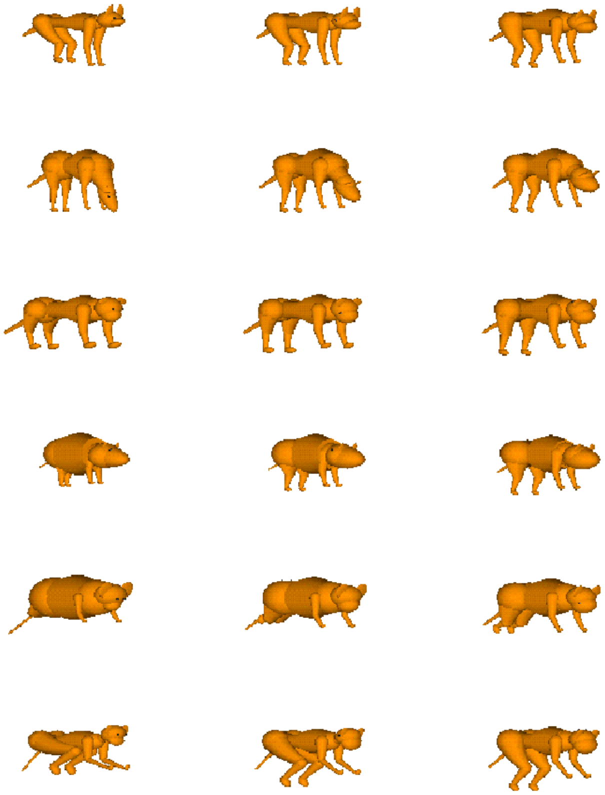

Figure 1: Three levels of similarity (columns) for the six animal-like computer-generated objects. The left column shows the original animals. The similarity between different animals, i.e. within one column, is increasingly larger in the middle and right column. The shapes in the top and the bottom rows in this figure are the original objects used in (Edelman, 1995a); the others are parametric variations, courtesy of T. Sugihara, RIKEN Institute.

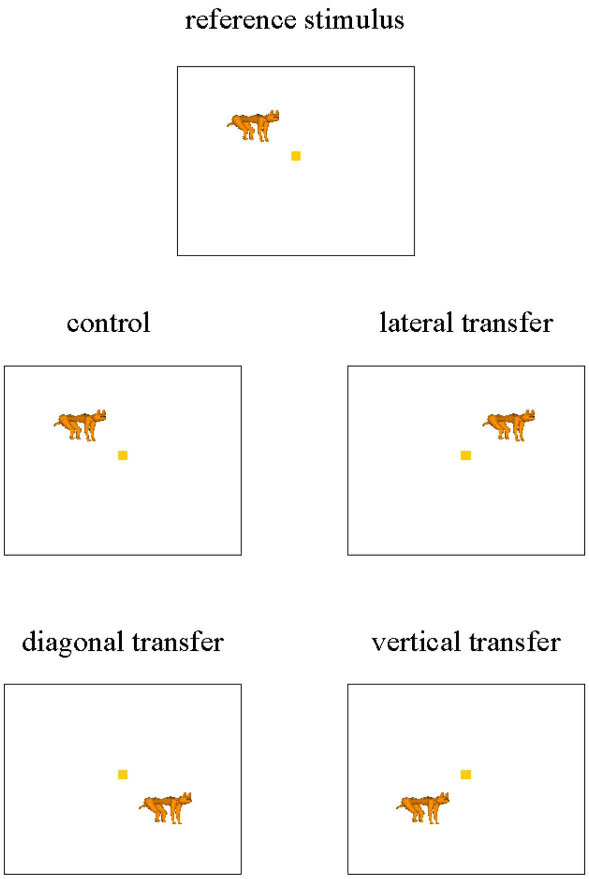

Figure 2: The four transfer conditions used in our experiments. In this illustration, the first stimulus in a trial appears above and to the left of fixation. The second stimulus may appear in the same location (*control* condition), or it may be translated to generate the *lateral*, *vertical* and *diagonal* conditions.

or corrected-to-normal vision. At the beginning of a session, observers were shown examples of the animal stimuli and were informed about the design of the experiment (type and locations of stimuli, presentation sequence and task). They were instructed to keep steady fixation throughout each trial. All subjects were explicitly told that their decisions on pattern identity in each trial should be independent of the stimulus position and should rely only on the identity of the animal.

Apparatus and stimuli. The stimuli were generated and displayed on a Silicon Graphics Indigo workstation (19″ color monitor; refresh rate 120 Hz). The display was viewed binocularly at a distance of 60 *cm*. Each animal-like stimulus shape was defined by a set of 56 parameters representing characteristics such as length, diameter, and orientation of individual limbs (Edelman, 1995a). Six animal classes were used throughout the experiments; see Figure 1. Stimulus images were about 3° wide and 2° high, and could appear at four locations in the upper left, lower left, upper right or lower right quadrant (always at an eccentricity of about 4°). The objects in the images were always tilted and slanted in depth at 45° relative to the observer. The surface color of the animal objects was yellow, the background was dark gray and covered the entire computer screen. The stimuli were presented for 100 *ms*, a time too short to foveate the stimulus by a rapid saccade (Saslow, 1967). To avoid afterimages due to delayed phosphor decay, a stimulus presentation was always immediately followed by four masks. These were composed of 20 random cylinders each, and were presented simultaneously at the four possible stimulus locations, for 300 *ms*. Fixation was aided by a yellow spot of about 0.1° diameter in the middle of the screen. Decisions were communicated by pressing the left (for "same") or the right (for "different") mouse button. A computer beep provided negative feedback immediately after incorrect responses.

Experimental design. At the beginning of a trial, the fixation spot appeared for 1 *s*, followed by the brief display of the first animal stimulus at one location, and the random-cylinder masks displayed at all four locations. After the second presentation of the fixation spot (1 *sec*), the second animal either appeared at the same location (*control*) or at one of the other three positions corresponding to *lateral*, *vertical*, and *diagonal* transfer (Figure 2). Lateral and vertical transfer corresponded to displacements of about 5.5°, while the diagonal displacement was 8°. The onset asynchrony of the two animal stimuli was 1.4 *s*. This long interval and the employment of masks after the first and second stimuli abolish the effects of apparent motion and iconic afterimages (Phillips, 1974). For each *same* trial, the computer randomly chose one of the six animals; in *different* trials, two different animals were randomly selected. Successive trials were separated by a 1 *s* interval.

Experiment 1 comprised 288 trials,[2] divided into three blocks. Observers initiated a block by pressing a mouse button. Trials in each block were balanced for identity (*same* vs. *different*), quadrant in the visual field, and four displacement conditions (control and lateral, vertical or

[2] This number stems from a counterbalancing procedure, whose use was dictated by a cap on the length of the experiment (simply multiplying the number of levels of the different factors would result in too many trials).

diagonal transfer), which were presented in a randomized order.

2.2 Results

For each of the subjects in this and all the following experiments, percentages of correct responses and mean response times (RT) were calculated separately for each of the four displacement (control, lateral, vertical, diagonal) and two identity (*same* vs. *different*) conditions. Trials with RTs longer than 3 s (0.42%) were discarded prior to any further analysis. The mean RT was 488 ms; the correct response rates ranged from 77.0% to 94.5% (mean 85.0%).

Figure 3 shows the correct response rates, averaged across the ten observers. For *same* trials, a 6.9% difference was observed between the *control* condition, i.e., when both animals were presented at the same location, and the mean of the three transfer conditions (89.4% compared to 82.5%). For *different* trials, all conditions yielded approximately the same performance. These qualitative observations were confirmed by a two-way analysis of variance (ANOVA), testing the influence of TRANSLATION (*control*, *lateral*, *vertical*, *diagonal*) and IDENTITY (*same*, *different*) variables. Neither of the main effects was significant, but there was a marginal interaction (F[3,72]=2.17; $p < 0.1$), reflecting differential effects of transfer in the *same* vs. *different* conditions. Post-hoc effects[3] estimated separately for the *same* condition turned out to be significant for the following contrasts: control vs. others (F=4.1, $p < 0.05$), control vs. diagonal (F=4.5, $p < 0.04$); the contrast for control vs. lateral was marginal (F=2.8, $p < 0.10$).

In this and the other experiments described below, we also computed, for each level of TRANSLATION, a bias-free measure of performance derived from Signal Detection Theory (Green and Swets, 1966): $d' = z(H) - z(FA)$.[4] Here, H is the "hit" rate (in the context of the present task, the proportion of correct responses in IDENTITY=*same* trials), FA – the "false alarm" rate (proportion of errors in IDENTITY=*different* trials), and z is the inverse normal cumulative probability function ("z-score"). The mean d' in experiment 1 was 2.2, with no discernible effects of translation. Finally, there were no significant RT effects, and no indication of a speed-accuracy tradeoff.

[3] These can be estimated as an option in the SAS General Linear Models (GLM) procedure in (SAS, 1989).

[4] This "classical" estimate of d' is highly conservative (by a factor of about $\sqrt{2}$), when applied to the analysis of data from a two-interval *same-different* (2IAX) task such as the present one (Macmillan et al., 1977); see also (Macmillan and Creelman, 1991), p.157.

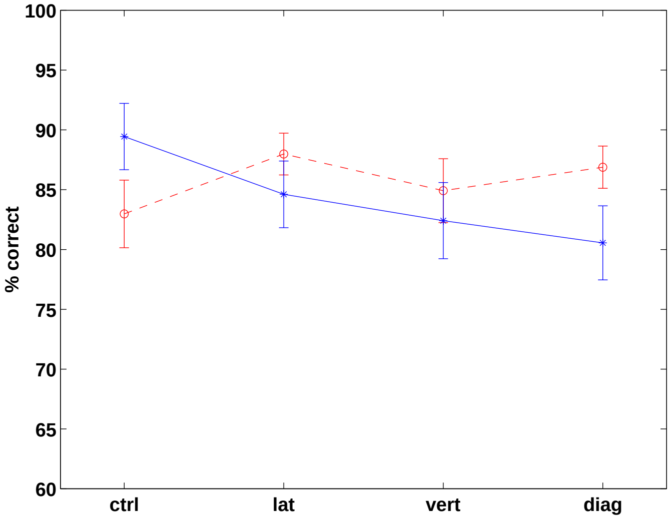

Figure 3: correct response rates by condition in experiment 1 (animal-like shapes). Solid line: *same* trials; dashed line: *different* trials. The points show mean correct rates ±1 standard error ($n = 10$) for the four TRANSLATION conditions (control, lateral, vertical and diagonal).

2.3 Discussion

In experiment 1, a weak effect of translation could be discerned only when the analysis was restricted to *same* trials. The overall influence of translation was not significant, as indicated by the lack of effect on d'. It should be noted that in many previous studies of invariance, analysis has been restricted to *same* trials. Following a variety of arguments (e.g., that a *different* trial does not uniquely correspond to a particular kind of *same* trial, or that recognition can only be investigated for matches, but not for non-matches), *different* trials were either discarded completely or only mentioned in footnotes or appendices. Given the complex nature of decision processes in *same-different* experiments, such omission of *different* trials may lead to an overestimation of the effects, and may result in a wrong interpretation of the available data.

We chose to concentrate on two possible reasons for the difference between our results and the published findings of incomplete translation invariance for dot cloud and checkerboard stimuli. First, the task in experiment 1 may have been too easy. Dill and Fahle (1997a) report that increasing the similarity between stimuli leads to more pronounced positional specificity. Likewise, Edelman (1995b) found that detrimental effects of changes in orientation are larger for similar than for more distinct objects. The animal shapes may have been too easy to discriminate to allow for any significant effect of translation. Second, although the shapes generated by our computer graphics program may not look entirely natural, the class of real animal objects which inspired them is familiar to human observers. It is likely that our subjects had had prior exposure to thousands of animal images, and they may have seen these images at many different locations in the visual field. Any positional specificity that may be observed with novel stimuli could long be lost for a familiar object class, due to this pre-experimental learning process. In the remaining experiments, we examine each of these possibilities in turn.

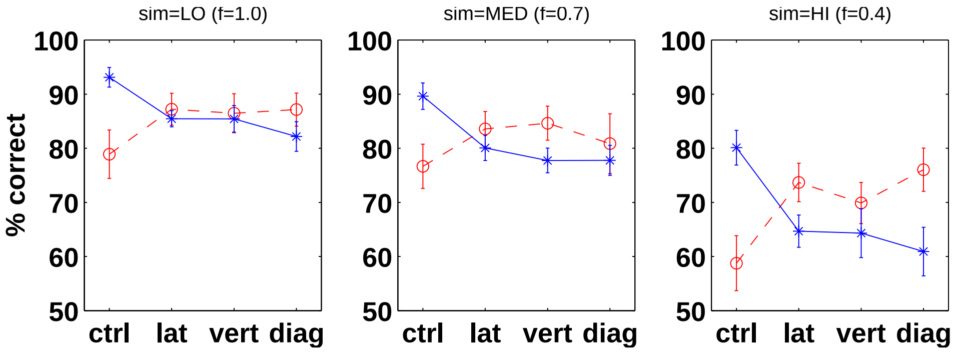

Figure 4: correct response rates by condition in experiment 2 (animal-like shapes, controlled similarity). Solid line: *same* trials; dashed line: *different* trials. The points show mean correct rates ±1 standard error ($n = 16$) for the four TRANSLATION conditions (control, lateral, vertical and diagonal). The three plots (left to right) correspond to the three levels of similarity (low, medium, high) among the stimuli, as explained in section 3.

3 Experiment 2: similarity and invariance

Experiment 2 investigated the influence of similarity among animal-like shapes on translation invariance. As pointed out above, evidence from translation studies with dot clouds indicates that a higher degree of stimulus similarity can lead to a stronger effect of stimulus displacement. Because Edelman (1995a) also found an interaction between similarity and invariance for animal-like shapes, we expected to detect a more pronounced positional specificity following an increase in the similarity of the six animals. To test this idea, we created three sets of six animals by interpolating between the original six animals and a "mean" shape that was computed by averaging each of the

56 model parameters across the six class prototypes. We tested each subject with all three levels of similarity (corresponding to the three columns in Figure 1). To neutralize serial presentation effects, half of the subjects started with the easiest discrimination task, then proceeded to the intermediate and the most difficult tests. The remaining observers were tested with the difficult (highly similar) stimulus set first.

3.1 Method

Stimuli. The same apparatus and stimulus conditions as in experiment 1 were used. To control the level of similarity, we varied the parametric difference between the six animals. For that purpose, the mean 56-parameter vector was computed by averaging the six animal vectors. The experimental objects were then made by interpolating between each of the six original parameter vectors and the mean-animal vector. Under this scheme, the smaller the distance between the interpolated objects and the mean animal, the higher the similarity between the interpolated shapes. We varied this distance by multiplying the parametric difference between the mean and the original vectors by a constant (dis)similarity factor f. Three different factor values were used for the experiment: $f = 1$ (corresponding to the original animals), $f = 0.7$, and $f = 0.4$ (note that $f = 0$ would have produce six interpolated animals identical to the mean).

Experimental design. Each subject was tested in three partial experiments, each with stimuli of a single similarity level only. Half of the 16 subjects started with the original animals (low similarity), followed by medium and high similarity levels, while the remaining subjects were tested in the opposite order. Each part of the experiment consisted of 192 trials, separated into two blocks, and lasted about 15 minutes. Between successive parts, the subjects were offered a short break. Individual trials followed exactly the same design as in experiment 1. Except for the first author, none of the subjects in this experiment had participated in experiment 1.

3.2 Results

Trials with RTs longer than 3 s (1.2%) were discarded prior to any further analysis. The mean RT was 548 ms; the correct response rates ranged from 61.1% to 91.5% (mean 78.6%).

The mean accuracy results are shown in Figure 4, which suggests a decrement in the mean correct rate and a strengthening of the effects of translation with increased similarity. A three-way ANOVA (TRANSLATION × IDENTITY × SIMILARITY) indicated that similarity of the animals strongly affected performance (F[2,360]=52.8; $p < 0.001$). Not surprisingly, performance was the best when animal shapes were the least similar to each other (dissimilarity factor $f = 1.0$). As in experiment 1, TRANSLATION and IDENTITY had no significant main effects (F< 1), but interacted strongly with each other (F[3,360]=15.1; $p < 0.001$). Other interactions were n.s.

Separate 2-way ANOVAs (TRANSLATION × SIMILARITY by IDENTITY) revealed, for IDENTITY=*different*, a significant effect of TRANSLATION only in the high-similarity condition (F[3,60]=3.4, $p < 0.02$). For IDENTITY=*same*, the effect of TRANSLATION was significant at all three similarity levels (for $f = 0.4$, or high similarity: F[3,60]=4.9, $p < 0.004$; for $f = 0.7$: F[3,60]=5.3, $p < 0.0026$; for $f = 1$: F[3,60]=4.5, $p < 0.0065$). These results were confirmed by post-hoc contrast analysis.

Although translation had strong effects on correct rate in this experiment, the effects were opposite for *same* and for *different* trials. This conclusion is consistent with the observed pattern of d' values. A General Linear Models analysis (used instead of regular ANOVA because of the presence of unbalanced cells in the d' data) revealed only one significant effect, that of SIMILARITY: F[2,163]=36.0, $p < .0001$. The mean d' values were 1.1 for $f = 0.4$, 1.9 for $f = 0.7$, and 2.2 for $f = 1.0$. SIMILARITY was also the only significant effect for RT (F[2,360]= 3.9, $p < 0.0215$).

As noted above, we had separated our pool of subjects into two, to control for possible serial adaptation effects: eight observers proceeded from easy to difficult tasks, and the other eight were tested in the opposite order. The effects described above for the complete data set were identical for the two subgroups.

3.3 Discussion

Even more than in experiment 1, the effect of translation in *same* trials in experiment 2 were offset by nearly opposite effect in *different* trials. Increasing the similarity among the stimuli made both these effects. This result is different from the observations made by Dill and Fahle (1997a), who found that positional specificity increased with a decrease in the discriminability of random dot clouds and checkerboard patterns. In this sense, the recognition of novel, complex patterns seems to be qualitatively different from the recognition of more familiar objects such as our animal-like stimuli, regardless of the similarity of the latter to each other.

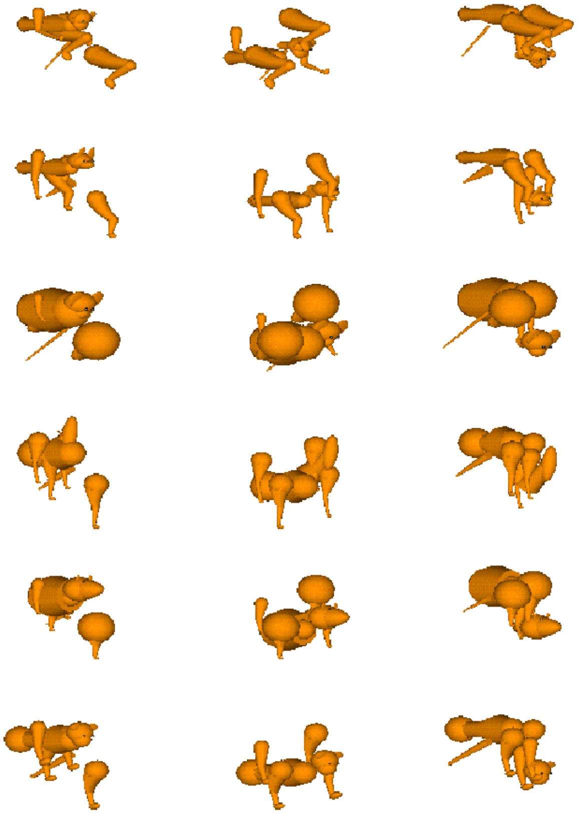

Figure 5: Examples of scrambled animals used in experiments 3 and 4. Each column shows different basic shapes, scrambled in the same manner; each row corresponds to the same basic shape, scrambled in different manners.

4 Experiment 3: "scrambled" animals, local feature cues

One major difference between our first two experiments and the earlier studies with complex random patterns (Foster and Kahn, 1985; Dill and Fahle, 1997a) was the general prior familiarity of the subjects with animal-like shapes. Although our computer-rendered shapes were not naturalistic copies of real animals, subjects readily named the animals when being introduced to the experiment and the stimuli. Experiment 3 was designed to test whether the familiarity of the objects, i.e., their resemblance to already experienced real or toy animals, leads to robust translation invariance that is not observed for novel patterns. To reduce familiarity and still be able to compare results directly with the above two experiments, we rendered the six animals as sets of "limbs," while randomizing the location of limbs relative to each other. This produced "scrambled" animals that contained the same "local" features (limbs) as the original ones, but did not form a meaningful object (see Figure 5). Additionally, since the configuration of the limbs could be changed from trial to trial, repetition of the stimuli and the possible concomitant learning effects were avoided.

4.1 Method

The same apparatus and stimulus conditions as in experiment 1 were used. Scrambled animals were designed from the same set of limbs as the animal models in experiment 1. Instead of composing the seven parts (head, body, two forelegs, two hind legs, tail) into complete 3D animal objects, each one was translated by small random amounts in three mutually orthogonal directions. In different trials, the second scrambled animal differed from the first one parametrically, in the *shapes* of its parts. The random scrambling, however, was the same for both animals: homologous parts (e.g., the heads) were shifted by the same amount in both stimuli. Thus, the subjects could base their discrimination decision on local shape cues. For each trial, the displacement of part types was newly randomized. The design of the individual trials, presentation times, masking, etc., were exactly as in experiment 1. Eight subjects were tested in three blocks of 96 trials each. All but two subjects had not participated in experiments 1 or 2.

4.2 Results and discussion

Trials with RTs longer than 3 s (0.47%) were discarded prior to any further analysis. The mean RT was 485 ms; the correct response rates ranged from 62.8% to 80.6% (mean 71.4%).

A two-way ANOVA (TRANSLATION × IDENTITY) revealed, as before, a strong interaction between these two variables (F[3,56]=8.3, $p < 0.0001$), and a main effect of IDENTITY (F[1,56]=20.6, $p < 0.0001$); the main effect of TRANSLATION was n.s. As can be seen in Figure 6, this stemmed from opposing tendencies at the TRANSLATION=*control* level, where in *different* trials the performance was at chance, whereas in *same* trials it was as high as 86% (in the three translation conditions, correct rates were flat and nearly the same in *same* and *different* trials).

The difficulty of the task involving scrambled animal shapes is indicated by the low mean d' value of 1.2 (*diagonal*: 1.1; *control*: 1.2; *horizontal*: 1.3; *vertical*: 1.4). The effect of TRANSLATION on d' was n.s. The RT data also showed no significant effects. As in the other experiments, there

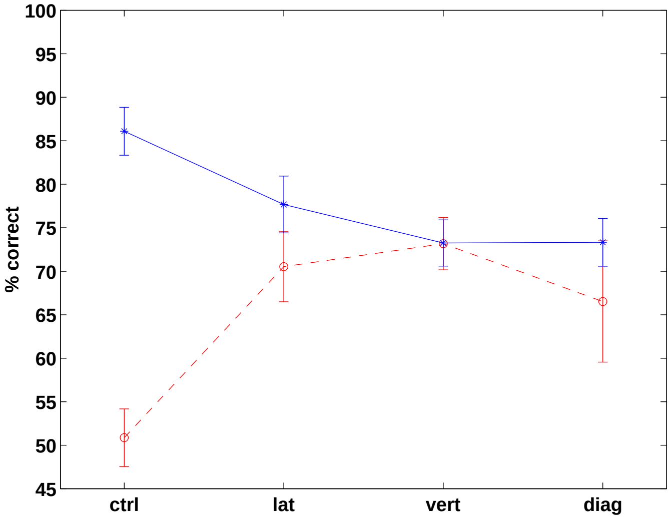

Figure 6: correct response rates by condition in experiment 3 (scrambled animals, local feature cues). Solid line: *same* trials; dashed line: *different* trials. The points show mean correct rates ±1 standard error ($n = 8$) for the four Translation conditions (control, lateral, vertical and diagonal).

was no indication of a speed-accuracy tradeoff.

The results of experiment 3 suggest that the meaningful shape of the animal-like stimuli was unlikely to have been responsible for the pattern of performance found in experiments 1 and 2 (namely, a decrease of correct rate with translation only in *same* trials, and no effect on d'). The scrambling of the animal-like shapes did not, by itself, result in positional specificity. In the next section, we introduce a simple variation on the scrambling method that does lead to a significant overall effect of translation.

5 Experiment 4: “scrambled” animals, global configuration cues

Both the identities of local features of an object and their spatial relations can help discriminate it from other objects. In experiment 3, the spatial relations among the parts were identical for the two stimuli in any given trial; the pair only differed in the shapes of the parts employed. In experiment 4, we created a complementary situation: both scrambled animals in a given trial were now composed of identically shaped parts, and only differed in their spatial arrangement. If, for example, the first stimulus was a particular scrambled monkey, then the second stimulus was a differently scrambled monkey (cf. the *rows* in Figure 5). In comparison, in experiment 3, the second object would have been, for example, a scrambled dog or mouse (cf. the *columns* in Figure 5). Both experiments, therefore, employed the same type of scrambled objects, but separated the effects of local features (part shapes) from those of their global layout (part configuration).

5.1 Method

Experiment 4 involved exactly the same experimental procedure as experiment 3, including the same kind of scrambled animals. However, unlike in experiment 3, the stimuli in each trial always consisted of the same set of parts, scrambled in two different manners. Experiment 4 has been carried out at two separate locations, with two distinct subject populations, as detailed below. Experiment 4a, conducted at MIT, involved nine subjects, three of whom were new to this experiment series; the remaining six had already participated in this study. Experiment 4b, conducted at Cornell University,[5] involved 13 subjects, all undergraduates enrolled in a summer course. The data from these two experiments are presented separately in the appendix; an analysis of the pooled data appears next.

5.2 Results

Data from the 16 subjects whose correct rate exceeded 55% were included in this analysis. Trials with RTs longer than 3 *s* (1.0%) were omitted from further analysis. The mean RT was 494 *ms*; the correct response rates for the seven subjects ranged from 58.0% to 81.5% (mean 69.9%).

A two-way ANOVA (TRANSLATION × IDENTITY) yielded a marginal main effect of TRANSLATION (correct rate dropping from 73.7% for *control* to 66.7% for *diagonal*; F[3,120]=2.5, $p < 0.06$); a main effect of IDENTITY (correct rate of 67.7% for *different* vs. 72.1% for *same*; F[1,120]=5.6, $p < 0.02$), and the by now familiar very strong interaction (F[3,120]=21.8, $p < 0.0001$). Figure 7 shows a pattern similar to that of experiments 4a and 4b (cf. Figure 10 in the appendix): in *same* trials correct rate in the *control* condition was better than in the three translation conditions, with a relatively uniform performance the across the four conditions in *different* trials.

[5] In response to the comments of a reviewer, who pointed out the importance of a replication of the results of experiment 4a with naive subjects.

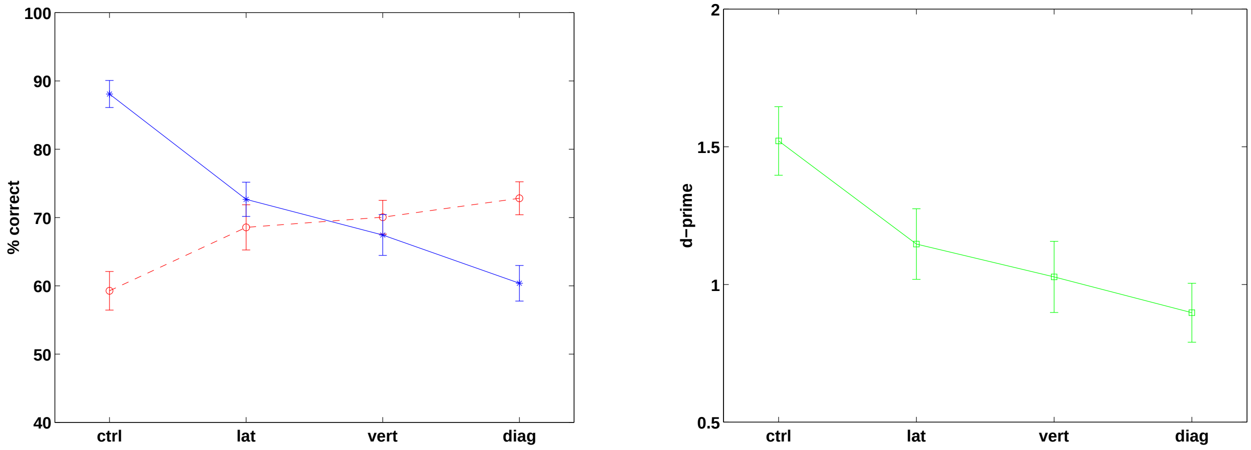

Figure 7: experiment 4 (scrambled animals, configuration cues). *Left:* correct rates. Solid line: *same* trials; dashed line: *different* trials. *Right:* d'. The points show mean correct rates ±1 standard error ($n = 16$) for the four TRANSLATION conditions (control, lateral, vertical and diagonal).

The mean d' in experiment 4 was 1.2. The mean for the *diagonal* condition was 0.9, for *horizontal* 1.0, for *vertical* 1.1, and for *control* 1.5. Importantly, the effect of TRANSLATION on d' estimated by ANOVA was significant (F[3,60]=4.8, $p < 0.005$). Post-hoc contrasts between *control* and the other conditions, which were all significant, confirmed this outcome. The RT data showed no significant effects. As in the other experiments, there was no indication of a speed-accuracy tradeoff.

5.3 Discussion

The seemingly slight modification of the task between experiments 3 and 4 — from discrimination by local features to discrimination by their spatial configuration — produced a considerable difference in the results. In experiment 4, the occurrence of a particular part was not diagnostic for discrimination, unless via a chance occlusion. Unlike the discrimination by local features in experiment 3, the performance based on configurational (structural) cues was not completely invariant to translation.

It is tempting to attribute this distinction to two different subsystems (or stages) of object vision: one that is translation invariant and allows recognition of local features and one that is at least partially position-specific and is responsible for the processing of configurational or structural information. Note that achieving translation invariant recognition of a particular stimulus feature implies downplaying its position in the visual field. To be able to discriminate objects solely on the basis of the spatial relations of some simpler features, the system may have to rely on evidence from mechanisms that are not fully shift-invariant. We shall return to this point in the general discussion.

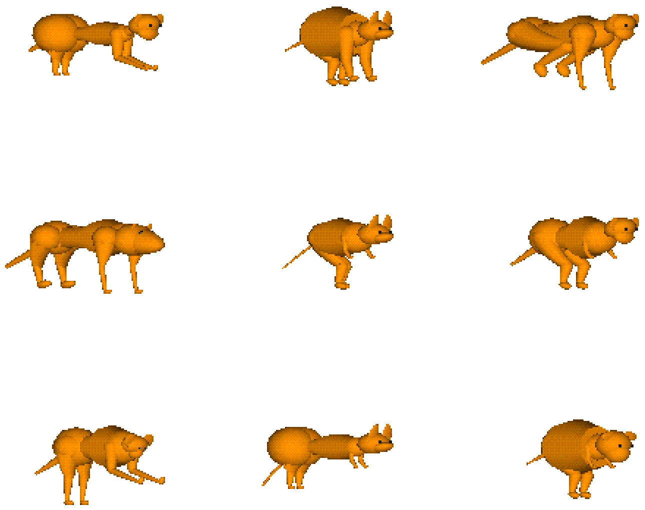

Figure 8: chimera shapes used in experiment 5. Each object contains parts taken from various animals.

6 Experiment 5: chimerae

To explore further the distinction between local and global or structural shape representation, in experiment 5 we used a new class of chimeric objects by randomly combining parts from different animals (see Figure 8). Aside from random similarity with “regular” animals, these chimerae were difficult to categorize into familiar animal classes. Another major difference compared to experiments 1 and 2 was that new chimerae could be created for each new trial, thereby avoiding the development of a classification scheme by the subject. Note that the identification of a particular feature (e.g., the head) in two chimerae does not necessarily indicate that they are identical, because all the other features may still be different. Subjects, therefore, were forced to attend to the entire configuration of each chimera.

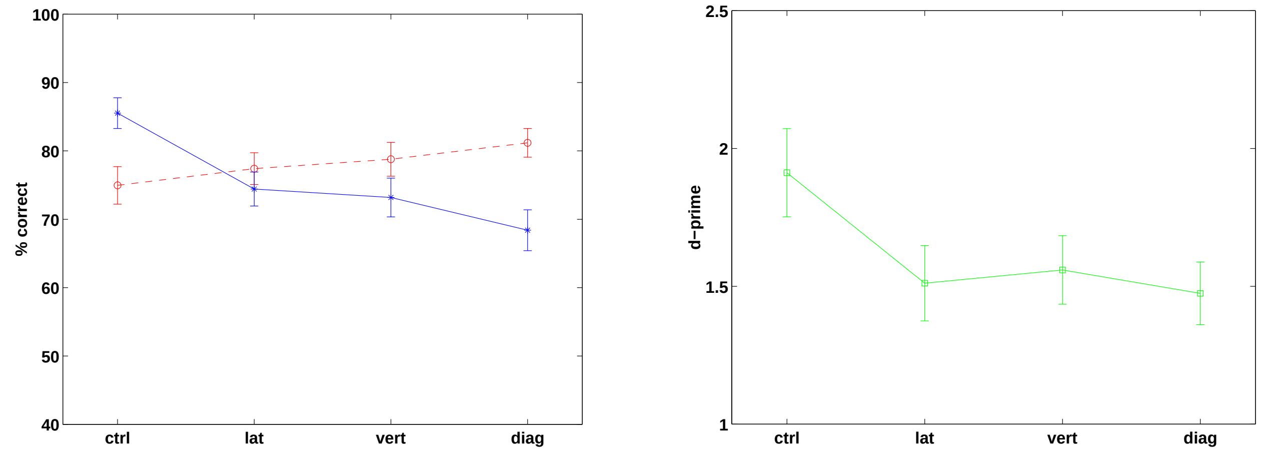

Figure 9: experiment 5 (chimera shapes). *Left:* correct rates. Solid line: *same* trials; dashed line: *different* trials. *Right:* d'. The points show mean correct rates ±1 standard error ($n = 22$) for the four TRANSLATION conditions (control, lateral, vertical and diagonal).

6.1 Method

Experiment 5 followed the same basic design as experiment 1. The single difference between the two experiments was that while in experiment 1 only the original set of six animals was used, random mixtures of the original models were composed for experiment 5. Each chimera was produced by randomly choosing four components (head, body and tail, forelegs, hind legs) each from one of the six animals. For example, a stimulus could consist of the head of the tiger, body and tail of the monkey, forelegs of the mouse, and hind legs of the horse. In each trial, new components were chosen at random. In *different* trials, both chimerae were randomly different.

Experiment 5, like the previous one, has been carried out at two separate locations, with two distinct subject populations. Experiment 5a, conducted at MIT, involved eight subjects. Experiment 5b, conducted at Cornell University, involved 14 undergraduates enrolled in a summer course. As before, the data from these two experiments are presented separately in the appendix. An analysis of the pooled data from the 22 subjects appears next.

6.2 Results

Trials with RT longer than 3 s (1.8%) were discarded prior to further analysis. The mean RT was 486 ms; the correct response rates ranged from 64.1% to 87.8% (mean 76.7%).

A two-way ANOVA (TRANSLATION × IDENTITY) showed only the interaction as significant (F[3,168]=7.4, $p < .0001$). Figure 9 reveals a decrease in correct rate with translation in *same* trials (and a slight increase in *different* trials). As before, we conducted separate ANOVAs by IDENTITY; for *different* trials, the effect of TRANSLATION was n.s. ($p = 0.3$); in comparison, an ANOVA for IDENTITY=*same* resulted in a highly significant effect of TRANSLATION (F[3,84]=7.5, $p < 0.0002$).

The mean d' in experiment 5 was 1.6. The mean for the *diagonal* condition was 1.5, for *horizontal* 1.6, for *vertical* 1.5, and for *control* 1.9. The effect of TRANSLATION on d' was marginal (F[3,84]=2.2, $p < 0.09$); post-hoc contrasts also showed differences between *control* and the other conditions (at $p < 0.06$ or better). The RT data showed no significant effects. As in the other experiments, there was no indication of a speed-accuracy tradeoff.

7 General discussion

The present study examined the degree of translation invariance in *same/different* discrimination of complex 3D objects. Our major findings can be summarized as follows. First, translation invariance is more likely to hold in trials where the correct response is *different*, compared to *same* trials. In other words, it is more difficult to label two objects as same when they are spatially offset; labeling objects as different depends less on the spatial displacement. Second, translation invariance holds for 3D stimuli (both familiar and unfamiliar) that can be discriminated on the basis of local shape information, but not for stimuli whose only distinguishing cues are configurational or structural.

Rather than confirming or refuting *in toto* earlier reports of translation invariance (Biederman and Cooper, 1991; Bricolo and Bülthoff, 1992; Bricolo, 1996) or of position specificity (Foster and Kahn, 1985; Nazir and O'Regan, 1990; Cave et al., 1994; Dill and Fahle, 1997a; Dill and Fahle, 1997b; Larsen and Bundesen, 1998), our results provide a certain insight into the mechanism that underlies the comparison of objects related through translation. Apparently, this mechanism can produce behavior that is more invariant or less so, depending on the conditions. The pattern of results of experiments 1 through 5 suggests that this mechanism treats local cues and configurational or structural information differentially.

It is difficult to reconcile this conclusion either with the structural theories of object representation, which predict complete invariance to translation (Biederman, 1987), or with the holistic appearance-based theories, which predict imperfect invariance across the board (Edelman, 1995b). There does exist, however, a computational approach to object representation that embodies a distinction between local features and structural information analogous to the one that emerges from our data.

In computer vision, this approach takes the form of appearance-based methods modified to treat structure explicitly. For example, Burl et al. (1998) combine "local photometry" (features that are basically templates for small snippets of images) with "global geometry" (a probabilistic quantification of spatial relations between pairs or triplets of features). Likewise, Camps et al. (1998) represent objects in terms of appearance-based "parts"[6] and their approximate relations. In both these methods, recognition and categorization are based on an interplay of local shape cues and approximate location information. Such hybrid methods constitute an attractive alternative to holistic appearance-based models, if only because they may eventually meet the systematicity[7] challenge for shape representation, without opting for the problematic structural approach (Edelman, 1999; Edelman and Intrator, 2000).

In the present context, a separate treatment of local and structural cues may explain why the former, but not the latter, support translation-invariant processing (cf. our experiments 1 through 3 on the one hand, and experiments 4 and 5 on the other hand). Suppose that "units" selectively responding to local cues (according to the models mentioned above, such cues can be as simple as random snippets of images) are replicated throughout the visual field, and, moreover, that the receptive field of each such unit is spatially localized (rather than extending over all of the central visual field). The response of an ensemble of such units would signal "where" in addition to "what" the stimulus components are, and would, therefore, carry information sufficient for recognition and categorization, as well as for other tasks that may require explicit representation of shape structure (Edelman and Intrator, 2000). Importantly, in such a system structure would be represented in a distributed fashion; unlike local features (presumably replicated all over the visual field), it would not, therefore, be amenable to translation-invariant priming — precisely what the present study

[6] Actually, projections of image fragments onto the principal components of stacks of such fragments.

[7] The problem of systematicity (Fodor, 1998) refers to the need to represent and manipulate structure explicitly, so that making sense of an object composed of, say, a cube positioned above a sphere would entail an equally successful processing of a sphere above a cube (Hummel, 2000).

found.

Encouragingly, neuronal mechanisms corresponding functionally to the shape-tuned "units" have been described by a number of groups, e.g., (Fujita et al., 1992; Logothetis et al., 1995); see the reviews in (Logothetis and Sheinberg, 1996; Rolls, 1996; Tanaka, 1996). In line with the human psychophysics, most of the shape-tuned cells in the monkey respond selectively to some particular views of an object, and nearly equally to a range of stimulus sizes and locations (Tovee et al., 1994; Ito et al., 1995; Logothetis et al., 1995). Finally, the "what+where" cells needed specifically for implementing the structure representation scheme outlined above (Edelman and Intrator, 2000) have also been found, in cortical areas V4 and posterior IT (Kobatake and Tanaka, 1994; Ito et al., 1995), and in the prefrontal cortex (Rao et al., 1997; Rainer et al., 1998). A practical test of these intriguing parallels between psychophysics, neurobiology and computational considerations should be possible, once the "what+where" scheme is implemented as a working model.

Acknowledgments

This research was conducted at the Center for Biological and Computational Learning, at the Department of Brain and Cognitive Sciences, MIT, while SE was a visiting scientist there. M. Dill was supported by the Boehringer Ingelheim Fonds. We thank J. E. Cutting, M. Fahle, G. Geiger, T. Poggio, M. Riesenhuber, and P. Sinha for valuable discussions, B. P. Hiles for help with running the experiments at Cornell, and S. Kaufmann for bringing to our attention the 1954 paper by Wallach and Austin-Adams.

Appendix

Experiment 4a

Trials with RTs longer than 3 s (0.88%) were discarded prior to any further analysis. The mean RT was 486 ms; the correct response rates ranged from 58.0% to 79.5% (mean 69.5%).

A two-way ANOVA (TRANSLATION × IDENTITY) revealed a significant main effect of IDENTITY (F[1,64]=4.8, $p < 0.03$), no main effect of TRANSLATION, and a strong interaction (F[3,64]=11.1, $p < .0001$). Figure 10, left, shows that in *same* trials correct rate in the *control* condition was better than in the three translation conditions, while in *different* trials the performance was relatively

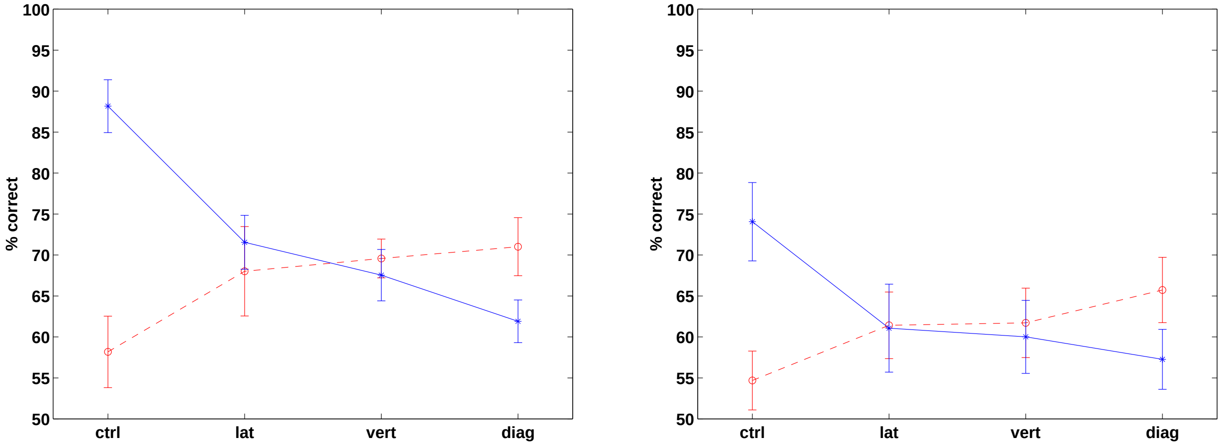

Figure 10: correct response rates by condition in experiment 4a (left) and 4b (right). Solid line: *same* trials; dashed line: *different* trials. The points show mean correct rates ±1 standard error (left: $n = 9$; right: $n = 7$) for the four TRANSLATION conditions (control, lateral, vertical and diagonal).

uniform across the four conditions. In a separate ANOVA for IDENTITY=*different*, the effect of TRANSLATION did not reach significance (F[3,32]=2.0, $p = 0.13$); in comparison, an ANOVA for IDENTITY=*same* resulted in a very strong effect of TRANSLATION (F[3,32]=13.5, $p < 0.0001$).

The mean d' in experiment 4a was 1.13 (nearly as low as in experiment 3). The mean for the *diagonal* condition was 1.13, for *horizontal* 0.99, for *vertical* 1.1, and for *control* 1.53. Importantly, and in contrast to experiment 3, the effect of TRANSLATION on d' was significant (F[3,32]= 3.1, $p < 0.04$). The RT data showed no significant effects. As in the other experiments, there was no indication of a speed-accuracy tradeoff.

Experiment 4b

Data from six out of the 13 subjects whose mean correct rate did not reach 55% were discarded. Data from the other seven subjects were processed as before. Trials with RTs longer than 3 s (1.2%) were omitted from further analysis. The mean RT was 695 ms; the correct response rates for the seven subjects ranged from 60.0% to 81.5% (mean 62.0%).

A two-way ANOVA (TRANSLATION × IDENTITY) showed only the interaction to be significant (F[3,96]=3.9, $p < 0.01$). Figure 10, right shows a pattern similar to that of experiment 4a (cf.

Figure 10, left): in *same* trials correct rate in the *control* condition was better than in the three translation conditions, with a relatively uniform performance the across the four conditions in *different* trials. Also as in experiment 4a, a separate ANOVA for Identity=*different* yielded no effect of Translation ($p > 0.1$), while an ANOVA for Identity=*same* did show a marginal effect of Translation (F[3,48]=2.6, $p < 0.06$).

The mean d' in experiment 4b was 1.18. The mean for the *diagonal* condition was 0.92, for *horizontal* 1.1, for *vertical* 1.2, and for *control* 1.51. Although the effect of Translation on d' estimated by ANOVA did not reach significance ($p > 0.23$), a post-hoc contrast between *control* and *diagonal* conditions was significant at a $p < 0.05$ level. The RT data showed no significant effects. As in the other experiments, there was no indication of a speed-accuracy tradeoff.

Experiment 5a

Eight observers participated in experiment 5a, which consisted of three blocks of 96 trials each. Experiments 4a and 5a were run with the same subjects on a single day, separated by a $5 - 10$ *min* break. Four of the subjects were tested with scrambled animals first, and another four started with the chimerae (one of the subjects participated only in experiment 4). No difference between the two groups could be detected in the results.

The single trial with RT longer than 3 *s* was discarded prior to further analysis. The mean RT was 396 *ms*; the correct response rates ranged from 67.7% to 87.8% (mean 78.8%).

A two-way ANOVA (Translation × Identity) revealed a significant main effect of Identity (F[1,56]=12.9, $p < 0.0007$), no main effect of Translation, and a significant interaction (F[3,56]=4.1, $p < .01$). Figure 11, left shows that, as in experiment 4, in *same* trials correct rate in the *control* condition was better than in the three translation conditions; in *different* trials the performance was quite uniform across the four conditions. In a separate ANOVA for Identity=*different*, the effect of Translation did not reach significance (F< 1); in comparison, an ANOVA for Identity=*same* resulted in a strong effect of Translation (F[3,28]=4.9, $p < 0.008$).

The mean d' in experiment 5a was 1.8. The mean for the *diagonal* condition was 1.6, for *horizontal* 1.8, for *vertical* 1.6, and for *control* 2.1. The effect of Translation on d' was n.s. ($p = 0.3$), but post-hoc contrasts did show a marginal difference between *control* and the other conditions ($p < 0.09$). The RT data showed no significant effects. As in the other experiments,

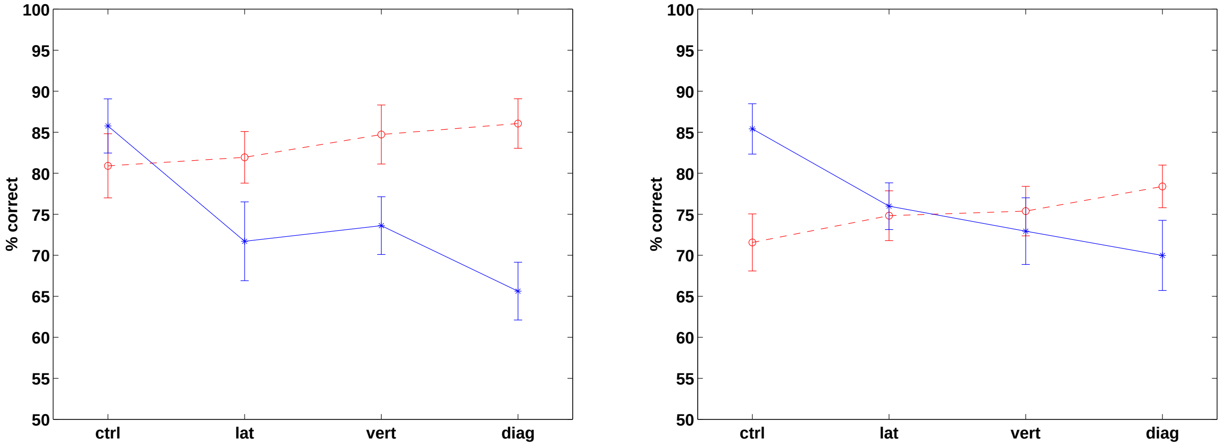

Figure 11: correct response rates by condition in experiment 5a (left) and 5b (right). Solid line: *same* trials; dashed line: *different* trials. The points show mean correct rates ±1 standard error (left: $n = 8$; right: $n = 14$) for the four TRANSLATION conditions (control, lateral, vertical and diagonal).

there was no indication of a speed-accuracy tradeoff.

Experiment 5b

Fourteen observers participated in experiment 5a, which consisted of three blocks of 96 trials each. Trials with RT longer than 3 *s* (2.7%) were discarded prior to further analysis. The mean RT was 537 *ms*; the correct response rates ranged from 64.1% to 86.7% (mean 75.6%).

A two-way ANOVA (TRANSLATION × IDENTITY) showed only the interaction as significant (F[3,104]=4.0, $p < 0.01$). Figure 11, right reveals the same pattern as in experiment 5a, albeit with larger standard errors. As before, we conducted separate ANOVAs by IDENTITY; for *different* trials, the effect of TRANSLATION was n.s. (F< 1); in comparison, an ANOVA for IDENTITY=*same* resulted in a significant effect of TRANSLATION (F[3,52]=3.4, $p < 0.02$).

The mean d' in experiment 5b was 1.5. The mean for the *diagonal* condition was 1.4, for *horizontal* 1.4, for *vertical* 1.5, and for *control* 1.8. The effect of TRANSLATION on d' was n.s. ($p = 0.3$), but post-hoc contrasts did show a marginal difference between *control* and the other conditions ($p < 0.07$), as in experiment 5a. The RT data showed no significant effects. As in the other experiments, there was no indication of a speed-accuracy tradeoff.

On pooling the data

A visual comparison between the results of experiments 4a and 4b (Figure 10, left and right) and between those of 5a and 5b (Figure 11, left and right) reveals a qualitative similarity between the performance patterns of the different groups of subjects involved in the original and repeated experiments. Specifically, while the performance in *different* trials (dashed lines) showed little dependence on *Translation*, the percentage of correct responses in *same* trials (solid lines) decreased with *Translation*. This visual impression is supported by quantitative analyses: the ANOVA produced the same significant *Identity-Translation* interaction in experiment 4a as in 4b, and again in 5a as in 5b. Thus, the pooling of the data that we reported in sections 5 and 6 is perfectly reasonable.

References

Bar, M. and Biederman, I. (1998). Subliminal visual priming. *Psychological Science*, 9(6):464–469.

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

Biederman, I. and Cooper, E. E. (1991). Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, 20:585–593.

Bricolo, E. (1996). *On the Representation of Novel Objects: Human Psychophysics, Monkey Physiology and Computational Models.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Bricolo, E. and Bülthoff, H. H. (1992). Translation-invariant features for object recognition. *Perception*, 21 (supp.2):59.

Bricolo, E. and Bülthoff, H. H. (1993). Rotation, translation, size and illumination invariances in 3D object recognition. *Invest. Ophthalm. Vis. Science*, 34(4):1081.

Bülthoff, H. H. and Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science*, 89:60–64.

Burl, M. C., Weber, M., and Perona, P. (1998). A probabilistic approach to object recognition using local photometry and global geometry. In *Proc. 4^{th} Europ. Conf. Comput. Vision, H. Burkhardt and B. Neumann (Eds.), LNCS-Series Vol. 1406–1407, Springer-Verlag*, pages 628–641.

Camps, O. I., Huang, C.-Y., and Kanungo, T. (1998). Hierarchical organization of appearance-based parts and relations for object recognition. In *Proc. ICCV*, pages 685–691. IEEE.

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

Dill, M. and Fahle, M. (1997a). Limited translation invariance of human visual pattern recognition. *Perception & Psychophysics*, 60:65–81.

Dill, M. and Fahle, M. (1997b). The role of visual field position in pattern-discrimination learning. *Proceedings of the Royal Society of London B*, 264:1031–1036.

Edelman, S. (1995a). Class similarity and viewpoint invariance in the recognition of 3D objects. *Biological Cybernetics*, 72:207–220.

Edelman, S. (1995b). Representation, Similarity, and the Chorus of Prototypes. *Minds and Machines*, 5:45–68.

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

Edelman, S. and Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, 32:2385–2400.

Edelman, S. and Intrator, N. (2000). (Coarse Coding of Shape Fragments) + (Retinotopy) ≈ Representation of Structure. *Spatial Vision*, 13:255–264.

Fodor, J. A. (1998). *Concepts: where cognitive science went wrong.* Clarendon Press, Oxford.

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

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

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

Hummel, J. E. (2000). Where view-based theories of human object recognition break down: the role of structure in human shape perception. In Dietrich, E. and Markman, A., editors, *Cognitive Dynamics: conceptual change in humans and machines*, chapter 7. Erlbaum, Hillsdale, NJ.

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.*, 73:218–226.

Jolicoeur, P. and Humphrey, G. K. (1998). Perception of rotated two-dimensional and three-dimensional objects and visual shapes. In Walsh, V. and Kulikowski, J., editors, *Perceptual constancies*, chapter 10, pages 69–123. Cambridge University Press, Cambridge, UK.

Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.*, 71:856–867.

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

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

Logothetis, N. K. and Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, 19:577–621.

Macmillan, N. A. and Creelman, C. D. (1991). *Detection theory: a user's guide.* Cambridge University Press, Cambridge.

Macmillan, N. A., Kaplan, H. L., and Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychological Review*, 84:452–471.

Nazir, T. and O'Regan, J. K. (1990). Some results on translation invariance in the human visual system. *Spatial vision*, 5:81–100.

Neisser, U. (1967). *Cognitive Psychology.* Appleton-Century-Crofts, New York, NY.

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics*, 16:283–290.

Rainer, G., Asaad, W., and Miller, E. K. (1998). Memory fields of neurons in the primate prefrontal cortex. *Proceedings of the National Academy of Science*, 95:15008–15013.

Rao, S. C., Rainer, G., and Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, 276:821–824.

Rolls, E. T. (1996). Visual processing in the temporal lobe for invariant object recognition. In Torre, V. and Conti, T., editors, *Neurobiology*, pages 325–353. Plenum Press, New York.

SAS (1989). *User's Guide, Version 6.* SAS Institute Inc., Cary, NC.

Saslow, M. G. (1967). Latency for saccadic eye movement. *Journal of the Optical Society of America*, 57:1030–1036.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–139.

Tovee, M. J., Rolls, E. T., and Azzopardi, P. (1994). Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert monkey. *J. of Neurophysiology*, 72:1049–1060.

Wallach, H. and Austin-Adams, P. (1954). Recognition and the localization of visual traces. *American Journal of Psychology*, 67:338–340.