MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1610 June, 1997
C.B.C.L. Paper No. 150

Translation invariance in object recognition, and its relation to other visual transformations

Marcus Dill and Shimon Edelman

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1500-1999/AIM-1610.ps.Z

Abstract

Human object recognition is generally considered to tolerate changes of the stimulus position in the visual field. A number of recent studies, however, have cast doubt on the completeness of translation invariance. In a new series of experiments we tried to investigate whether positional specificity of short-term memory is a general property of visual perception. We tested *same-different* discrimination of computer graphics models that were displayed at the same or at different locations of the visual field, and found complete translation invariance, regardless of the similarity of the animals and irrespective of direction and size of the displacement (Exp. 1 and 2). Decisions were strongly biased towards *same* decisions if stimuli appeared at a constant location, while after translation subjects displayed a tendency towards *different* decisions. Even if the spatial order of animal limbs was randomized ("scrambled animals"), no deteriorating effect of shifts in the field of view could be detected (Exp. 3). However, if the influence of single features was reduced (Exp. 4 and 5) small but significant effects of translation could be obtained. Under conditions that do not reveal an influence of translation, rotation in depth strongly interferes with recognition (Exp. 6). Changes of stimulus size did not reduce performance (Exp. 7). Tolerance to these object transformations seems to rely on different brain mechanisms, with translation and scale invariance being achieved in principle, while rotation invariance is not.

This report describes research done at the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences at the Massachusetts Institute of Technology.

1 Introduction

When trying to recognize an object, our brain is confronted with the problem that the projection of this object on the retina can vary considerably between different instances. Among the numerous transformations our visual system has to cope with, tolerance to translation in the visual field has been often considered as the least problematic. The finding that lower animals such as flies do not exhibit position-invariant processing (Dill et al. 1993; Dill & Heisenberg 1995), and the difficulty of simpler neural networks such as the Perceptron to learn translation invariance (Minsky & Pappert 1969), have cast doubt on this simplicity assumption.[1]

A number of recent studies (Foster & Kahn 1985; Nazir & O'Regan 1990; Dill & Fahle 1997a; Dill & Fahle 1997b) have shown that in humans recognition of novel complex stimuli is not completely translation invariant. If, for example, subjects have to discriminate whether two sequentially flashed random-dot clouds are *same* or *different*, decisions are faster and more frequently correct when both stimuli are presented to the same rather than to different locations in the visual field (Foster & Kahn 1985; Dill & 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 rule out explanations in terms of afterimages, eye movements, and shifts of spatial attention (Dill & Fahle 1997a).

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 target 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. Recently, Dill and Fahle (1997b) have isolated two components of training performance. 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 succeeded to show 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 – perceptual learning – is much more specific to the location of training.

The basic constraint on the stimuli in psychophysical studies of invariant recognition 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, the typical stimuli both in *same-different* matching and in learning studies tended to be highly unnatural and complex. With the employment of more familiar patterns, one might suspect performance to prove to be insensitive to retinal shifts. Indeed, priming experiments with more natural stimulus types showed complete invariance. For instance, 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 largely 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 Humphreys (1997), the visual part of the priming effect may be too small to detect an influence of position, size or other transformations.

1.1 The Present Study

The main purpose of the present study was to investigate whether partial positional specificity is a general characteristic of visual recognition processes, or is peculiar to complex and unnatural stimulus types that are difficult to process and store. For that purpose, we adapted computer graphics stimuli (cf. Figure 1) that have been used in the past to investigate the influence of rotation in depth on object recognition (Edelman 1995). In a series of experiments, Edelman (1995) showed that changing the orientation of these 3D objects relative to the observer severely reduces their recognition rate. The impact of rotation could be detected both in training and in *same-different* matching experiments, and was more pronounced for similar than for easily discriminable objects. The latter result is consistent with the finding that positional specificity in discrimination of random patterns is influenced by the similarity between stimuli, as mentioned above (Dill & Fahle 1997a).

2 Experiment1: Discrimination of animal objects

In our first experiment we tested positional specificity of *same-different* discrimination among six computer-graphics animal-like shapes (the left column in Figure 1). These stimuli were adapted from an earlier study (Edelman 1995) that had shown highly significant effects of changing the orientation in depth on the discrimination of these objects. 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 paid for their

[1] Viable neural network models for learning translation invariance started to emerge only fairly recently (Földiák 1991; O'Reilly & Johnson 1994).

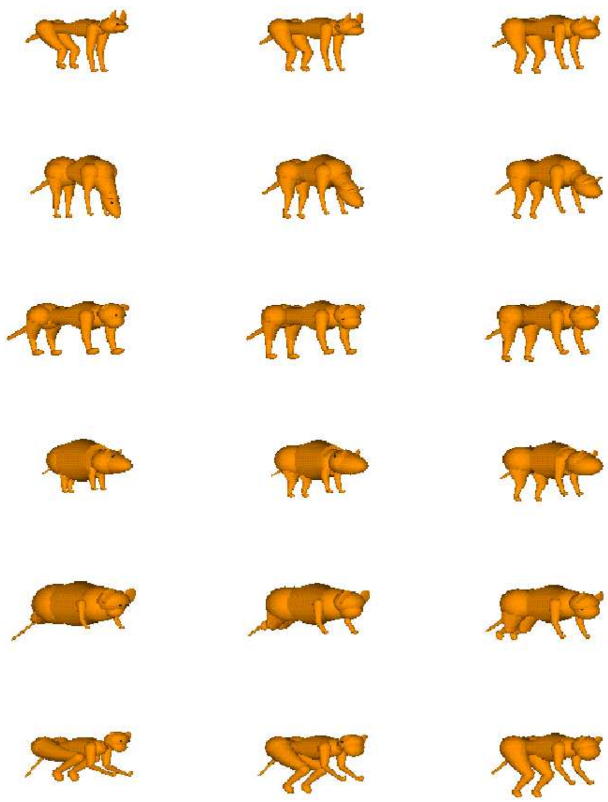

Figure 1: Three levels of similarity (columns) for the six animal-like computer-generated objects. The left column shows the original animals. The similarity between different animals, i.e. within one column, is increasingly larger in the middle and right column.

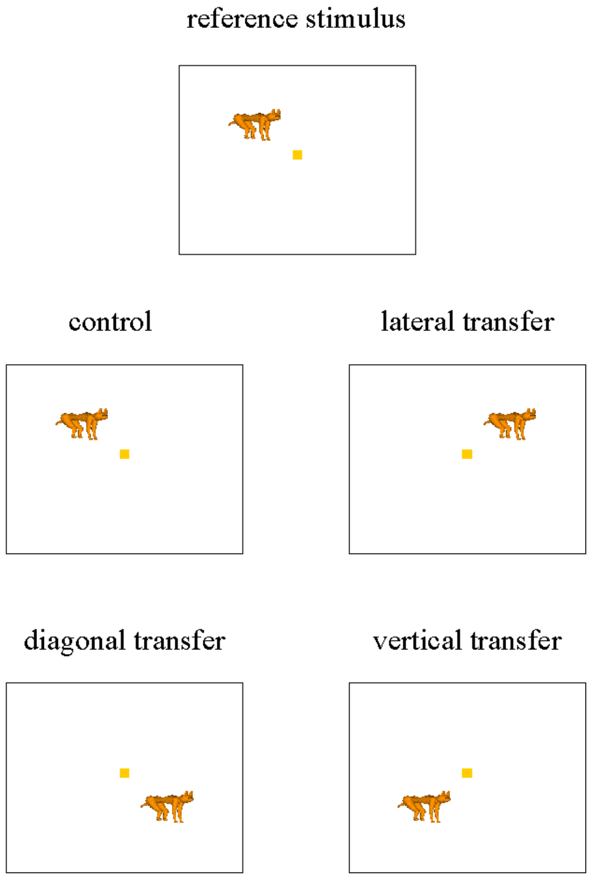

Figure 2: The four transfer conditions.

participation in one-hour sessions. Each observer had normal 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. Stimuli were produced and displayed on a Silicon Graphics workstation ($19''$ color monitor; refresh rate 120 Hz). The display was viewed binocularly at a distance of 0.6 *m*. The stimuli were 3D computer-graphics animal-like objects, adapted from an earlier study on object recognition (Edelman 1995). Each animal was defined by a set of 56 parameters representing characteristics such as length, diameter, and orientation of individual limbs. Six animal classes were used throughout the experiments; see Figure 1.[2] Stimulus images were about 3 deg wide and 2 deg 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 deg). The main body axis of an animal was always oriented in depth at 45 deg 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 only ca. 100 *msec*, 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 comprised 20 random cylinders each, and were presented simultaneously at the four possible stimulus locations, for 300 *msec*. Fixation was aided by a yellow spot of about 0.1 deg diameter at the middle of the screen. Decisions were communicated by pressing left (for "same") or right (for "different") mouse buttons. A computer beep provided negative feedback immediately after incorrect responses.

[2]The top and the bottom shapes in this figure are the original objects used in (Edelman 1995); the others are parametric variations, courtesy of T. Sugihara, RIKEN Institute.

Experimental Design. At the beginning of a trial, the fixation spot appeared for 1 *sec*, 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 deg, while the diagonal displacement was 8 deg. The onset asynchrony of the two animal stimuli was 1.4 *sec*. 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 *sec* interval.

Experiment 1 comprised 3 blocks of 96 trials each. 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 diagonal transfer), which were presented in a randomized order.

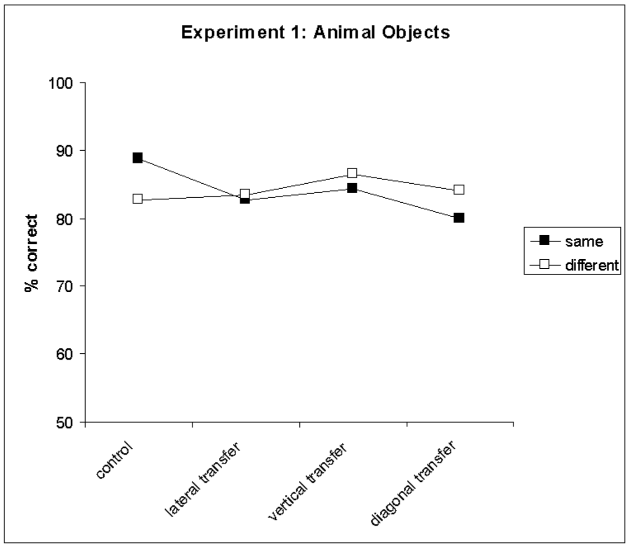

Figure 3: *Same-different* discrimination of animal objects.

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 *sec* were discarded prior to the calculation.

Figure 3 represents accuracy and RT averaged across the ten observers. For *same* trials, a 6% difference was observed between the control condition, i.e. when both animals are presented at the same location, and the mean of the three transfer conditions (88.7% compared to 82.4%). For *different* trials, however, accuracy without retinal shifts was even slightly below transfer results. The overall effect of translation, therefore, is small. These qualitative observations were confirmed by the results of two-way analysis of variance (ANOVA), testing the influence of TRANSLATION (control, lateral, vertical, diagonal) and IDENTITY (same, different). Both main factors did not contribute significantly to variance (TRANSLATION: $F[3,27]=1.93$; $p > 0.1$. IDENTITY: $F[1,9]=0.01$; $p > 0.1$). However, they interacted strongly ($F[3,27]=4.00$; $p < 0.1$), reflecting differential effects of transfer in *same* and *different* trials. Performance was relatively homogenous among the three transfer conditions. A separate ANOVA with only the three transfer conditions did not reveal any significant effect or interaction ($p > 0.1$). Clearly, the size and the direction of the displacement had no influence on performance.

RT data showed only small and non-significant effects ($p > 0.1$ for all main effects and interactions). In general, however, the RT tendencies are consistent with those in accuracy results. There was no indication of a speed-accuracy tradeoff.

2.3 Discussion

In *same-different* experiments with complex random patterns, significant effects of translation have been found repeatedly by different experimenters (Foster & Kahn 1985; Dill & Fahle 1997a). Similarly, rotating animal objects in depth had a strong impact on performance (Edelman 1995). We had, therefore, expected to find in Experiment 1 a clear deficit of transfer compared to control comparisons. Positional specificity, however, was only small. Only when the analysis was restricted to *same* trials did the effect of translation become significant. The slight opposite tendency in *different* trials largely cancels this effect. The overall influence of translation came out, therefore, as not significant, except that the decisions seemed to have been biased by the relative location of stimuli.

It should be noted that in many previous invariance studies, analysis have been restricted to *same* trials. Following a variety of arguments (e.g., that *different* trials do 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.

There are several possible reasons for the difference between our new results and the published findings of incomplete translation invariance for dot cloud and checkerboard stimuli. First, recognition of animal objects and identification of complex random patterns may involve different processes. Specifically, one may suspect that higher cognitive levels become involved when a meaningful object is detected. These levels may be less sensitive to the location of the object. Contributions from non-visual, i.e. conceptual or verbal, levels could have obscured positional specificity that would have been observed if subjects relied on visual processes only. For example, one may imagine a subject employing a labeling strategy in which each instance of an animal is stored as belonging to a certain abstract category. When comparing two animal stimuli, referral to the category labels might be sufficient, while the actual visual information is only used to access this verbal representation. Even within the visual processing and memory systems, a hierarchy of position-specific and translation-invariant stages may exist. For recognition of random patterns, contributions from lower levels may be stronger than for identification of meaningful objects.

A second explanation for nearly complete translation invariance of object discrimination could be that, although our computer graphics models did not look entirely natural, the class of real animal objects is familiar to all humans. Most likely, 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.

Finally, the task in Experiment 1 may have been too easy. Dill and Fahle (1997) report that increasing the similarity between stimuli leads to more pronounced positional specificity. Similarly, Edelman (1995) found that deteriorating effects of changes in orientation are larger for similar than for more distinct objects. The animal models may have been too easy to discriminate, to detect any significant effect of translation.

3 Experiment 2: Similarity and invariance

Experiment 2 investigated the influence of similarity between animal stimuli on positional specificity. 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 (1995) also found an interaction between similarity and invariance for animal-like shapes, we expected to detect 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" animal that was computed by averaging each of the 56 model parameters across the six animals. We tested each subject with all three levels of similarity (corresponding to the three columns in Figure 1). To avoid 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 (similar) stimulus set first.

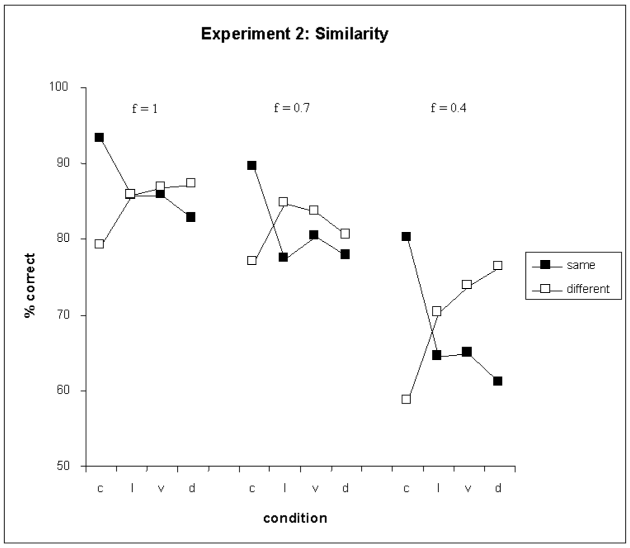

Figure 4: Similarity and invariance. The three plots correspond to the three levels of similarity among the six stimuli, as explained in section 3. *c*: control; *l*: lateral transfer; *v*: vertical transfer; *d*: diagonal transfer.

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 factor f. Three different similarity factors 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

The mean accuracy results are shown in Figure 4. Three-way ANOVAs (TRANSLATION $\times$ IDENTITY $\times$ SIMILARITY) indicated that similarity of the animals strongly affected performance (F[2,30]=49.44; $p <$ 0.001). Not surprisingly, performance was the best when animals were the least similar to each other. As in Experiment 1, TRANSLATION (F[3,45]=0.93; $p > 0.1$) and IDENTITY (F[1,15]=0.01; $p > 0.1$) had no significant main effect, but interacted strongly with each other (F[3,45]=23.45; $p < 0.001$). SIMILARITY, however, did not interact with TRANSLATION (F[6,90]=0.31; $p > 0.1$), indicating that increasing the similarity does not increase positional specificity. Even for the most difficult condition with nearly identical animals, a displacement effect was not obvious (see Figure 4).

The position-specific discrepancy between *same* and *different* trials was much more pronounced for the more similar than for the easily discriminable stimulus set. This was confirmed statistically by the significant three-way interaction (F[6,90]=2.48; $p < 0.05$). A direct interaction between SIMILARITY and IDENTITY was not observed (F[2,30]=1.23; $p > 0.1$).

RT results showed similar tendencies, although, as in Experiment 1, the effects were much less pronounced. SIMILARITY had a weak influence on latencies (F[2,30]=2.53; $p < 0.1$). Except for the TRANSLATION $\times$ IDENTITY interaction (F[3,45]=4.01; $p <$ 0.05), no main effects or interaction approached significance ($p > 0.1$).

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 very similar for the two subgroups. Most importantly, in both groups similarity influenced overall performance, but did not interact significantly with translation. Subjects starting with the easy discrimination condition were more accurate in all three experimental subsessions than those starting with the difficult task. It is unclear whether this reflects individual differences among subjects, or is due to different discrimination strategies.

3.3 Discussion

As in Experiment 1, no significant positional specificity was observed in Experiment 2, except for the prominent interaction with *same-different* identity. Furthermore, increasing the similarity among the stimuli did not interfere with translation invariance. This result is clearly different from the observations by Dill and Fahle (1997), who found that positional specificity increased with a decrease in the discriminability of random dot clouds and checkerboard patterns. In this sense, recognition of novel, complex patterns is qualitatively different from recognition of more familiar objects such as our animal-like stimuli, regardless of the similarity of the latter to each other.

Edelman's (1995) finding of an interaction between similarity and invariance with a very similar set of objects indicates that the effects of the transformation he studied – rotation in depth – and translation are not equivalent. While rotating an object relative to the observer strongly reduces accuracy and increases RT in a similarity-dependent way, translating it proved to have only a minor influence.

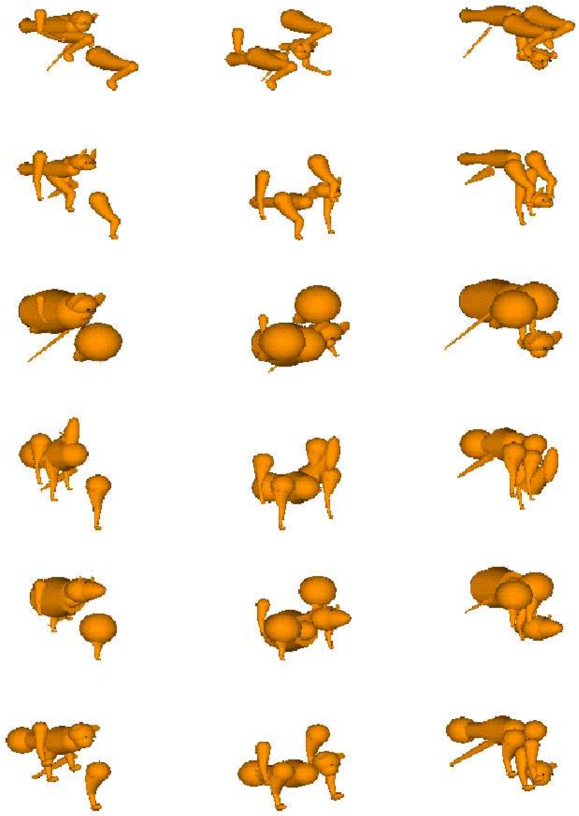

Figure 5: Examples of scrambled animals as used in Experiments 3 and 4.

4 Experiment 3: "Scrambled" animals - features

One major difference between our first two experiments and the earlier studies with complex random patterns (Foster & Kahn 1985; Dill & Fahle 1997a) was the general prior familiarity of the subjects with animal-like shapes. Although our computer-graphics models were not naturalistic copies of real animals, subjects readily named the animals when being introduced to the experiment and the stimuli. Notably, even the most similar animals are still meaningful objects that may be interpreted and classified by higher cognitive levels, because subjects are presumably over-trained to animal-like objects in everyday life. All positional specificity that is observed for novel stimuli may have been lost for such familiar objects, due to the decades of visual learning.

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 complete 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 basic features (limbs) as the original ones, but did not form a meaningful object (cf. Figure 5). Additionally, since the configuration of the limbs could be changed from trial to trial, repetition of stimuli and possible resulting 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 limbs (head, body, 2 forelegs, 2 hind legs, tail) into complete 3D animal objects, each limb 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 limbs. The random scrambling, however, was the same for both animals: homologous limbs (e.g., the heads) were shifted by the same amount in both stimuli. For each trial, the displacement of limb 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. Except for two subjects, the observers had not participated in Experiments 1 or 2.

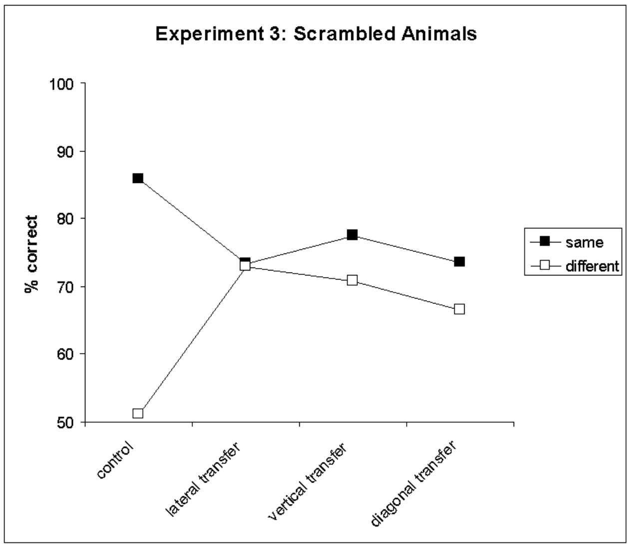

Figure 6: Scrambled animals - features (Experiment 3).

4.2 Results and Discussion

TRANSLATION had only a small insignificant effect on *same-different* discrimination of scrambled animals (F[3,21]=2.12; $p > 0.1$). As can be seen in Figure 6, this effect did not even have the correct sign that one would expect if visual short-term memory were position specific. Consistent with the above experiments, TRANSLATION strongly interacted with IDENTITY (F[3,21]=13.42; $p < 0.001$). Additionally, however, subjects displayed a tendency towards *same* decisions (F[1,7]=8.85; $p < 0.05$).

RT data again showed only minor effects that were consistent with the findings for accuracy, although neither the main effects nor the interactions approached significance ($p > 0.1$). As in the other experiments, there was no indication of a speed-accuracy tradeoff.

The results of Experiment 3 clearly show that the meaningful content of animal objects is not responsible for the complete translation invariance of discrimination performance in Experiments 1 and 2. Our reduction of the interpretability of the objects did not lead to positional specificity. Thus, discrimination of objects is translation invariant even if these objects are highly unfamiliar and difficult to label.

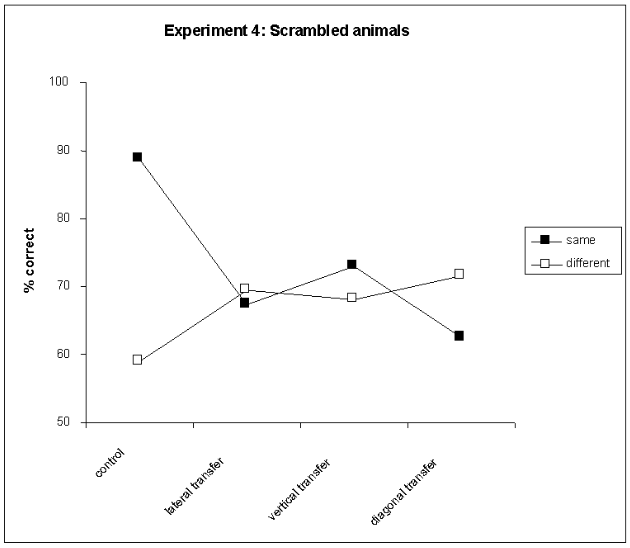

Figure 7: Scrambled animals - configuration (Experiment 4).

5 Experiment 4: "Scrambled" animals - configuration

Both the identity 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 limbs were identical for the two stimuli of a given trial, which only differed with respect to the shapes of the limbs employed. In Experiment 4, we created a complementary situation: both scrambled animals in a given trial were now composed of identically shaped limbs, and only differed in their spatial arrangement. If, for example, the first stimulus was a particular scrambled monkey, then the second stimulus in Experiment 4 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. columns in Figure 5). Both experiments, therefore, employed the same type of scrambled objects, but separated the effects of features (limb shapes) and feature relations (limb configuration).

5.1 Method

Experiment 4 was performed with exactly the same experimental procedure as Experiment 3, including the same kind of scrambled animals. However, the stimuli in each trial always consisted of the same set of limbs, scrambled in two different manners. Eight subjects participated in Experiment 4. Three of them were new to this experiment series; the remaining five had already participated in one or two of the first three experiments.

5.2 Results

Despite the employment of the same experimental procedure and stimulus type, results in Experiments 3 and 4 were different. In Experiment 3 (different features, same spatial arrangement) no effect had been obtained. For comparison, in Experiment 4 (same features, different arrangement), the main effect of TRANSLATION was significant (F[3,21]=4.66; $p < 0.05$). The reduction of accuracy in transfer trials was accompanied by a stronger interaction with IDENTITY (F[3,21]=13.32; $p < 0.001$). As a result, in *same* trials control was better than transfer performance, while it was even slightly worse in *different* trials. IDENTITY had no significant effect (F[1,7]=2.09; $p > 0.1$). Naive observers and those that already had participated in one of the earlier experiments did not show any obvious difference.

Effects on RT were again small. Only the interaction TRANSLATION × IDENTITY was significant (F[3,21]=4.84; $p < 0.05$). The main effects were not reliable ($p > 0.1$); their sign was inconsistent with a speed-accuracy tradeoff.

5.3 Discussion

The slight modification of the task between Experiments 3 and 4 — from discrimination by features to discrimination by their spatial relations — produced a considerable difference in the results. In Experiment 4, the occurrence of a particular limb was not diagnostic for discrimination, unless via a chance occlusion. Apparently, performance under these conditions was not completely invariant to translation, while discrimination by features in Experiment 3 was.

It is tempting to attribute this distinction to two different subsystems of object vision: one that is translation invariant and allows recognition of features and one that is at least partially position specific and is responsible for identification of feature relations. Alternatively, achieving translation invariant recognition of a particular stimulus feature may be bought by some uncertainty about its position in the visual field. To be able to discriminate objects on the basis of the spatial relations of some simpler features only, the system may have to rely on evidence from lower or intermediate processing levels that are not fully shift-invariant.

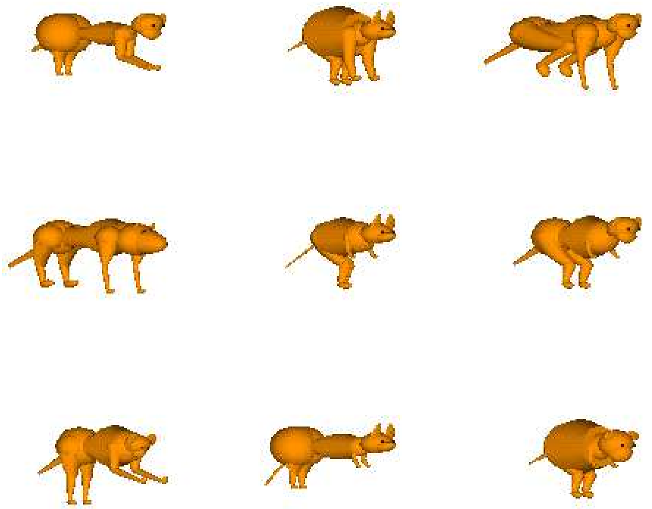

Figure 8: Example chimerae (Experiment 5).

6 Experiment 5: Chimerae

For Experiment 5, we created new animal objects by randomly combining limbs from different animals (cf. Figure 8). Aside from random similarity with "regular" animals, these chimerae were difficult to categorize into familiar animal classes. A second major difference compared to Experiments 1 and 2 was that new chimerae could be created for each new trial, thereby avoiding the development of their classification by the subject. Note that identification of a particular feature (e.g. the head) in two chimerae does not necessarily indicate that both stimuli are identical, because all other features may still be different. Subjects, therefore, were forced to attend to the entire configuration of each chimera.

6.1 Method

Experiment 5 followed the same basic design as Experiment 1. The only difference between the two experiments was that while in Experiment 1 only the original stimulus 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. Eight observers participated in Experiment 5, which consisted of three blocks of 96 trials each. Experiments 4 and 5 were run on 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 the other four started with the chimerae. No difference between the two groups be detected in the results.

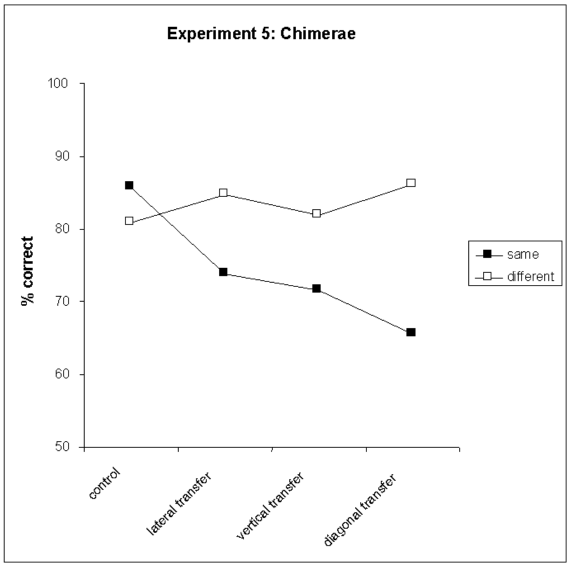

Figure 9: Chimerae (Experiment 5).

6.2 Results and Discussion

The mean accuracy results for this experiment are shown in Figure 9. As in all previous experiments, the most striking observation is that of the differential effect of displacement on *same* and *different* trials. The interaction TRANSLATION × IDENTITY again was highly significant (F[3,21]=5.41; $p < 0.01$). In addition to this interaction, however, both factors also displayed independent main effects. *Different* trials were generally more accurate than *same* trials (F[1,7]=9.74; $p < 0.05$). More importantly in the present context, translation significantly reduced performance (F[3,21]=3.67; $p < 0.05$). Both the influence of TRANSLATION (F[3,21]=3.62; $p < 0.05$) and its interaction with IDENTITY (F[3,21]=3.57; $p < 0.05$) were reliable. The effects of displacement on RT and accuracy are consistent: after transfer to a new location recognition is not only worse but also slower. Therefore, the displacement effect cannot be attributed to a speed-accuracy tradeoff.

Taken together, the results of Experiments 4 and 5 show that visual object discrimination is not completely invariant to translation. At least under certain experimental conditions, performance clearly profits if the stimuli that are to be compared appear at the same location of the visual field.

7 Experiment 6: Translation and Rotation in Depth

Translation in the image plane seems to have no deteriorating effect on the recognition of computer-graphics animal-like objects, unless certain stimulus and task manipulations are performed (cf. Experiments 4 and 5). A straightforward interpretation of these findings is that the visual system is able to recognize an object independently of its location, but, when forced to do so, it also relies on position-specific information that may be present at intermediate stages of visual processing.

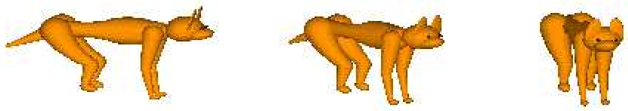

Figure 10: Three different orientations of an animal (Experiment 6).

Interestingly, Edelman (1995) found an influence of rotation in depth with animal stimuli that were very similar to the ones in Experiment 1. No additional manipulations were required to find strong orientational specificity of visual short-term memory. This indicates that while translation is in principle tolerated by the visual system, rotation invariance is not achieved in very similar tasks. Different brain processes may, therefore, be involved in compensating for the two object transformations, translation and rotation. However, the experiments described above and Edelman's earlier study differ slightly in their experimental procedures. We, therefore, decided to directly compare translation and rotation in a combined experiment with the set of six animal objects that we had employed in Experiment 1.

7.1 Method

Most conditions were exactly as in the previous experiments. The original set of six 3D animal objects was used in Experiment 6. As in Experiment 1, the location of the animal stimuli relative to the fixation spot was varied. Animals were located at 2 deg eccentricity on one of the four diagonal axes. Additionally, the objects in Experiment 6 could be presented at six different azimuth orientations in depth. Specifically, they could be rotated by 15, 45, or 75 deg to the left or right relative to the frontal view (cf. Figure 10). The elevation of the virtual camera was set to 10 deg above the horizon.

In each trial, the computer randomly selected one of the four possible locations and one of the six orientations for the first stimulus. The second stimulus again could be either the same or a different animal. It could appear at the same or the diagonally opposite location (corresponding to 4 deg displacement; lateral and vertical transfer were not tested) and could be rotated by 0, 30, or 60 deg relative to the first presentation. Eight subjects were tested in 8 × 72 trials each. Trials within each of the eight blocks were balanced for *same-different* identity and degree of translation and rotation. Subjects were explicitly told to ignore both position and orientation when deciding whether both stimuli represented same or different animal objects.

7.2 Results and Discussion

When combining the two visual transformations – rotation in depth and translation in the image plane –

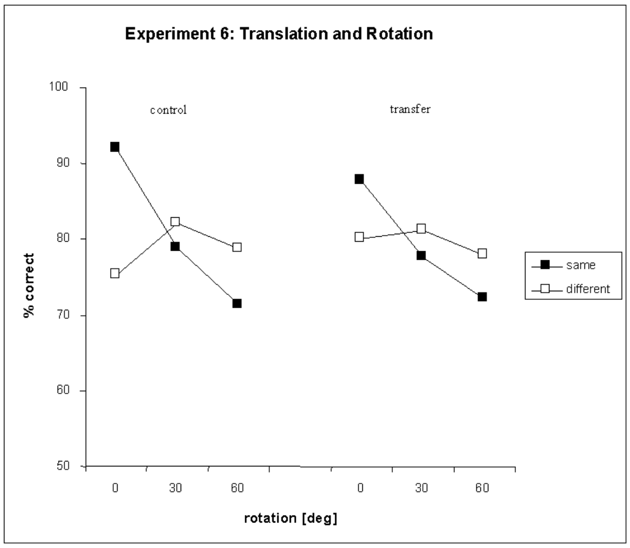

Figure 11: Translation and rotation (Experiment 6).

only orientation changes had a significant influence on accuracy (cf. Figure 11) and RT; position in the visual field did not affect discrimination performance. This was confirmed by the results of three-way ANOVAs (TRANSLATION × ROTATION × IDENTITY), indicating that only the main factor, ROTATION, had a significant effect (accuracy: $F[2,14]=11.69$; $p < 0.01$. RT: $F[2,14]=20.91$; $p < 0.001$). The other two main factors were negligible for the task ($p > 0.1$). A significant interaction between the two visual transformations was not observed (accuracy: $F[2,14]=0.12$; $p > 0.1$; RT: $F[2,14]=2.41$; $p > 0.1$).

Although in most of the above experiments translation did not interfere with overall performance, it had at least a strong differential impact on *same* and *different* trials, as indicated by the significant interactions TRANSLATION × IDENTITY. Interestingly, this interaction was very small in Experiment 6. Remnants of it may be apparent when only trials without rotation are considered. They were too weak, however, to show up as a reliable two-way (accuracy: $F[1,7]=1.60$; $p > 0.1$. RT: $F[1,7]=3.39$; $p > 0.1$) or three-way interaction (accuracy: $F[2,14]=2.16$; $p > 0.1$. RT: $F[2,14]=1.99$; $p > 0.1$). Instead, the influence of rotation on performance was largely specific to *same* trials, while *different* trials were hardly affected (interaction ROTATION × IDENTITY for accuracy: $F[2,14]=12.31$; $p < 0.001$. RT: $F[2,14]=3.30$; $p < 0.1$).

Experiment 6, therefore, clearly confirms both our earlier finding of full translation invariance, and Edelman's (1995) earlier result of incomplete invariance to rotation in depth for the same set of animal objects. This strongly indicates a qualitative difference between the brain mechanisms responsible for tolerance to these two visual transformations, confirming similar reports on recognition of wirelike objects (Rock & DiVita 1987; Bülthoff & Edelman 1992). While the visual system can in principle recognize a stimulus independently of its location, rotating objects in depth strongly hinders performance.

8 Experiment 7: Scale and Rotation in Depth

Given the result of Experiment 6, namely, that object recognition is invariant to translation but not to rotation in depth, it is tempting to ask for the influence of other transformations. Does, for example, a change in size affect perception of the same object type? Is invariance a unique property that is only achieved for translation? Or is orientation in depth special in that an object is not easily recognized after rotation even under conditions that allow tolerance to other transformations? To explore these issues, we designed Experiment 7, which is analogous to the previous experiment, except that animal objects now had to be compared across changes in size (instead of position) and orientation.

8.1 Method

The experimental procedure was generally similar to that of Experiment 6. However, animal objects in Experiment 7 always appeared at the same location in the center of the visual field. Instead of manipulating the position, their size could take one of eight values, differing by an isotropic scaling with respect to the center (factor of 1.19). The smallest size extended over about 2.5 × 1.5 deg, while the largest animals were approximately 8.5 deg wide and 5.5 deg high.

In each trial, the computer randomly selected one of the eight possible sizes and one of the six orientations for the first stimulus. The second stimulus differed in scale from the first one by a size ratio of 1, 1.41, or 2, and was rotated by 0, 30, or 60 deg relative to the first presentation. Eight subjects were tested in eight blocks of 72 trials each. Trials within a block were balanced for *same-different* identity, degree of rotation, and scale factor. Subjects were explicitly told to ignore both relative size and orientation when deciding whether both stimuli represented same or different animal objects.

8.2 Results

As in the previous experiment, rotation in azimuth strongly reduced the accuracy of the *same-different* discrimination ($F[2,14]=17.85$; $p < 0.001$) and increased the response latencies ($F[2,14]=59.47$; $p < 0.001$). As can be seen in Figure 12, ROTATION also interacted with *same-different* IDENTITY (accuracy: $F[2,14]=5.11$; $p < 0.05$. RT: $F[2,14]=7.74$; $p < 0.01$): effects were strong in *same* trials, and less homogeneous in *different* trials.

Interestingly, changing the size of the animals had no general influence on performance (accuracy: $F[2,14]=0.49$; $p > 0.1$. RT: $F[2,14]=0.85$; $p > 0.1$), nor did it interact with IDENTITY (accuracy: $F[2,14]=0.01$; $p > 0.1$. RT: $F[2,14]=3.04$; $p > 0.05$) or with ROTATION (accuracy: $F[4,28]=0.44$; $p > 0.1$. RT: $F[4,28]=1.11$; $p > 0.1$). Within the range of size ratios we tested, scaling animal-like objects, therefore, seems to be tolerated by visual recognition processes.

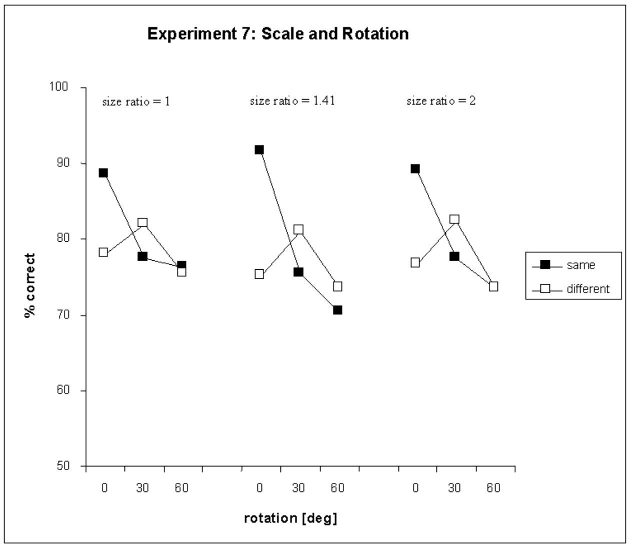

Figure 12: Size and rotation (Experiment 6).

8.3 Discussion

As for translation in Experiment 6, we found no significant influence of scaling. Apparently, the visual system is able to tolerate changes in size, at least to the extent we tested. We did not probe the limits of size invariance, nor did we test whether or not there is a similar stimulus and task dependence that we had found for translation invariance. Thus, the present findings indicate that size and translation invariance constitute problems that are relatively easily solved by the brain.

Rotations in depth, on the other hand, seem to be much more difficult to compensate for. Whereas some studies reported orientation-invariant recognition (Biederman 1987), others found orientation-dependent recognition for a wide variety of stimuli (Rock & DiVita 1987; Bülthoff & Edelman 1992; Jolicoeur & Humphreys 1997). Studies specifically manipulating object familiarity found that to achieve a significant 3D orientation invariance, the visual system has to go through a learning process, whose results, moreover, do not fully transfer to novel stimuli (Moses *et al.* 1996; Gauthier & Tarr 1997).

The influence of different visual transformations on recognition of the same object has been compared before: Bricolo and Bülthoff (1993) reported preliminary evidence that accuracy of recognition of wirelike objects is reduced after rotation in depth, but not after changes in position, size, and illumination (Bricolo & Bülthoff 1993; Bricolo 1996). Two earlier studies (Kubovy & Podgorny 1981; Larsen 1985) testing the influence of scaling and rotation in the image plane yielded very similar results, although the reported data are incomplete in that the authors focused on response times, and for most experiments discarded *different* trials. From the available information it is obvious, however, that rotation had a strong effect, while size changes were largely or even completely tolerated. Larsen (1985) further showed that by manipulating the size of the stimulus set small effects of scaling can be obtained. This dependence on the particulars of the task parallels our present findings for translation. While transformation of both position and size may result in reduced performance under some circumstances, the visual system seems to have found a way to tolerate them in principle. A lack of invariance only shows up if the stimuli are specifically tailored to uncover it. However, achieving invariance to rotation – both in the image plane and in depth – seems to be a more fundamental problem, which may be solved only after extensive learning, specific to the stimulus class.

9 General Discussion

The aim of our study was twofold: to test translation invariance of object vision and to compare the effects of translation with those of other transformations for the same set of stimuli. Our results show that translation invariance is in principle achieved by the visual system. Notably, complete translation (and scale) invariance is obtained with a set of 3D objects known to evoke strong orientational specificity in *same-different judgment*. These findings are largely consistent with Biederman and Cooper's (1991) claim that visual priming is translation-invariant. They also confirm reports that recognition of wirelike objects is sensitive to rotation in depth, but not to changes in position, size, and illumination (Bricolo & Bülthoff 1993; Bricolo 1996).

From the computational point of view the difference between 3D rotation and a shift or scale change in the image plane may seem plausible. As pointed out by Vetter *et al.* (1995), specific object-knowledge is required to generate rotated virtual examples from a single view of that object. For simple image-plane transformations like translation or scaling no additional information is needed. It is not surprising, therefore, that many electrophysiological investigations on higher visual areas of the brain found neurons that respond to a specific stimulus largely independent of its size and position in the visual field (Schwartz *et al.* 1983; Tovee *et al.* 1994; Ito *et al.* 1995; Logothetis *et al.* 1995), while in many cases these cells are highly selective to only one orientation of this stimulus (Perrett *et al.* 1985; Wachsmuth *et al.* 1994; Logothetis *et al.* 1995; Wang *et al.* 1996).

Our experiments, however, demonstrate that discrimination performance *does* suffer from object translation when distinguishing among objects that differ in their structure rather than in their local features. This finding, along with the earlier results concerning the effects of translation on dot-cloud and checkerboard-pattern discrimination (Foster & Kahn 1985; Nazir & O'Regan 1990; Dill & Fahle 1997a; Dill & Fahle 1997b), indicates clearly that the mechanisms that allow visual objects to be perceived and remembered independently of the location are far from universal in the kinds of objects for which they are effective.

In general, the effects of the various transformations as they emerge from the psychophysical studies conducted in the past decade support the notion that ascent in the visual pathway brings about an increase in the invariance of the representations, accompanied by a decrease in the amount of information available for distinguishing between similar stimuli. This notion is intuitively acceptable, and is broadly compatible with the electrophysiological characterization of the receptive field properties in the mammalian visual pathway (Desimone *et al.* 1984; Desimone *et al.* 1985; Logothetis & Sheinberg 1996; Rolls 1996; Tanaka 1996). The partial failure of invariance for complex unfamiliar stimuli and for stimuli that differ structurally but not locally suggests the possibility of a theoretical refinement of that general picture.

An intriguing candidate theory, whose full consideration is beyond the scope of the present note, is the involvement of a common principle in the visual system's quest for invariance. The principle we propose is reliance on mechanisms trained for particular object classes (Edelman 1997), operating on top of a universal preprocessing stage that provides a pedestal performance, irrespective of the object identity (Edelman *et al.* 1997; Riesenhuber & Poggio 1997). This proposal is related to Ullman's (1995) idea of a division of object transformations into two kinds. Transformations of the first kind can be compensated for irrespective of object identity, in a bottom-up fashion; for example, translation in the image plane can be countered by centering the luminance distribution, an operation that does not require knowledge of the object that gave rise to this distribution. Transformations of the second kind can only be compensated for if the object's geometry is taken into account. For example, the effect of rotation in depth on the appearance of an object can be countered if prior experience with similar objects is put to use (Moses *et al.* 1996; Lando & Edelman 1995; Gauthier & Tarr 1997). We note that a recognition mechanism operating on this principle would perform worse with objects that are radically novel in their structure, or very complex, because such objects would not activate differentially the "modules" tuned to different familiar objects, impeding discrimination (especially among highly similar stimuli). Putting this scheme for recognition and categorization of transformed objects to an explicit psychophysical test is left for future work.

Acknowledgments M. Dill was supported by Boehringer Ingelheim Fonds. We thank M. Fahle, G. Geiger, T. Poggio, M. Riesenhuber, and P. Sinha for valuable discussions.

References

I. Biederman (1991) Recognition by components: a theory of human image understanding. *Psychological Review* 84:115-147.

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

E. Bricolo (1996) On the Representation of Novel Objects: Human Psychophysics, Monkey Physiology and Computational Models. Ph.D. Thesis. Massachusetts Institute of Technology.

E. Bricolo and H.H. Bülthoff (1993) Rotation, Translation, Size and illumination invariances in 3D object recognition. *Investigative Ophthalmology & Visual Science (Suppl.)* 34:1081.

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

R. Desimone, T.D. Albright, C.G. Gross, and C.J. Bruce (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience* 4:2051-2062.

R. Desimone, S.J. Schein, J. Moran, and L.G. Ungerleider (1985) Contour, color and shape analysis beyond the striate cortex. *Vision Research* 25:441-452.

M. Dill and M. Fahle (1997a) Limited Translation Invariance of Human Visual Pattern Recognition. *Perception & Psychophysics* (in press).

M. Dill and M. Fahle (1997b) The Role of Visual Field Position in Pattern-Discrimination Learning. *Proceedings of the Royal Society London B* (in press).

M. Dill, R. Wolf, and M. Heisenberg (1993) Visual pattern recognition in Drosophila involves retinotopic matching. *Nature* 365:751-753.

M. Dill and M. Heisenberg (1993) Visual pattern memory without shape recognition. *Philosophical Transactions of the Royal Society* B349:143-152.

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

S. Edelman (1997) Representation is representation of similarity. *Behavioral and Brain Sciences* (in press).

S. Edelman, N. Intrator, and T. Poggio (1997) Complex cells and object recognition. (submitted).

P. Földiák (1991) Learning invariance from transformation sequences. *Neural Computation* 3:194-200.

D.H. Foster and J.I. Kahn (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.

D.H. Foster and J.I. Kahn (1985) Becoming a *Greeble* Expert: Exploring the Face Recognition Mechanism. *Vision Research* 37:1673-1682.

M. Ito, H. Tamura, I.Fujita, and K. Tanaka (1995) Size and Position Invariance of Neural Responses in Monkey Inferotemporal Cortex. *Journal of Neurophysiology* 73:218-226.

P. Jolicoeur and G.K. Humphreys (1997) Perception of rotated two-dimensional and three-dimensional objects and visual shapes. In *Visual Constancies: Why things look as they do.* (Eds.: V. Walsh

& J. Kulikowski) Cambridge University Press (in press).

M. Kubovy and P. Podgorny (1981) Does pattern matching require the normalization of size and orientation? *Perception & Psychophysics* 30:24-28.

A. Larsen (1985) Pattern matching: Effects of size ratio, angular difference in orientation, and familiarity. *Perception & Psychophysics* 38:63-68.

M. Lando and S. Edelman (1995) Receptive field spaces and class-based generalization from a single view in face recognition. *Network* 6:551-576.

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

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

M. Minsky and S. Pappert (1969) Perceptrons. MIT press.

Y. Moses, S. Ullman, and S. Edelman (1996) Generalization to novel images in upright and inverted faces. *Perception* 25:443-46.

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

R.C. O'Reilly and M.H. Johnson (1994) Object Recognition and Sensitive Periods: A Computational Analysis of Visual Imprinting. *Neural Computation* 6:357-389.

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

D.I. Perrett, P.A.J. Smith, D.D. Potter, A.J. Mistlin, A.S. Head, A.D. Milner, and M.A. Jeeves (1985) Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society London B* 223:293-317.

M. Riesenhuber and T. Poggio (1997) Just One View: Invariances in Inferotemporal Cell Tuning. (submitted).

I. Rock and J. DiVita (1987) A Case of Viewer-Centered Object Perception. *Cognitive Psychology* 19:280-293.

E. T. Rolls (1996) Visual processing in the temporal lobe for invariant object recognition. In *Neurobiology.* (Eds.: V. Torre & T. Conti). p. 325-353. New York: Plenum Press.

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

E.L. Schwartz, R. Desimone, T.D. Albright, and C.G. Gross (1983) Shape recognition and inferior temporal neurons. *Proceedings of the National Academy of Sciences* 80:5776-5778.

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

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

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

T. Vetter, A. Hurlbert, and T. Poggio (1995) View-based Models of 3D Object Recognition: Invariance to Imaging Transformations. *Cerebral Cortex* 5:261-269.

E. Wachsmuth, M.W. Oram, and D.I. Perrett (1994) Recognition of Objects and Their Component Parts: Responses of Single Units in the Temporal Cortex of the Macaque. *Cerebral Cortex* 5:509-522.

G. Wang, K. Tanaka, and M. Tanifuji (1996) Optical Imaging of Functional Organization in the Monkey Inferotemporal Cortex. *Science* 272:1665-1668.