Class similarity and viewpoint invariance in the recognition of 3D objects

Shimon Edelman
Dept. of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100, Israel
Internet: edelman@wisdom.weizmann.ac.il

Abstract

In human vision, the processes and the representations involved in identifying specific individuals are frequently assumed to be different from those used for basic-level classification, because classification is largely viewpoint-invariant, but identification is not. This assumption was tested in psychophysical experiments, in which objective similarity between stimuli (and, consequently, the level of their distinction) varied in a controlled fashion. Subjects were trained to discriminate between two classes of computer generated 3D objects, one resembling monkeys, and the other dogs. Both classes were defined by the same set of 56 parameters, which encoded sizes, shapes, and placement of the limbs, the ears, the snout, etc. Interpolation between parameter vectors of the class prototypes yielded shapes that changed smoothly between monkey and dog. Within-class variation was induced in each trial by randomly perturbing all the parameters. After the subjects reached 90% correct performance on a fixed canonical view of each object, discrimination performance was tested for novel views that differed by up to 60° from the training view. In experiment 1 (in which the distribution of parameters in each class was unimodal) and in experiment 2 (bimodal classes), the stimuli differed only parametrically and consisted of the same geons (parts), yet were recognized virtually independently of viewpoint in the low-similarity condition. In experiment 3, the prototypes differed in their complement of geons, yet the subjects' performance depended significantly on viewpoint in the high-similarity condition. In all three experiments, higher inter-stimulus similarity was associated with an increase in the mean error rate and, for misorientation of up to 45°, with an increase in the degree of viewpoint dependence. These results suggest that a geon-level difference between stimuli is neither strictly necessary nor sufficient for viewpoint-invariant performance. Thus, the standard characterization of basic and subordinate-level processes in visual recognition may need a revision.

1 Features of recognition

The issue of representation is of central importance in recognition, as it is in other areas of vision. Consequently, the development of successful recognition schemes may be aided by progress in finding out how objects and object classes are represented in human vision. Theories of recognition have proposed different approaches to the representation problem. A prominent recent example is structural description in terms of geons (generalized cylinders representing object parts) in the RBC, or recognition by components, scheme (Biederman, 1987). An alternative approach (Ullman and Basri, 1991; Poggio and Edelman, 1990) calls for representing objects by small collections of 2D images. It has been shown how recognition can be performed using such representations, e.g., by a process of interpolation between the stored 2D images.

In human vision, the representations used for identifying specific instances of object classes (the so-called "subordinate level") are frequently postulated to be different from those used for basic-level classification (Jolicoeur, 1990). On one hand, the recently demonstrated viewpoint dependency of subordinate-level recognition (Rock and DiVita, 1987; Tarr and Pinker, 1989; Edelman and Bülthoff, 1992; Bülthoff and Edelman, 1992) is consistent with theories that hold that the visual system represents three-dimensional objects by storing several of their two-dimensional views. Developments in computational vision, and especially new approaches to model-based recognition (Ullman and Basri, 1991; Poggio and Edelman, 1990; Edelman and Poggio, 1992), support such a possibility. On the other hand, psychophysical findings on basic-level recognition seem to point towards the representation of object prototypes in a more symbolic manner, for example by collections of volumetric primitives or components (Biederman, 1987).

Does the visual system rely on distinct sets of representations and processes for subordinate and basic levels of recognition? A recent proposal (Edelman, 1991) outlined a possible unified approach to recognition at both levels, based on the notion of "features of recognition." The central tenet of the proposed account is that recognition normally requires neither 3D reconstruction of the stimulus, nor the maintenance of a library of 3D models of objects. Instead, information sufficient for recognition can be found in the 2D image locations of object *features*. The choice of features and their complexity may vary among objects. For example, a pineapple can be recognized by its characteristic pattern of spiny scales. The main feature in this case is textural and is distributed over the object's surface. In comparison, the relevant features of a peanut are both its texture and a characteristic outline. To consider a more complex example, an aircraft can be classified as such by the presence of wings, which may be considered as complex features. At the same time, for the image of an aircraft to be recognized, e.g., as a fighter or a passenger jet, more basic features such as contour elements and corners must be appropriately situated in the image (in the vicinity of the locations of corresponding features in the image of a prototypical aircraft).

The highlights of the approach that was proposed in (Edelman, 1991) were as follows:

1. *Versatility:* Recognition starts with the extraction of a large variety of image-based features.

2. *Plasticity:* Recognition procedure for a given object at a given category level is synthesized at need and is optimized with practice.

3. *Hierarchy:* One of the ways of optimizing recognition performance involves formation of compound features out of simpler ones, and subsequent reliance on such features.

4. *Invariance/diagnosticity tradeoff:* Some of the features are well-localized within the 2D image-based reference frame. Exclusive reliance on such features under certain circumstances makes recognition viewpoint-dependent. In comparison, features defined over extended regions are likely to support viewpoint-independent performance, at the expense

of the ability of the system to discriminate among members of the same basic category that differ only in local details.[1]

The predicted tradeoff between viewpoint invariance of a feature and its diagnosticity, or the degree of discrimination among object instances that it affords, deserves some clarification. Consider, for instance, a domain of objects composed of generalized cylinders and polyhedra. To recognize such objects, a visual system can use local features such as image-plane positions of object corners or edges, as well as extended features such as patterns of shading over object surfaces. The two types of features will, in general, lead to different performance. When the pose of the object relative to the viewer changes, the projected locations of the corners will shift. Unless this shift is compensated for (e.g., by pose recovery and model alignment (Ullman, 1989)), recognition of unfamiliar views of the object will be poor. In comparison, the shape of a shaded patch can in principle be extracted regardless of the pose of the object to which it belongs (as long as the patch is visible). At the same time, when the pose is fixed, projected corners, edges, or other localized features offer better discrimination among similar shapes than shading (cf. Bülthoff and Mallot, 1988).

In human subjects, difficulty in recognizing novel views is a central characteristic of performance in tasks that require discrimination among members of the same basic category (Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992). When the objects are to be classified at the basic level, recognition performance depends on viewpoint to a much lesser extent (Biederman, 1987; Biederman and Gerhardstein, 1993). These observations suggest that the two patterns of performance emerge in response to the different levels of detail that must be addressed in subordinate and basic-level recognition. If this is true, then one would expect the extent of viewpoint invariance in subjects' performance to be affected by a manipulation of the relevant level of detail, determined by inter-object similarity. The present paper reports an experimental demonstration of this effect in human subjects, and its computational modeling and analysis.

***** Figure 1 here *****

2 Psychophysics

2.1 Experimental methods

In three experiments designed to demonstrate the tradeoff between viewpoint invariance and discriminative power of features, subjects were trained to tell apart parameterized computer-generated three-dimensional monkey and dog-like objects (see Figure 1). The subjects were shown a succession of isolated static images of objects belonging to these two classes, which had to be discriminated by pressing one of two buttons on a computer mouse. Each trial was initiated by displaying a fixation aid in the center of the screen for $250 msec$. Immediately afterwards, the stimulus image was displayed for $20 msec$, and was followed by a mask (an image of a collection of 20 tapered cylinders, with size, orientation, location, and taper factor randomized anew for each trial). The subsequent trial followed $500 msec$ after the subject's response. The display interval was short, to prevent the subjects from employing a conscious discrimination strategy,

[1] Essentially non-spatial features such as distinctive color that are viewpoint-invariant because their perception has little to do with viewing geometry are not considered here.

and to keep performance below ceiling, so that manipulation of independent variables in the experiments would have an opportunity for a discernible effect.

***** Figure 2 here *****

During training the objects always appeared at a limited range of attitudes ($\pm 10°$) around a fixed orientation, corresponding roughly to the "three quarters" frontal view as defined by Palmer et al. (1981). Auditory feedback was given for incorrect responses, until the subject reached 90% correct performance in the trailing 20 trials. At that point, the subject was notified by an auditory signal that the testing stage was about to begin. During testing there was no feedback, and the objects appeared at attitudes that differed from the training attitude by a rotation in depth around either the horizontal or the vertical axis. The subjects' performance was measured by the combined percentage of correct positive responses to the two objects. This measure of performance is non-parametric, and is not affected by the subject's bias towards either of the two possible responses (Green and Swets, 1966, p.404).

2.2 Experiment 1

In the first experiment, the parameter distribution corresponding to each of the two object classes was unimodal (each parameter had an independent Gaussian distribution, with a standard deviation of 0.075 times the mean value of that parameter.). The main independent variable in this experiment was the distance between the centers of the two distributions (see Figure 2, top). The two distributions were closer in the first than in the second session for four subjects, and vice versa for another four subjects. The mean response time was $741 \pm 17 msec$, and the mean percentage of correct responses (CR) was $78.5 \pm 1.6\%$. The lack of speed-accuracy tradeoff was signified by a negative correlation between response time and percent correct ($t(1, 68) = -3.7$, $p < 0.0005$). A multiple-range Duncan test of the CR means by subject divided the subjects into two non-overlapping groups, with CR $> 80\%$ in the larger group (five subjects) and CR $< 70\%$ in the smaller group (three subjects). The data from the three subjects in the poor-performance group were omitted from subsequent consideration (see section 2.5 for a discussion). The mean CR after this deletion was 83%.

***** Figure 3 here *****

The data were then subjected to a homogeneity-of-slopes analysis of variance (ANOVA), using the SAS GLM procedure, with Subject specified as a random class effect, Similarity as a fixed class effect, and D (misorientation of the stimulus relative to the training attitude) — as a continuous effect. The analysis revealed significant effects of Similarity ($F(1, 22) = 27.7$, $p < 0.0001$), of the misorientation D ($F(1, 22) = 8.0$, $p < 0.01$), and the interaction Similarity $\times$ D ($F(1, 22) = 3.1$, $p < 0.09$). The main effect of Subject was n.s. ($F < 1$). Separate analyses for the two levels of Similarity (NEAR and FAR prototypes in the parameter space) showed a marginal effect of D in the FAR condition ($F(1, 9) = 2.9$, $p = 0.12$), and a significant effect of D in the NEAR condition ($F(1, 9) = 8.0$, $p < 0.02$).[2]

Linear regression analysis (SAS procedure REG) revealed a similar pattern of different slopes in the two similarity conditions. The regression in the FAR condition was n.s. (the slope not significantly different from 0). The slope of the linear regression in the NEAR condition was

[2]The effect of Subject was n.s. in the NEAR condition, but was present in the FAR condition ($F(4, 9) = 6.2$, $p < 0.01$). The interaction between Subject and D when separated by Similarity levels could not be estimated from the present design.

-0.31 ± 0.14 (regression significant at $p < 0.035$). These figures support the notion of the invariance-discrimination tradeoff predicted by the features of recognition theory.

2.3 Experiment 2

In the second experiment, each of the two classes of objects consisted of two subpopulations, or modes (see Figure 2, bottom). Each of the two modes in a class was Gaussian, with the same standard deviation as in the previous experiment. The distance between the means of the two modes was always 0.15 times the distance between the "reference" points in the parameter space that corresponded to the prototypical monkey and dog. Of the two modes, one was always situated at the appropriate reference point, and the other was either in between the reference ones (on the line in the parameter space connecting the prototypes), or outside them.

***** Figure 4 here *****

This arrangement of stimuli was reported by the subjects to be more difficult to learn (this difficulty was also reflected in the longer training sessions), possibly due to the bimodal distribution of parameter values within each object class.[3] Ten subjects participated in this experiment. The mean response time was $850 \pm 15msec$, and the mean percentage of correct responses was $80.5 \pm 0.8\%$. The lack of speed-accuracy tradeoff was signified by a negative correlation between response time and percent correct ($t(1,198) = -4.0$, $p < 0.0001$). As in the previous experiment, data from subjects whose mean CR was below 80% and who were grouped in the lowest performance interval by the Duncan test were discarded (there were four such subjects).

The performance of the remaining six subjects as a function of the misorientation of the stimulus with respect to the training attitude is plotted in Figure 4. The plot of CR vs. D for the inner mode of the NEAR condition revealed a "knee" at $D = 45°$. A comparison of least-squares adjusted means, produced by the GLM procedure, confirmed this impression (means for $D = 15°, 30°, 45°$ all different from each other at $p < 0.01$; means for $D = 45°, 60°$ not significantly different from each other). Thus, the subsequent analysis was carried out for $D \in [15°, 45°]$ only (this decision is discussed in section 2.5).

The analysis was carried out by a homogeneity-of-slopes ANOVA (Subject $\times$ Mode $\times$ D; for an illustration of the four levels of Mode, see Figure 4, left panel), using the GLM procedure. There was a strong main effect of D ($F(1,39) = 70.4$, $p < 0.0001$), and a significant Subject $\times$ Mode interaction. A hint of a $D \times$ Mode interaction was also present ($F(3,39) = 1.5$, $p = 0.22$), which prompted a separate by-Mode linear regression analysis. The regression results were highly significant in all four modes, and the slopes were, respectively, -0.38 ± 0.14, -0.29 ± 0.15, -0.38 ± 0.09, and -0.57 ± 0.11. The last result shows that the dependence of CR on D was indeed higher under high similarity (the inner mode of the NEAR condition) than in the other three conditions.

[3] Research by Holyoak and others showed that subjects are sensitive not only to the mean tendencies of the distributions of stimulus parameters, but also to their variability (Fried and Holyoak, 1984). Bimodal distributions are initially treated as if they were unimodal, leading to impaired performance on the exemplars that are deemed to be "outliers." It may take many hundreds of trials for the tacit assumption of unimodality to be dropped (Flannagan et al., 1986).

2.4 Experiment 3

The results of experiments 1 and 2 indicate that it is not necessary for two objects to differ in their part (geon) structure to obtain performance that varies little with viewpoint. Specifically, such performance was obtained for the monkey and dog stimuli, provided that they were sufficiently widely separated in parameter space, even though those two object classes had exactly the same geons in the same nonaccidental relationships with respect to each other. Experiment 3 was an attempt to show that geon difference, in addition to not being strictly necessary, is also not always sufficient for obtaining viewpoint-invariant performance.

***** Figure 5 here *****

The stimuli in experiment 3 were versions of the monkey and dog shapes that were modified so as to have five separate nonaccidental contrasting features. These features were obtained by turning parts that were previously cylindrical or ellipsoidal in both objects into different geons in each of them. For example, the tapered-cylinder foreleg in the monkey became a concave generalized cylinder, and in the dog it became convex. The entire set of newly introduced contrasts was such that the objects could not be distinguished merely by the presence or absence in the image of a geon of a given type. The degree of nonaccidental contrast was varied with the blending parameter α, so that, for instance, the convexity and the concavity of the generalized cylinders became less or more prominent. Still, because at their closest separation (of 0.70 times the distance between the reference modes) the two Near modes were distinct enough, the nonaccidental contrasts remained noticeable under all conditions.

Five subjects participated in this experiment. The mean response time was $781 \pm 24msec$, and the mean percentage of correct responses was $85.5 \pm 1.5\%$. The lack of speed-accuracy tradeoff was signified by the lack of a positive correlation between response time and percent correct ($t(1, 98) < 1$, n.s.). One of the five subjects was rejected by the mean CR criterion (same as the one employed in the analysis of the previous two experiments).

Figure 5 shows a plot of CR vs. D for the remaining four subjects. The curve for the Near condition, but not for the Far condition, revealed an upward concavity, therefore the subsequent analysis was carried out only for $D \in [15°, 45°]$ to facilitate the comparison of the linear trends in the two conditions (see the discussion in section 2.5).

A homogeneity-of-slopes GLM analysis of variance ($D \times$ Similarity $\times$ Subject) showed a significant main effect of D ($F(1, 11) = 10.3$, $p < 0.008$), and a significant Similarity $\times$ Subject interaction ($F(3, 11) = 11.3$, $p < 0.001$). Data were pooled over levels of Mode (the effects of Mode were n.s.). As in experiment 2, there was an indication of possible $D \times$ Similarity interaction ($F(1, 11) = 1.7$, $p = 0.22$), and a separate by-Similarity linear regression analysis was carried out. The slopes in the Far and the Near conditions were, respectively, -0.11 ± 0.10 (regression n.s.), and -0.27 ± 0.14 (regression significant at $p < 0.06$). Thus, here as in experiment 2, the dependence of CR on D was higher under high similarity (in the Near condition).

2.5 Discussion

2.5.1 Qualifications

The conclusions that may be drawn from the data presented in this section are subject to two qualifications. The first of these has to do with the rejection of data from poorly performing subjects. Each experimental session started with a training phase, and the subject could only pass on to the testing phase if his or her performance on the training images was better than 90%. The rejection criterion for the data from the testing phase was then set at a mean correct rate of 80%, which allowed for the lower performance in generalization to novel views, while discarding data from subjects who, in a sense, failed to learn the task.

The second qualification is concerned with the range of misorientation D for which the linear trends reported above hold. In experiment 1, the range of D was 15° to 45°, and the trends there were clearly discernible. In the other two experiments, however, D was in the interval $[15°, 60°]$, and a separate consideration of low and high misorientation effects had to be made. Specifically, the linear trends in these experiments were only apparent up to $D = 45°$. The bottoming out of the effects of D on CR for $D \geq 60°$ may be attributed to the onset of a different viewpoint-dependence (or rather, viewpoint invariance) mechanism than the one that is at work for smaller values of D. This phenomenon warrants further investigation, but it does not preclude obtaining a meaningful characterization of human performance for $D \leq 45°$.

2.5.2 Summary of psychophysical findings

Subject to the above qualifications, the results of the three experiments can be summarized as follows:

- A geon-level difference between stimuli was not *necessary* for nearly viewpoint-invariant performance: the two stimuli in experiments 1 and 2 differed only parametrically, and had the same complement of geons, yet were recognized relatively independently of viewpoint in the FAR condition of experiment 1.
- A geon-level difference between stimuli was not *sufficient* for achieving viewpoint invariance, as indicated by the significantly viewpoint-dependent performance of subjects in the NEAR condition of experiment 3.
- In all three experiments, increasing the inter-stimulus similarity affected two characteristics of recognition performance:
  - mean percentage of correct responses CR deteriorated;
  - the degree of viewpoint dependence, as reflected in the slope of the regression of CR on stimulus orientation relative to training, increased (the slope became more negative).

Altogether, the performance of the 15 (out of the total of 23) subjects whose performance passed the acceptance criterion described above suggests that the standard characterization of basic and subordinate-level processes in visual recognition may need a revision. The next section describes computational simulations that hint at a possible direction such a revision could take.

3 Simulations

According to the psychophysical data described above, the dependence of human recognition performance on viewpoint varies with inter-stimulus similarity in a manner that is compatible with predictions of the features of recognition (FOR) approach, outlined in section 1. One way to gain a computational understanding of these psychophysical results is through simulation of the experiments. This section presents such a simulation. The model used to replicate the psychophysical experiments was based on an interpolating classifier that represents 3D objects by collections of their 2D views. The classifier used Radial Basis Functions to interpolate among the stored 2D views of objects (Poggio and Edelman, 1990). Previous experience with this approach to the modeling of subordinate-level recognition of 3D objects had been positive (see, e.g., Bülthoff and Edelman, 1992). However, so far the problem of representation of individual views of objects has been circumvented by supplying the classifier with representations assumed to be computed by a separate mechanism. The present work makes a step towards clarifying the nature of representation of individual views, by preceding the classification stage of the RBF model with a simple feature extraction stage, and by investigating the extent to which human performance in recognition can be replicated by the resulting scheme.

***** Figure 6 here *****

3.1 Transduction stage

The feature extraction method used in the simulations was chosen to satisfy two criteria. First, the method had to be generic rather than elaborate and specific (employing any of the more sophisticated available approaches to feature extraction developed in computer vision would have amounted to forcing an answer to the question of the features of recognition). Second, the method had to be computationally viable. Feature extraction by convolution of the input image with a bank of localized receptive fields (RFs) meets both of the above requirements: it does not commit the entire simulation framework to a particular choice of high-level features, and it has a record of success in modeling low-level visual functions such as hyperacuity (Poggio et al., 1992), as well as higher-level functions such as face recognition (Edelman et al., 1992). Related methods of feature extraction have been proposed repeatedly in the past (e.g., Amari, 1978; Nishihara and Poggio, 1984; Snippe and Koenderink, 1992; Edelman, 1992). In the simulations described here I assumed the individual RFs to possess a Gaussian profile, with an x/y aspect ratio distributed uniformly between 0.1 and 10.0. The density of the coverage of the input image by RFs decreased from the center outward, also according to a normal distribution. Separate simulations were run for three values of the number of the RFs (which determines the dimensionality of the representation passed on to the subsequent classification stage): 150, 200, and 400. A typical arrangement of the RFs with respect to an input image is illustrated in Figure 6.

3.2 Classification stage

Classification of the input represented by the vector of activities of the transducer receptive fields was performed by a Radial Basis Function (RBF) classifier (e.g., Moody and Darken, 1989). The classifier was first trained on ten images of each of the two stimuli objects (these images were taken from the same range of viewpoints that was used in the training of human subjects). The twenty training images were taken to be the twenty basis function centers of the classifier, which was trained to output $+1$ for the images of the monkey, and -1 for the images of the dog. The classifier's performance was then tested using exactly the same images as seen by the human subjects in experiment 2. The outcome of a trial was considered to be correct if the output of the classifier had the correct sign (there was no attempt to model noise at the decision stage). The simulated experiment was repeated for three different values of the parameter that determined the width σ of the basis functions.[4] The three values of the σ factor, combined with the three values of the number of receptive fields in the transduction stage, yielded a 9-fold replication of the simulated experiment. The results of the simulations are presented below as means and standard errors of the percentage of correct responses of the classifier over these nine runs. Note that each simulation run consisted in fact of two testing blocks, each preceded by its own training stage, just as in the real experiments there were always two sessions, each with a separate training and testing stage.

3.3 Simulation results

The results of the simulations are summarized in Figure 7. A comparison of this figure with the corresponding summary of human data that appears in Figure 4 reveals some similarities, as well as major discrepancies. Two apparent similarities are the order of performance levels in the different modes (the best for the outer mode in the FAR condition, and the worst for the inner mode in the NEAR condition), and the slower deterioration of performance with D in the outer mode in the FAR condition, up to and including $D = 45°$, relative to the other three modes. The most noticeable discrepancy is in the absolute level of the performance floor: human performance remained well above chance for all tested values of D, while the model's performance dropped to chance at $D = 60°$.

***** Figure 7 here *****

3.4 Discussion

The simulated experiment described above attempted to replicate human performance in the real experiments using a simple two-stage model of recognition, in which transduction was followed by interpolating classification. Despite the much stronger sensitivity of the model to viewpoint, it did surprisingly well, given that it only stored images of the stimuli represented by collections of locally averaged intensity values, whereas both the objective specification of the stimuli and their intuitive description involved 3D volumetric primitives. This may be taken as an indication that recognition in the human visual system relies at least to some extent on view interpolation (cf. Poggio and Edelman, 1990; Bülthoff and Edelman, 1992).

Following Biederman (1987), it may be conjectured that using nonaccidental features such as combinations of receptive fields that signal the presence of collinear or parallel contour segments will lead to a better performance at large misorientations than that exhibited by the simple model. Guidelines for enhancing the feature extraction ability of this model can be found in the recent psychophysical data on interactions between spatial filters in human vision (Polat and Sagi, 1993), and in neurobiological evidence for long-range lateral connections in the primary visual area in mammals (Gilbert, 1988; Katz and Callaway, 1992). It remains to be seen whether endowing the two-stage model with nonaccidental feature detectors will enable it to perform above chance at large misorientations, without resorting to an elaborate multistage approach such as Hummel's recent implementation of the Recognition By Components theory (Hummel and Biederman, 1992).

[4] The value of σ was computed as follows. First, the mean separation S of the vectors in the entire ensemble of inputs was computed (this is the value of σ recommended by (Saha and Keeler, 1990)). Second, σ was set to either 0.5, 1.0, or 2.0 times S.

4 Conclusions and future work

The present work studied the interaction between class similarity and viewpoint dependence in recognition. Intuitively, one expects an increased sensitivity to viewpoint in the discrimination between highly similar objects, if these can only be distinguished by paying attention to small details that change appearance or become occluded when the objects rotate in space. This intuition was a major motivation in the development of the features of recognition framework, and is supported by the psychophysical results presented above. These results, however, clearly warrant further work. The following open issues seem to be the most important at the present stage:

4.1 The contribution of the primary visual areas to recognition

A full functional simulation of the parvo stream in the primary visual areas of the mammalian cortex (V1 and V2), including spatial filters at multiple scales (Wilson and Bergen, 1979), log-polar mapping (Cavanagh, 1985), contour completion (von der Heydt et al., 1984), and lateral interactions between filters (Polat and Sagi, 1993; Edelman, 1992), should be employed in the transduction stage of the model. A success of the resulting model would signify that, in certain tasks, recognition may require little machinery beyond that available in V1 and V2 (the model's second stage – interpolating classifier – can be implemented as well by a weighted sum of receptive fields such as those found in the primary visual cortex).

4.2 The structure of the psychological representation space

The modeling approach adopted here assumed that views of 3D objects are represented by points in a multidimensional metric space of the activities of a collection of receptive fields. Psychophysical methods such as multidimensional scaling (Kruskal and Wish, 1978) should be employed to explore the structure of this space.

4.3 Psychology of decision-making in recognition

Another assumption was that the identity of the input view is decided by comparison with internally represented views, according to a variant of the nearest-neighbor criterion. Some of the relevant open questions here are concerned with the details of the decision criteria adopted by human subjects, and with the influence of learning and of the stimulus parameters on these criteria. Specifically, it would be interesting to determine whether the feature vector for a new input is compared with an explicitly represented decision surface constructed in the feature space, or with representations of previously encountered exemplars corresponding to the familiar views, or with a set of prototypes each of which stands for a class of familiar objects (Nosofsky, 1988; Edelman, 1993; Maddox and Ashby, 1993).

4.4 Computational analysis

A computational analysis of possible sources of the observed orientation effects in recognition appears in appendix A. The results stated there indicate that the dependence of performance on stimulus orientation can be explained by either one of two computational mechanisms: a single-basis RBF preprocessor, or a Bayesian decision module (both used in conjunction with the nearest-neighbor approach in the representation space). Future work will address computational issues left open by appendix A, and will explore ways to distinguish between the possible explanations psychophysically.

4.5 The generalization of the results to other object classes

The lack of software tools for the automatic generation of object classes jointly parameterized by isomorphic sets of variables hampers the extension of the results reported here to additional objects. One way around this technical difficulty may be to use as stimuli collections of 3D geon-like parts in random configurations (cf. Biederman and Gerhardstein, 1993). If the random objects are evolved via a controlled perturbation from real 3D objects, this approach can also help clarify the role of prior everyday exposure in the recognition process, possibly through the demonstration of an object superiority effect (Weisstein and Harris, 1974) for real objects.

4.6 Summary

The psychophysical results reported in this paper suggest that viewpoint invariance, characteristic of basic-level classification, and viewpoint dependence, a trait of subordinate-level recognition, may be more closely related than previously thought. The possibility of varying the degree of viewpoint dependence of the subjects' performance by manipulating objective similarity between the stimuli indicates that a unified account of recognition, suggested in section 1, may be feasible. Moreover, the extent to which such a unified account may be based on feature extraction coupled with exemplar-based classification appears to depend on further developments of feature extraction methods, beyond the simple approach taken by the simulations described here.

Acknowledgments

I thank Shabtai Barash, Irving Biederman, Heinrich Bülthoff, Nathan Intrator, Yael Moses, Tomaso Poggio, and Shimon Ullman for useful discussions, Devra Goldberg for programming assistance in preparing appendix A, and the subjects of the psychophysical experiments for their time and patience. The work described here is supported by a grant from the Basic Research Foundation, administered by the Israel Academy of Arts and Sciences.

References

Amari, S. (1978). Feature spaces which admit and detect invariant signal transformations. In *Proc. 4th Intl. Conf. Pattern Recognition*, pages 452–456, Tokyo.

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

Biederman, I. and Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: evidence and conditions for 3D viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, 19. in press.

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.

Bülthoff, H. H. and Mallot, H. A. (1988). Interaction of depth modules: stereo and shading. *Journal of the Optical Society of America*, 5:1749–1758.

Cavanagh, P. (1985). Local log polar frequency analysis in the striate cortex as a basis for size and orientation invariance. In Rose, D. and Dobson, V. G., editors, *Models of the visual cortex*, pages 146–157. Wiley, New York, NY.

Edelman, S. (1991). Features of recognition. CS-TR 91-10, Weizmann Institute of Science.

Edelman, S. (1992). Representing 3D objects by sets of activities of receptive fields. CS-TR 92-19, Weizmann Institute of Science. Biol. Cybern. 1993, in press.

Edelman, S. (1993). Representation, similarity, and the chorus of prototypes. CS-TR 93-10, Weizmann Institute of Science.

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 Poggio, T. (1992). Bringing the Grandmother back into the picture: a memory-based view of object recognition. *Int. J. Pattern Recog. Artif. Intell.*, 6:37–62.

Edelman, S., Reisfeld, D., and Yeshurun, Y. (1992). Learning to recognize faces from examples. In Sandini, G., editor, *Proc. 2nd European Conf. on Computer Vision, Lecture Notes in Computer Science*, volume 588, pages 787–791. Springer Verlag.

Flannagan, M. J., Fried, L. S., and Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 12:241–256.

Fried, L. S. and Holyoak, K. J. (1984). Induction of category distributions: a framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10:234–257.

Gilbert, C. D. (1988). Neuronal and synaptic organization in the cortex. In Rakic, P. and Singer, W., editors, *Neurobiology of Neocortex*, pages 219–240. Wiley, New York, NY.

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

Hofstadter, D. R. (1985). *Metamagical themas*. Viking, Harmondsworth, England.

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

Jolicoeur, P. (1990). Identification of disoriented objects: a dual-systems theory. *Mind and Language*, 5:387–410.

Katz, L. C. and Callaway, E. M. (1992). Development of local circuits in mammalian visual cortex. *Ann. Rev. Neurosci.*, 15:31–56.

Knill, D. C. and Kersten, D. (1991). Ideal perceptual observers for computation, psychophysics and neural networks. In Watt, R., editor, *Vision and visual dysfunction*, volume 14, chapter 7, pages 83–97. Macmillan, London.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling.* Sage Piblications, Beverly Hills, CA.

Maddox, W. T. and Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics*, 53:49–70.

Moody, J. and Darken, C. (1989). Fast learning in networks of locally tuned processing units. *Neural Computation*, 1:281–289.

Moses, Y. and Ullman, S. (1992). Limitations of non model-based recognition schemes. In Sandini, G., editor, *Proc. 2nd European Conf. on Computer Vision, Lecture Notes in Computer Science*, volume 588, pages 820–828. Springer Verlag.

Nishihara, H. K. and Poggio, T. (1984). Stereo vision for robotics. In Brady, J. M. and Paul, R., editors, *Robotics research: the first international symposium*, pages 489–505. MIT Press, Cambridge, MA.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14:700–708.

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

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

Poggio, T., Fahle, M., and Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science*, 256:1018–1021.

Polat, U. and Sagi, D. (1993). Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments. *Vision Research*, 33:993–997.

Rock, I. and DiVita, J. (1987). A case of viewer-centered object perception. *Cognitive Psychology*, 19:280–293.

Saha, A. and Keeler, J. D. (1990). Algorithms for better representation and faster learning in Radial Basis Function networks. In Touretzky, D., editor, *Neural Information Processing Systems*, volume 2, pages 482–489. Morgan Kaufmann, San Mateo, CA.

Snippe, H. P. and Koenderink, J. J. (1992). Discrimination thresholds for channel-coded systems. *Biological Cybernetics*, 66:543–551.

Snodgrass, J. G. and Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6:174–215.

Tarr, M. and Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21:233–282.

Ullman, S. (1989). Aligning pictorial descriptions: an approach to object recognition. *Cognition*, 32:193–254.

Ullman, S. and Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1005.

von der Heydt, R., Peterhans, E., and Baumgartner, G. (1984). Illusory contours and cortical neurons' responses. *Science*, 224:1260–1262.

Weiss, Y. and Edelman, S. (1993). Representation with receptive fields: gearing up for recognition. CS-TR 93-09, Weizmann Institute of Science.

Weisstein, N. and Harris, C. S. (1974). Visual detection of line segments: an object-superiority effect. *Science*, 186:752–755.

Wilson, H. R. and Bergen, J. R. (1979). A four mechanism model for threshold spatial vision. *Vision Research*, 19:19–32.

A A computational formulation of the invariance / diagnosticity tradeoff

The tradeoff between relative invariance with respect to the viewing position supported by a given choice of features, and the degree of discrimination between objects that these features allow has been predicted in (Edelman, 1991) as a possible manifestation of the unity of the mechanisms underlying basic and subordinate-level recognition. In this section, I explore two possible general approaches to the analysis of this tradeoff. To simplify the analysis, it is assumed that individual views are represented as points $\mathbf{x} \in R^k$, and that the decision mechanism is a variant of the nearest-neighbor scheme.

Let $\mathbf{x}_{0,1}^{(A)}$ be any two distinct views of object A, and $\mathbf{x}_0^{(B)}$ be an arbitrary view of object B. Denote the action of the feature extraction stage by a vector-valued function $\mathbf{f}(\mathbf{x}) : R^k \rightarrow R^k$. Then, for the feature extraction process to lead to a gain in viewpoint invariance, it is required that

$$d\left(\mathbf{f}\left(\mathbf{x}_0^{(A)}\right), \mathbf{f}\left(\mathbf{x}_1^{(A)}\right)\right) < d\left(\mathbf{x}_0^{(A)}, \mathbf{x}_1^{(A)}\right) \tag{1}$$

where d is the metric on R^k used in the nearest-neighbor scheme. At the same time, for the feature extraction to lead to an increase in object discriminability, it is required that

$$d\left(\mathbf{f}\left(\mathbf{x}_0^{(A)}\right), \mathbf{f}\left(\mathbf{x}_0^{(B)}\right)\right) > d\left(\mathbf{x}_0^{(A)}, \mathbf{x}_0^{(B)}\right) \tag{2}$$

Generally, the feature space will be of a different dimensionality than the input space. Therefore, care must be taken when distances before feature extraction are compared to those after feature extraction: they may fall in different ranges merely because of the different dimensionalities. In section A.1 this problem is avoided by normalizing the distances before comparison.

Assuming that views are represented by orthographic projections of spatially localized features, and using the notion of reachability of projections of 3D objects defined in (Moses and Ullman, 1992), one can show that requirements 1 and 2 cannot be simultaneously satisfied for all views of all objects. This result, however, seems to be too weak to be of any practical significance: it would be more useful to find out, for example, whether the two conflicting requirements can still be satisfied simultaneously on the average, or for a majority of viewpoints.

A.1 The tradeoff as a by-product of the RBF classification

Let us assume now that that the nearest-neighbor classification stage is preceded by an RBF preprocessor, with a single Gaussian basis function, corresponding to the single stored view (see Figure 8). In that case, the recognition rate for other views of the same object can be improved merely by increasing the width σ of the basis function. This manipulation, however, will cause an increase in the false alarm rate, that is, in the tendency of the classifier to overgeneralize (Edelman and Poggio, 1992).

***** Figure 8 here *****

To quantify this effect, let $\mathbf{v}_r \in R^{2k}$, $\mathbf{v}_t \in R^{2k}$, and $\mathbf{y} \in R$ be, respectively, the reference view stored as the RBF center, a test view of the same object, and the output of the basis unit (k is the number of features in the object). We can compare the effect of the misorientation relative to the reference view on two distances, one computed before and the other one after the RBF stage. The "raw" distance is

$$d^2_{raw} = \|\mathbf{v}_r - \mathbf{v}_t\|^2 = \|P\mathbf{x}_r - PT(\alpha)\mathbf{x}_r\|^2 \tag{3}$$

where $\mathbf{x}_r \in R^{3k}$ is the feature vector for the reference view before projection, $P : R^{3k} \rightarrow R^{2k}$ is the orthographic projection matrix, and $T(\alpha)$ is the transformation matrix corresponding to a rotation by α with respect to the reference orientation.[5] The distance after the RBF stage is

$$d^2_{RBF} = (y_r - y_t)^2 = \left(\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\|P\mathbf{v}_r - PT(\alpha)\mathbf{v}_r\|}{\sigma}\right)^2}\right)^2 \tag{4}$$

where σ is the width of the Gaussian basis function used in the RBF stage. Note that d_{RBF} starts at 0 for $|\alpha| = 0$ and asymptotes at $1/\sqrt{2\pi}\sigma$ for $|\alpha| \rightarrow \pi$. Both the asymptotic value of d_{RBF} and its rate of change are influenced by σ. To assess this influence numerically, ten 10-vertex "objects" consisting of clouds of unlabeled feature points in 3D were created by choosing the x, y, and z coordinates of each feature independently and randomly according to a uniform probability density from the $[-10.0,\ 10.0]$ interval. Given those objects, a plausible value for σ can be set, e.g., by requiring that $d_{RBF}|_{|\alpha|=\pi/4} = 0.5 d_{RBF}|_{|\alpha|=\pi}$. A numerical solution of this equation yields $\sigma = 9.6$.[6]

One can now estimate the average effects of the object orientation α on the two distances, d_{raw} and d_{RBF}, by computing the appropriate partial derivatives and integrating over the relevant range of orientations, say, $[0..\pi/2]$ radians. Let

$$\mathcal{D}_{RBF} = \frac{\int_0^{\pi/2} \frac{\partial d_{RBF}}{\partial \alpha}(\sigma, \alpha)\, d\alpha}{\int_0^{\pi/2} d_{RBF}(\sigma, \alpha)\, d\alpha} \tag{5}$$

$$\mathcal{D}_{raw} = \frac{\int_0^{\pi/2} \frac{\partial d_{raw}}{\partial \alpha}(\sigma, \alpha)\, d\alpha}{\int_0^{\pi/2} d_{raw}(\sigma, \alpha)\, d\alpha} \tag{6}$$

Numerical evaluation of these expressions at the chosen value of $\sigma = 9.6$ yields $\mathcal{D}_{RBF} = 1.29 < \mathcal{D}_{raw} = 1.76$, that is, on the average, the dependence of d_{RBF} on misorientation to training is smaller than the dependence of d_{raw} (here and below all the numerical results are averages over the ten test objects).

***** Figure 9 here *****

How do the changes in σ affect the viewpoint dependence of a system that uses a single-center RBF preprocessor? Figure 9 shows a plot of the partial derivative $\frac{\partial d_{RBF}(\alpha,\sigma)}{\partial \alpha}$ vs. σ and $|\alpha|$. One can observe that for the chosen value of $\sigma = 9.6$ the viewpoint dependence of the system's

[5]The components of α can be, e.g., the Euler angles encoding the object's orientation. The object's misorientation relative to an arbitrary reference viewpoint can be measured by a single rotation around an axis whose orientation in space can be defined, but is of no importance if the object possesses no intrinsic or natural orientation due, e.g., to the presence of a major axis of symmetry. This single rotation is denoted in what follows as $|\alpha|$.

[6]The method for choosing σ based on average inter-object distance, mentioned in section 3.2, is less appropriate here, because each object is assumed here to be recognized by a separate module, as in (Poggio and Edelman, 1990).

performance increases with an decrease in σ. Recall that the value of σ affects the false alarm rate of the system, and assume that the system can adapt to changing conditions by adjusting this value. If the false alarm rate grows (e.g., due to an increased similarity between the objects, as in the transition between FAR and NEAR conditions in the psychophysical experiments), it can be reduced by decreasing the value of σ. According to Figure 9, this would cause an increase in viewpoint dependence of the system's performance, similar to what was found in the psychophysical experiments.

A.2 The tradeoff as a by-product of the decision-making step

The invariance/diagnosticity tradeoff may also arise as a by-product of the decision-making process. In deciding whether two views that produced two given feature vectors belong to the same object, the significance of the difference between the feature vectors must be assessed. Two factors supply the reference against which this significance is to be judged: (1) the intrinsic variability of exemplars within each of the object classes, and (2) the differences between the classes. Note that the first of these factors was present in the experiments reported above, but was kept constant. The second factor, however, was varied, and its variation was associated with a change in the percentage of correct responses, and a concomitant change in the degree of viewpoint invariance. From the preceding discussion, it may be expected that reducing the distance between object classes will make changes in the feature vector caused by a shift in the viewpoint relatively more prominent (this is indeed what was found in the psychophysical experiments).

This situation can be analyzed using an ideal-observer approach (see, e.g., Knill and Kersten, 1991). Let $\mathbf{v}_t$ be a test view that the observer is to classify as belonging either to an object whose reference view is $\bar{\mathbf{v}}_1$, or to another object whose reference view is $\bar{\mathbf{v}}_2$. Let us assume that the observer's decision, say, in favor of the first object, is based on the likelihood ratio:[7]

$$R = \frac{P(\bar{\mathbf{v}}_1|\mathbf{v}_t)}{P(\bar{\mathbf{v}}_2|\mathbf{v}_t)} = R \cdot R_{prior} = \frac{P(\mathbf{v}_t|\bar{\mathbf{v}}_1)}{P(\mathbf{v}_t|\bar{\mathbf{v}}_2)} \cdot \frac{P(\bar{\mathbf{v}}_1)}{P(\bar{\mathbf{v}}_2)} \tag{7}$$

In the present psychophysical experiments, the prior probabilities of the appearance of each of the two objects were the same, so that we only need to consider the ratio

$$L = \frac{P(\mathbf{v}_t|\bar{\mathbf{v}}_1)}{P(\mathbf{v}_t|\bar{\mathbf{v}}_2)} = \frac{P_1}{P_2} \tag{8}$$

that is, the probability of obtaining view $\mathbf{v}_t$ from object 1, divided by the probability of obtaining the same view from object 2. Assuming that a test view is attributed by the observer to the object whose prototypical view is the closest to it, we can compute P_1 and P_2:

$$\begin{aligned} P_1 &= \int_S p(\mathbf{v}, \bar{\mathbf{v}}_1; \sigma_n, T)\, d\mathbf{v} && (9) \\ P_2 &= \int_S p(\mathbf{v}, \bar{\mathbf{v}}_1 + \mathbf{d}; \sigma_n, T)\, d\mathbf{v} && (10) \end{aligned}$$

[7]Evidence to the effect that subjects assign an exemplar to a given category on the basis of its likelihood of having been generated by the category can be found in (Fried and Holyoak, 1984). For a more general treatment of likelihood ratios in the contexts of perceptual decisions, see (Green and Swets, 1966).

where σ_n is the standard deviation of the normally distributed noise in the parameter space, T represents the effect of object rotation, and the region of integration is $S = \{\mathbf{v}\ s.t.\ |\mathbf{v} - \bar{\mathbf{v}}_1| < |\mathbf{v} - \bar{\mathbf{v}}_2|\}$ (see Figure 10). Note that in equation 10 $\bar{\mathbf{v}}_1 + \mathbf{d}$ is substituted for $\bar{\mathbf{v}}_2$.

***** Figure 10 here *****

***** Figure 11 here *****

We can now use Monte-Carlo integration to estimate the dependence of $L = P_1/P_2$ on the variance in object appearance, and how it is influenced by the dissimilarity between the two object classes. In what follows, no distinction is made between the two factors that influence object appearance: parameter variation and orientation. The reason for this is that both these factors (modeled here by noise, normally distributed around $\bar{\mathbf{v}}_1$ or $\bar{\mathbf{v}}_1$, with std.dev.$=\sigma_n$) have the same qualitative influence on L; the quantitative details are irrelevant at the present level of analysis. The dissimilarity is manipulated below by blending object prototypes, as in the psychophysical experiments (see section B.1). Let α be the blending parameter that controls the distance between object prototypes:

$$\bar{\mathbf{v}}_1' = (1-\alpha)\bar{\mathbf{v}}_1 + \alpha\bar{\mathbf{v}}_2 \tag{11}$$

$$\bar{\mathbf{v}}_2' = \alpha\bar{\mathbf{v}}_1 + (1-\alpha)\bar{\mathbf{v}}_2 \tag{12}$$

According to the above definition, for $\alpha = 0$ the prototypes $\bar{\mathbf{v}}_1', \bar{\mathbf{v}}_2'$ are at their "original" separation, determined by $\bar{\mathbf{v}}_1, \bar{\mathbf{v}}_2$, while for $\alpha = 0.5$ they become equal: $\bar{\mathbf{v}}_1' = \bar{\mathbf{v}}_2'$. Figure 11 (left) shows a plot of the likelihood ratio L vs. α and σ_n. The likelihood ratio is large (appears clipped in the plot) for small α and small σ_n, and decreases when the values of both these parameters increase. A contour plot of a polynomial fit to $L(\alpha, \sigma_n)$ shows the rate of the fall-off of L with increasing σ_n to be higher for high values of α. That is, the influence of the factors that contribute to image appearance variability (noise and rotation) grows when the prototype objects become more similar to each other.

B Parameterization of the stimuli objects

B.1 3D graphics tools for the study of object representation

The features of recognition framework (see section 1) predicted a tradeoff between the discriminative power and the degree of viewpoint invariance of the basic features used in object representation. A crucial component in an experimental demonstration of this tradeoff is control over similarity between different stimuli. Such control is easily achieved, e.g., for the wire-like stimuli of Edelman and Bülthoff (1992). However, the wire objects, all of which belong to the same basic category, cannot serve as stimuli in an experiment that is to address the issue of basic-level classification.

***** Figure 12 here *****

Smooth control over shape is possible even for complex objects, if these are appropriately parameterized (cf. Hofstadter, 1985, p.241). Given a parameterization of two objects that can be described by two points $\mathbf{x}_1, \mathbf{x}_2 \in R^n$ (where R^n is the n-dimensional parameter space), an object that is a blend of the two can be defined as $\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2$, where α is the blending constant. If this linear combination is convex (that is, if $0 \le \alpha \le 1$), then the blended object is, in a sense, "in between" the two original ones. Allowing the parameters to vary randomly around the central values of $\mathbf{x}_1, \mathbf{x}_2$ (e.g., according to a normal distribution with moderate variance) leads to objects that are random variations on the central or prototypical themes.

Figure 1 shows a family of images of 3D objects that were obtained with the procedure outlined above and were used in the psychophysical experiments. There were two classes of objects, one resembling monkeys, and the other dogs (according to Snodgrass and Vanderwart 1980, monkey and dog belong to separate basic-level categories). Both classes were defined by the same set of 56 parameters, which encoded sizes, shapes, and placement of the limbs, the ears, the snout, etc. Applying to the prototypical (central) members of the two classes the blending formula with α changing by small steps between 0 and 1 caused the resulting object to change its shape smoothly between that of a monkey and that of a dog.

B.2 A list of the parameters used in creating object shapes

To illustrate the parametric relationship between the two object classes used in the experiments, monkey and dog, this section lists the names and the values of the 56 parameters that defined the object prototypes.

```
        Parameter Name          #       monkey      dog
-------------------------------------------------------
#define SIZE                    0        4.0        4.0
#define HEAD_LENGTH             1        0.2        0.2
#define HEAD_ECCENTRICITY_1     2        0.8        0.8
#define HEAD_ECCENTRICITY_2     3        1.0        1.0
#define SNOUT_LENGTH            4        0.1        0.15
#define SNOUT_ECCENTRICITY_1    5        1.3        0.5
#define SNOUT_ECCENTRICITY_2    6        1.5        0.8
#define EAR_LENGTH              7        0.01       0.015
#define EAR_ECCENTRICITY_1      8       10.0       10.0
#define EAR_ECCENTRICITY_2      9        8.0        5.0
#define NECK_LENGTH            10        0.4        0.3
#define NECK_RADIUS            11        0.6        0.06
#define NECK_TAPER             12        1.3        1.3
#define TAIL_LENGTH            13        0.8        0.6
#define TAIL_RADIUS            14        0.03       0.1
#define TAIL_TAPER             15        0.5        0.2
#define EYE_SIZE               16        0.02       0.02
#define EYE_THETA              17        3.6        3.6
#define EYE_PHI                18       -1.0       -1.5
#define THIGH_LENGTH           19        0.7        0.5
#define THIGH_RADIUS           20        0.2        0.16
#define THIGH_TAPER            21        0.5        0.5
#define LEG_LENGTH             22        0.5        0.25
#define LEG_RADIUS             23        0.1        0.08
#define LEG_TAPER              24        0.5        0.5
```

```
#define FOOT_LENGTH             25        0.07       0.05
#define FOOT_ECCENTRICITY_1     26        1.8        1.8
#define FOOT_ECCENTRICITY_2     27        1.2        1.2
#define SHOULDER_LENGTH         28        0.35       0.45
#define SHOULDER_RADIUS         29        0.1        0.11
#define SHOULDER_TAPER          30        0.5        0.5
#define ARM_LENGTH              31        0.5        0.35
#define ARM_RADIUS              32        0.05       0.055
#define ARM_TAPER               33        0.5        0.5
#define HAND_LENGTH             34        0.05       0.05
#define HAND_ECCENTRICITY_1     35        0.5        0.5
#define HAND_ECCENTRICITY_2     36        0.5        0.5
#define CHEST_LENGTH            37        0.5        0.45
#define CHEST_ECCENTRICITY_1    38        0.3        0.3
#define CHEST_ECCENTRICITY_2    39        0.3        0.3
#define ABDOMEN_LENGTH          40        0.5        0.55
#define ABDOMEN_ECCENTRICITY_1 41         0.3        0.3
#define ABDOMEN_ECCENTRICITY_2 42         0.3        0.3
#define CHEST_ABDOMEN_ANGLE     43        0          0
#define NECK_CHEST_ANGLE        44        0          0
#define HEAD_NECK_ANGLE         45       20         10
#define ABDOMEN_THIGH_ANGLE     46      170        110
#define THIGH_LEG_ANGLE         47     -120        -35
#define CHEST_SHOULDER_ANGLE    48       90         80
#define SHOULDER_ARM_ANGLE      49       70         20
#define HEAD_EAR_ANGLE_1        50      -30        -25
#define HEAD_EAR_ANGLE_2        51       15          0
#define ABDOMEN_TAIL_ANGLE      52       45         25
#define SNOUT_ANGLE             53      280        330
#define SNOUT_INSET             54        0.6        0.8
#define LEG_OFFSET              55        1.1        0.8
```

Figure Legends:

Figure 1: A family of images of two classes of parameterized 3D objects, obtained with the blending procedure described in appendix B. The objects were created and rendered using the GL language on a Silicon Graphics 4D/35GT workstation. The illustration shows the two class prototypes, four blended objects, and the effects of random perturbation of parameters (top) and of object rotation (bottom left).

Figure 2: The relationships between the two classes of stimuli objects in the 56-dimensional parameter space (illustrated here schematically as 1-dimensional). *Top:* In experiment 1 the controlled parameter was the distance (NEAR or FAR) between the distributions corresponding to the two classes. *Bottom:* In experiments 2 and 3 the distributions were bimodal, and for each of them one of the modes was fixed in the parameter space.

Figure 3: Experiment 1. *Left:* a schematic illustration of the experimental conditions. *Right:* Percentage of correct responses CR in the low and high inter-class similarity conditions (FAR and NEAR modes; upper and lower curves, respectively), plotted vs. the angular distance D to the training orientation (means and standard errors of five subjects).

Figure 4: Experiment 2. *Left:* a schematic illustration of the experimental conditions. RIGHT: Percentage of correct responses CR in the four experimental conditions (top to bottom curves: outer and inner modes under low inter-class similarity (FAR conditions); outer and inner modes under high inter-class similarity (NEAR conditions). The abscissa represents the angular distance D to the training orientation. Data are means and standard errors of six subjects.

Figure 5: Experiment 3: *Left:* a schematic illustration of the experimental conditions. *Right:* Percentage of correct responses CR for the two similarity conditions, FAR (upper curve) and NEAR (lower curve), plotted vs. the angular distance D to the training orientation. (see section 2.4 for details).

Figure 6: A snapshot of a typical distribution of receptive fields used in the transduction step of the simulated experiment (in this example, there are 150 receptive fields; see section 3.1), overlayed on an image of one of the two stimuli (monkey, shown in Figure 12; the overlay image here has been subjected to edge detection for presentation clarity).

Figure 7: Simulation: percentage of correct responses for the inner and outer subclasses, in the two experimental conditions — low inter-class similarity (FAR; curves marked by "I" and "O"), and high inter-class similarity (NEAR; curves marked by "i" and "o"). The abscissa represents the angular distance to the training orientation. Data are means and standard errors of 9 runs (see text). Compare with Figure 4.

Figure 8: A nearest-neighbor classifier preceded by an "invariance enhancer" module, implemented as an RBF network. Section A.1 analyzes the output of a simplified RBF module consisting of a single basis unit (indicated in the illustration by an arrow). Generally, the output of a vector-valued module is a linear combination of the vector of activities of basis units (Poggio and Edelman, 1990).

Figure 9: A plot of the dependence of performance on viewpoint, as reflected in the value of $\frac{\partial d_{RBF}}{\partial \alpha}(\alpha, \sigma)$. The dependence on viewpoint is seen to increase with decreasing σ around the chosen value of $\sigma = 9.6$, for all relevant values of $|\alpha|$.

Figure 10: A diagram of the decision situation in an experiment that involves discrimination between two objects. For illustration purposes, the space of images of all possible objects is shown here as two-dimensional. The points corresponding to the images of the two object prototypes are denoted by $\bar{\mathbf{v}}_1$ and $\bar{\mathbf{v}}_2$. The task in the experiment is to decide which of the two objects gave rise to the test image, $\mathbf{v}_t$.

Figure 11: *Left:* The likelihood ratio for the decision that the input view belongs to object #1, given that it falls closer to $\bar{\mathbf{v}}_1$ than to $\bar{\mathbf{v}}_2$ (see Figure 10), plotted vs. the blending parameter α and a measure of the variability of the object's appearance σ_n. The data are means over five pairs of objects (the same random objects used in the other simulations in the appendix). Each point in the plot was obtained with a 50-sample Monte-Carlo integration of equations 9 and 10. *Right:* A contour plot of a 3rd-degree bivariate polynomial fit to $L(\alpha, \sigma_n)$. For $\sigma_n < 15$, the density of contours in the direction of increasing σ_n grows with α.

Figure 12: An image of one of the two stimuli objects (the monkey).

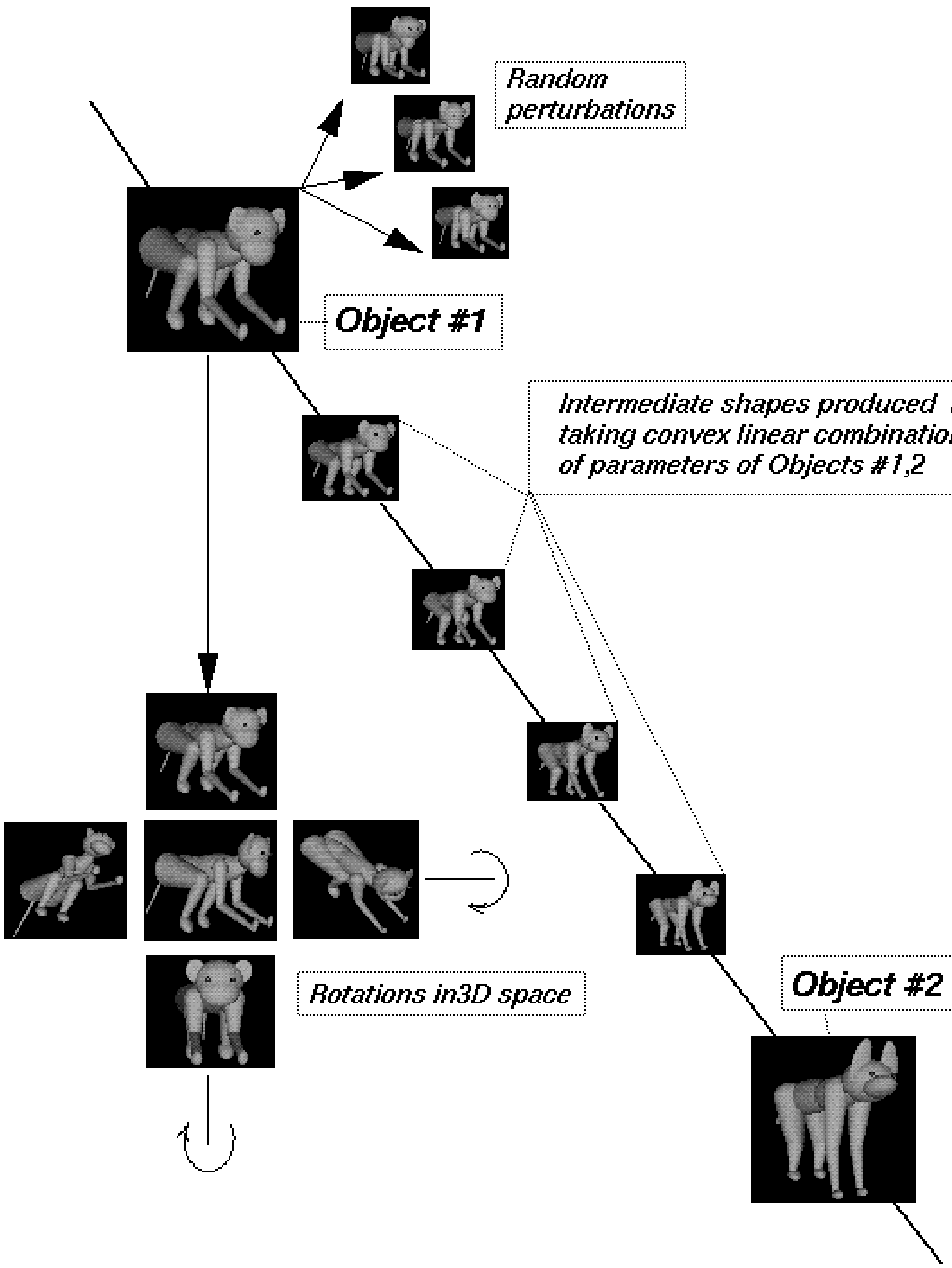

Random perturbations
Object #1
Intermediate shapes produced by taking convex linear combinations of parameters of Objects #1,2
Rotations in3D space
Object #2

Figure 1

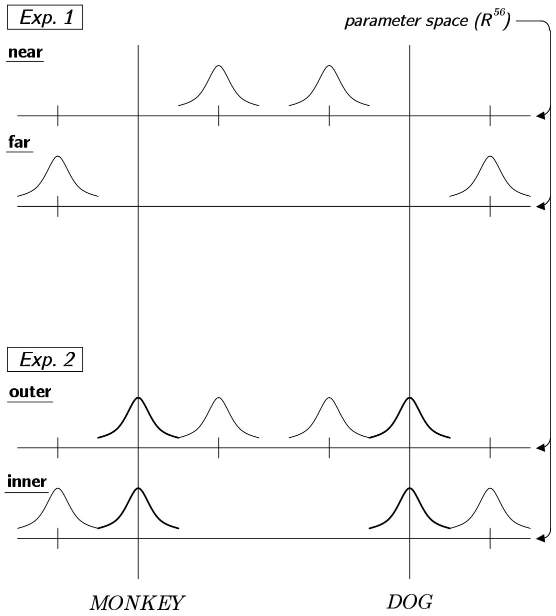
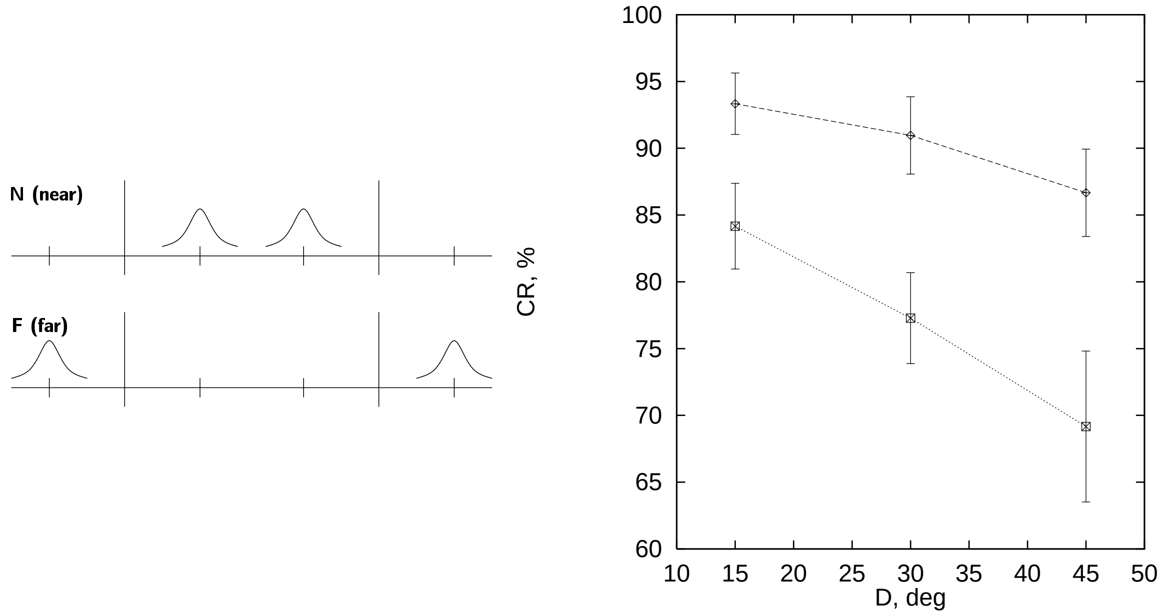

Exp. 1
parameter space (R^{56})
near
far
Exp. 2
outer
inner
MONKEY
DOG

Figure 2



N (near)
F (far)
100
95
90
85
80
75
70
65
60
10
15
20
25
30
35
40
45
50
CR, %
D, deg

Figure 3

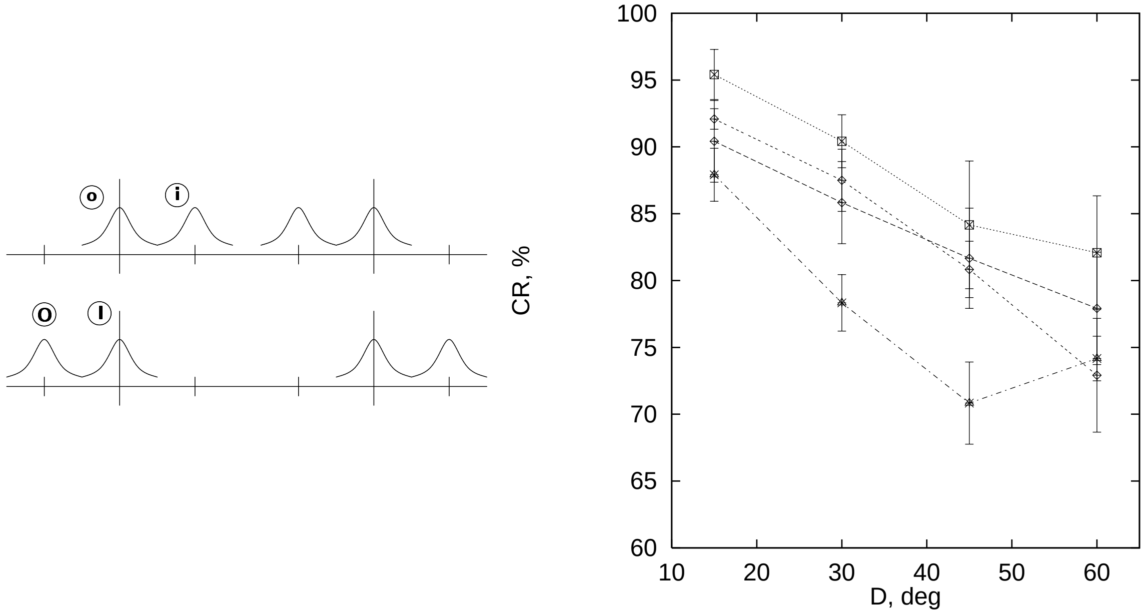


Figure 4

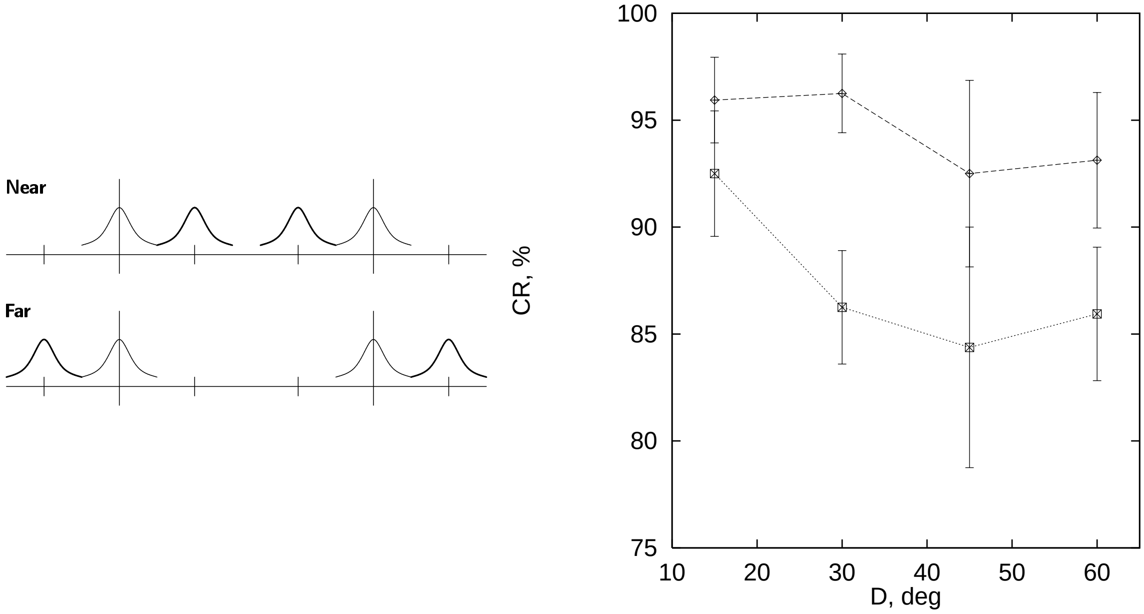


Figure 5

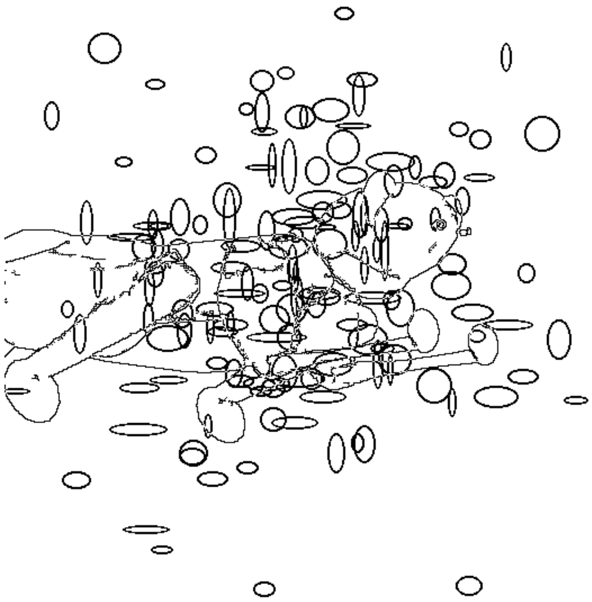
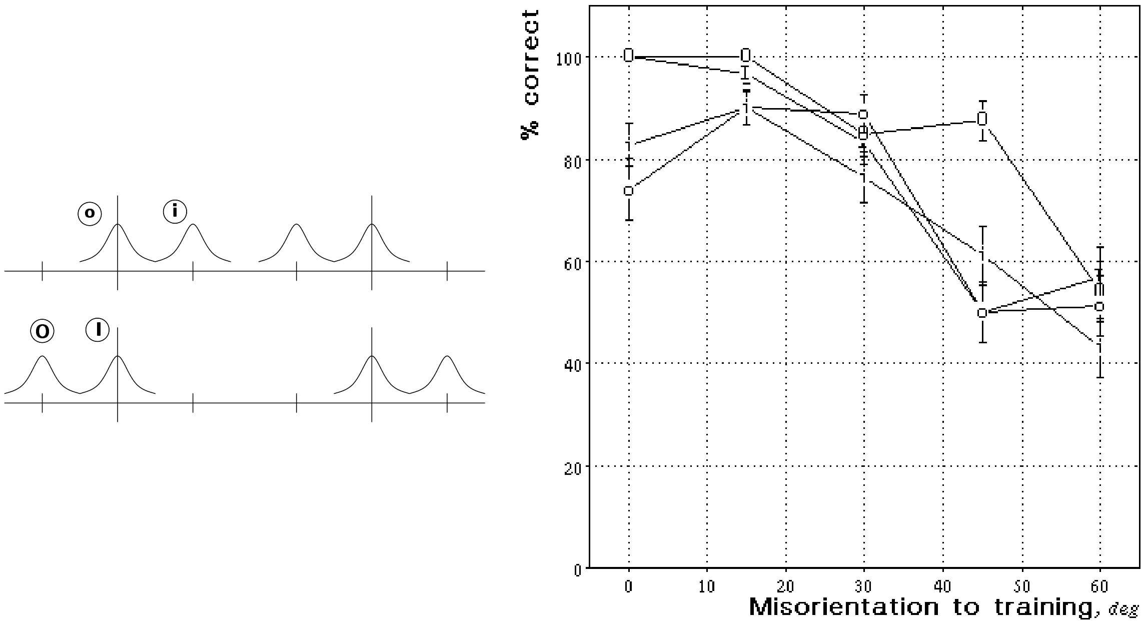
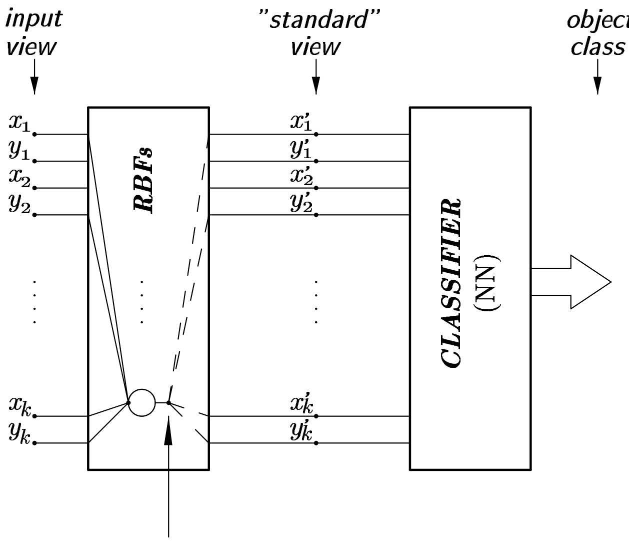


Figure 6



o
i
O
I
% correct
100
80
60
40
20
0
0
10
20
30
40
50
60
Misorientation to training, deg

Figure 7



input
view
"standard"
view
object
class
x_1
y_1
x_2
y_2
x_k
y_k
RBFs
x_1'
y_1'
x_2'
y_2'
x_k'
y_k'
CLASSIFIER
(NN)

Figure 8

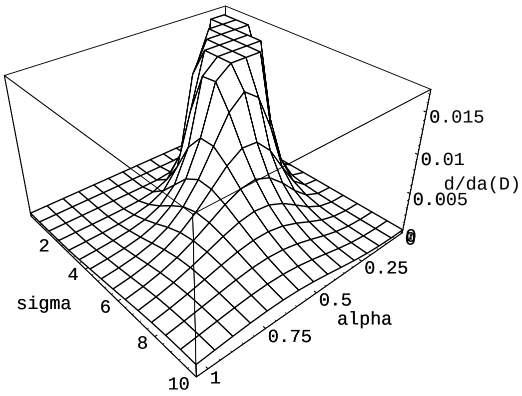


Figure 9

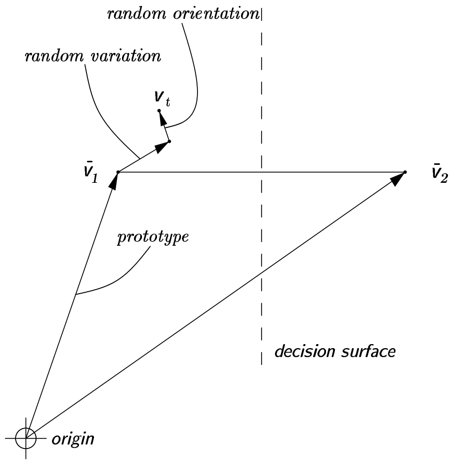

random orientation
random variation
v_t
$\bar{v}_1$
$\bar{v}_2$
prototype
decision surface
origin

Figure 10

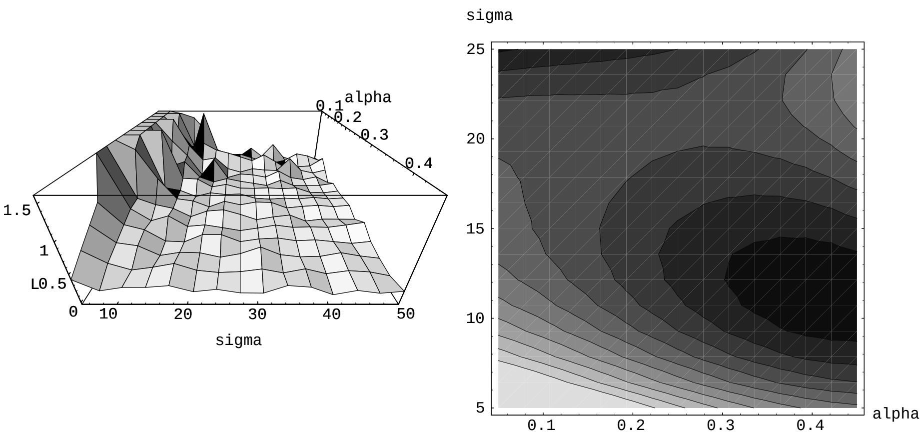

alpha
0.1
0.2
0.3
0.4
1.5
1
L0.5
0
10
20
30
40
50
sigma
sigma
25
20
15
10
5
0.1
0.2
0.3
0.4
alpha

Figure 11

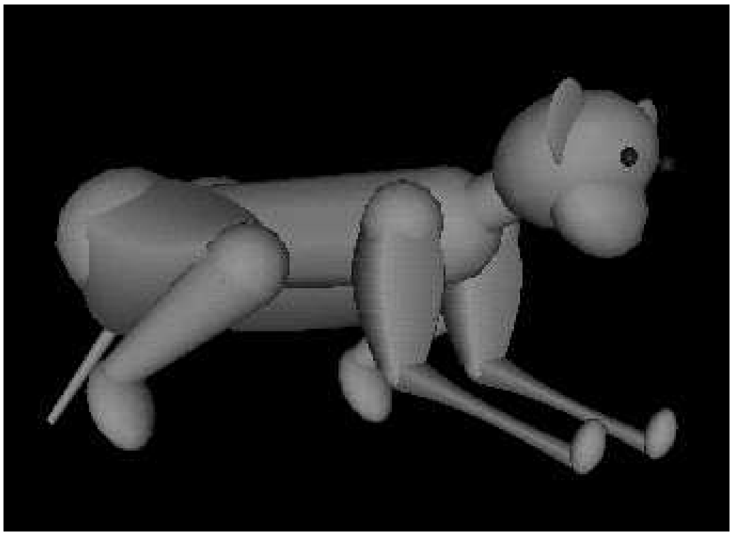


Figure 12