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

A.I. Memo No. 1682
C.B.C.L. Paper No. 185

May 1, 2000

The Individual is Nothing, the Class Everything: Psychophysics and Modeling of Recognition in Object Classes

Maximilian Riesenhuber and Tomaso Poggio

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.

Abstract

Most psychophysical studies of object recognition have focussed on the recognition and representation of individual objects subjects had previously explicitely been trained on. Correspondingly, modeling studies have often employed a “grandmother”-type representation where the objects to be recognized were represented by individual units. However, objects in the natural world are commonly members of a class containing a number of visually similar objects, such as faces, for which physiology studies have provided support for a representation based on a sparse population code, which permits generalization from the learned exemplars to novel objects of that class. In this paper, we present results from psychophysical and modeling studies intended to investigate object recognition in natural (“continuous”) object classes. In two experiments, subjects were trained to perform subordinate level discrimination in a continuous object class — images of computer-rendered cars — created using a 3D morphing system. By comparing the recognition performance of trained and untrained subjects we could estimate the effects of viewpoint-specific training and infer properties of the object class-specific representation learned as a result of training. We then compared the experimental findings to simulations, building on our recently presented HMAX model of object recognition in cortex, to investigate the computational properties of a population-based object class representation as outlined above. We find experimental evidence, supported by modeling results, that training builds a viewpoint- and class-specific representation that supplements a pre-existing representation with lower shape discriminability but possibly greater viewpoint invariance.

This report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. This research is sponsored by a grant from Office of Naval Research under contract No. N00014-93-1-3085, Office of Naval Research under contract No. N00014-95-1-0600, National Science Foundation under contract No. IIS-9800032, and National Science Foundation under contract No. DMS-9872936. Additional support is provided by: AT&T, Central Research Institute of Electric Power Industry, Eastman Kodak Company, DaimlerChrysler Corp., Digital Equipment Corporation, Honda R&D Co., Ltd., NEC Fund, Nippon Telegraph & Telephone, and Siemens Corporate Research, Inc. M.R. is supported by a Merck / MIT Fellowship in Bioinformatics.

1 Introduction

Object recognition performance crucially depends on previous visual experience. For instance, a number of studies have shown that recognition memory for unfamiliar faces is better for faces of one's own race than for faces of other races (the "Other Race-Effect") [14, 18]. Recognition performance for objects belonging to less familiar object classes such as "Greeble" objects [9] is rather poor without special training. These differences in performance are presumably the result of the differing degree of visual experience with the respective object classes: Extensive experience with an object class builds a representation of that object class that generalizes to unseen class members and facilitates their recognition.

Previous object recognition studies involving training on novel stimuli have mostly focussed on training subjects to recognize *individual* isolated objects [3, 9, 15, 31], usually by either familiarizing the subjects with a single object and then testing how well this object could be recognized among distractors [3, 15], or by training subjects on a small number of objects (*e.g.*, by having the subjects learn names for them [9, 31, 33]) and then testing how well these objects could be recognized among distractors. Thus, the objective of training was not to learn an object class (like faces) from which arbitrary novel examplars (unfamiliar faces) could be drawn during testing — rather subjects had to re-recognize the exact same objects as used in training.

An additional problem of several of the aforementioned studies was that the stimuli used belonged to rather artificial object classes such as "cube objects", "paperclips", or "amoebas" (Fig. 1b) differing from naturally occuring object classes — such as, *e.g.*, faces, human bodies, cats, dogs, or cars (Fig. 1a) — in that the objects did not share a common 3D structure (making them "not nice" object classes in the terminology of Vetter *et al.* [35]).

Even in studies where objects of a more natural appearance were used (such as the "Greeble" family of objects [9]), subjects were still trained to recognize individual representatives (*e.g.*, by naming them) whose recognition was later tested under various transformations [9, 21, 33]. Similarly, computational models of object recognition in cortex have almost exclusively focussed on the recognition of individual objects that had been learned explicitely [8, 22, 23, 26, 36]. These computational studies [8, 23, 27, 36] commonly feature an object representation where for each stimulus to be recognized, a unique "grandmother"-type unit is trained to respond to this individual object. While such a scheme (with one or more "grandmother" units per object [35]) may actually be used to represent highly overtrained objects [16] in situations where the subject has to recognize (a small number of) individual objects among a great number of similar distractor objects [3, 15, 16], the inefficiency and inflexibility of such a scheme makes it highly unlikely to be used in cortex to represent natural object classes.

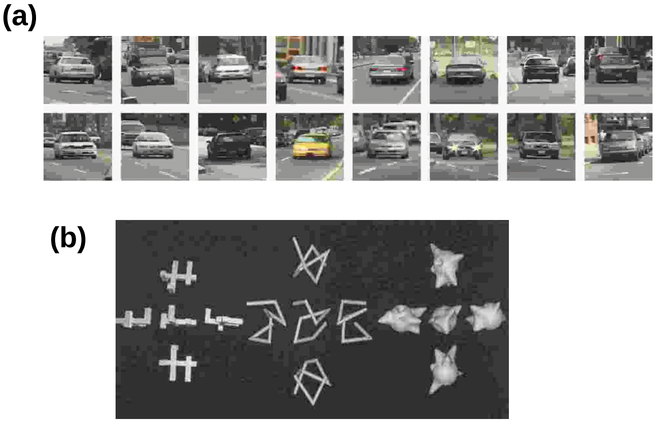

Figure 1: Natural objects, and artificial objects used in previous object recognition studies. **(a)** Members of natural object classes, such as pedestrians (not shown) and cars, usually share a common 3D structure, whereas stimuli popular in previous psychophysical studies of object recognition (from [4]), **(b)**, do not.

A different possibility to represent objects is a scheme where *a group of units, broadly tuned to representatives of the object class, code for the identity of a particular object by their combined activation pattern*. There exists some experimental evidence that is compatible with such a representation: Recordings from neurons in inferotemporal cortex (IT), a brain area believed to be essential for object recognition [17, 30], suggest that facial identity is represented by such a sparse, distributed code [38]. This is further supported by an optical imaging study in IT [37] that indicated an area of neurons selective for face stimuli.

Few studies,[*] experimental or theoretical, have investigated viewpoint-dependent recognition in a principled way for the more general (and natural) case of object *classes*, where training objects are used to build a distributed class representation that is then probed during testing using randomly chosen objects from the same class.[†]

[*] Edelman [5] in a recent study used simple classes (Gaussian blobs in parameter space) of geon-based "dog" and "monkey" stimuli. However, the focus of that study was object *categorization*.

[†] In a recent study, Tarr and Gauthier [33] trained subjects (in a naming task) to recognize a small number of individual objects seen from a single viewpoint. Subjects were then trained on additional viewpoints for a subset of the training objects. Subsequently, it was tested how recognition performance for rotated views transferred to the training objects that had only been seen at one viewpoint during training (the "cohort" objects). As the number of "cohort" and training objects was rather small (4–6 objects), however, it is unclear whether subjects actually learned a representation of the whole object class. Furthermore, two of the three experiments in [33] used

For the purpose of this paper we informally define a *continuous object class* as a set of visually similar objects in terms of 3D structure, that span a multidimensional space, *i.e.*, there is a continuous parametric representation of that class and objects can have arbitrarily similar shapes. Vetter and Blanz [34] (see also [11]), for instance, have shown that human faces can be well represented in such a way. This definition is related to the "nice" object classes of Vetter *et al.* [35]. Here, we stress the *shape* similarity of objects in the class, where two members can be arbitrarily similar to each other, which is of primary interest for recognition and discrimination.

The aim of this paper is to investigate if and how the results on view-dependent object recognition obtained for the individual object case and a "grandmother" representation transfer to continuous object classes represented through a distributed population code. We will first present results from a psychophysical study intended to investigate this question, in which subjects were trained to perform subordinate level discrimination in a continuous object class — images of computer-rendered cars — created using a 3D morphing system [29]. By comparing the recognition performance of trained and untrained subjects we can estimate the effects of viewpoint-specific training and infer properties of the object-class specific representation learned as a result of training.

We will then compare the experimental findings to simulations, building on our recently presented HMAX model of object recognition in cortex [27, 28, 32], to investigate the computational properties of a population-based object class representation as outlined above. In the process, we will demonstrate that the recognition performance of HMAX previously demonstrated for the class of paperclip objects is not special to this class of objects but also transfers to other object classes.

A second experiment was designed to test the model predictions and to investigate the viewpoint-dependency of object recognition in more detail. The results of this study will be compared to simulation results in the last section.

2 Experiment 1

Several psychophysical studies have reported above-chance recognition rates for up to $45°$ (and beyond [5]) viewpoint differences between sample and test object after presenting the sample object at a single viewpoint, for paperclip objects [3, 15] as well as geon-based dog and monkey stimuli [5]. However, these experiments controlled target/distractor similarity — which strongly influences recognition performance [5, 21, 33] — only very roughly (in two levels, [5]) or not at all ([3, 15]). Even more crucial, *these studies did not compare the recognition performance of trained subjects to naive subjects.* Hence, it is unclear how much of the recognition performance was due to training and how much was due to a pre-existing representation not specific to the class of training objects.

The aim of Experiment 1 was to train subjects on a recognition task involving stimuli chosen from a precisely defined continuous object class, presenting objects always at the same viewpoint, and then to probe this representation by testing recognition performance for varying viewpoints and match/nonmatch object similarities. The results of the trained group are compared to the performance of a naive group that did not receive any training on the object class prior to testing.

2.1 Methods

2.1.1 A Continuous Object Class: Morphed Cars

Stimuli for both experiment and modeling were generated using a novel automatic, 3D, multidimensional morphing system developed by Christian Shelton in our lab [29]. With this system we were able to create a large set of "intermediate" objects, made by blending characteristics of the different prototype objects (Viewpoint DataLabs, UT) spanning the class. This was done by specifying how much of each prototype the object to be created should contain, naturally defining a vector space over the prototype objects. Correspondences have been calculated for a system based on eight car prototypes (the "8 car system") and subsequently for 15 car prototypes (the "15 car system"). Thus, an advantage of the morphing system is that it allows *multi*dimensional morphing, *i.e.*, the creation of objects that are made up of mixtures of *several 3D* prototypes. Moreover, as the prototypes are three-dimensional graphics objects, morphed objects can be freely transformed, *e.g.*, through viewpoint or illumination changes.

In the initial morphing studies, we used the 8 car system, whose prototypes are shown in Fig. 2. While the prototypes are available as color models we chose to render all objects as "clay" models by setting the colors to gray values and decreasing surface reflectance (C. Shelton, personal communication). Objects were rendered with a lighting source located above the camera and equally strong ambient lighting, and normalized in size. This procedure was designed to reduce the influence of possibly confounding color and size cues in the experiment.

Stimulus space. Stimuli in Experiment 1 were drawn from a subspace of the 8 car system, a two-dimensional space spanned by the three marked prototypes shown

"cube" objects, which, as mentioned above, are not a good model for natural object classes.

[§] As monochrome printers produce gray values by dithering, these and the other grayscale images print best on a color printer.

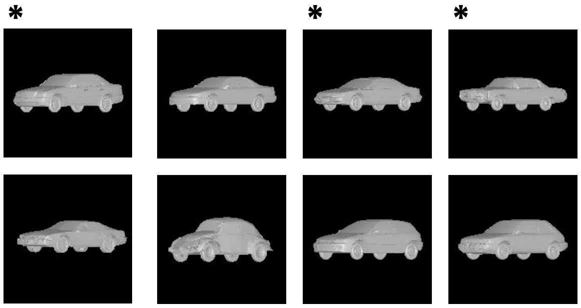

Figure 2: The eight prototype cars used in the 8 car system. The cars marked with an asterisk show the prototypes spanning the morph space used in Experiment 1.[§]

in Fig. 2 (the space spanned by three prototypes is two-dimensional since coefficients have to sum to one).

The advantage of using a well-defined object class spanned by few prototypes is that the class can be exhaustively covered during training and its extent is well known, which is not the case, *e.g.*, for the class of human faces.

2.1.2 Psychophysical Paradigm

Figures 3 and 5 illustrate the training and testing tasks, respectively. They follow a two alternative forced-choice (2AFC) design, with the two choices presented sequentially in time. The advantage of using such a task is that subjects only have to decide which of the two choices resembles the previously encountered sample stimulus more closely, thereby eliminating the influence of biases on the decision process that are associated with a yes/no task in which only one choice stimulus is presented. An additional important advantage of the 2AFC paradigm is that it transfers to simulations in a straightforward way.

Subjects sat in front of a computer monitor at a distance of about 70cm, with the stimuli subtending about 3 degrees of visual angle (128×128 pixels). Each trial was initiated by the appearance of the outline of a blue square (about 5° of visual angle) on the screen, at which time subjects had to push a button on a computer mouse to initiate the trial. Immediately after the button push, a randomly selected (see below) car appeared on the screen for 500ms, followed by a mask consisting of a randomly scrambled car image, presented for 50ms. After a delay of 2000ms, the first choice car appeared in the same location as the sample car, for 500ms, followed by a 500ms delay and the presentation of the second sample car. After the presentation of the second car, the outline of a green square appeared, cueing subjects to make a response (by pressing a mouse button), indicating whether the first (left button) or the second (right button) choice car was equal to the sample car. In the training task (in Experiment 1 as well as Experiment 2, see below), subjects received auditory feedback on incorrect responses.

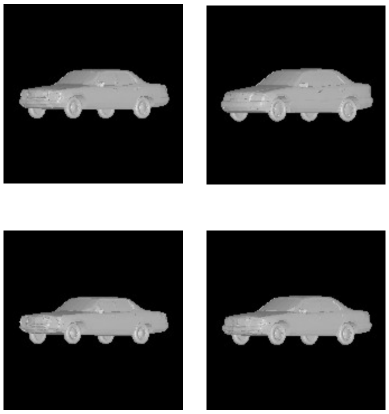

Figure 4: Illustration of match/nonmatch object pairs for Experiment 1. The top shows a pair a distance $d = 0.6$ apart in morph space while the lower pair is separated by $d = 0.4$.

In the training task, sample and test objects were all presented at the same 225° viewpoint on all trials, a 3/4 view (termed the *training view, TV*). New, randomly chosen target (sample) and distractor objects were chosen on each trial by picking coefficient vectors from a uniform distribution followed by subsequent coefficient sum normalization. The purpose of the training task was to induce subjects to build a detailed viewpoint-specific representation of the object class. The (Euclidean) distance d in morph space between target and distractor (nonmatch) objects was decreased over the course of training: Initially, distractor objects were chosen to be very dissimilar to the target objects, $d = 0.6$, making the task comparatively easy (Fig. 4, top). Subjects performed trials at this level of task difficulty until performance reached 80% correct. Then d was decreased by 0.1 and the training repeated with new stimuli down to $d = 0.4$ (Fig. 4, bottom). At the time they were tested, each subject in the trained group performed $> 80\%$ correct on the $d = 0.4$ set (on a block of 50 match and 50 nonmatch trials, randomly interleaved).

After subjects in the training group reached the performance criterion on the training task, they were tested in a task similar to the training task but in which the viewpoint of match and non-match choice stimuli differed by 45°, corresponding to a rotation of the car towards the viewer (as shown in Fig. 5 for a 22.5° rotation, as used in Experiment 2). This direction of rotation was chosen arbitrarily. For each viewpoint and distance combination, subjects were tested on 30 match and 30 nonmatch trials, for a total of 240 trials (with 120 unique match/nonmatch pairs), which were presented in random order. The high number of trials was chosen to mitigate possible effects of morph space anisotropy with respect to subjects' perceptual similarity judgments. Subjects received no feedback on their performance in the testing task.

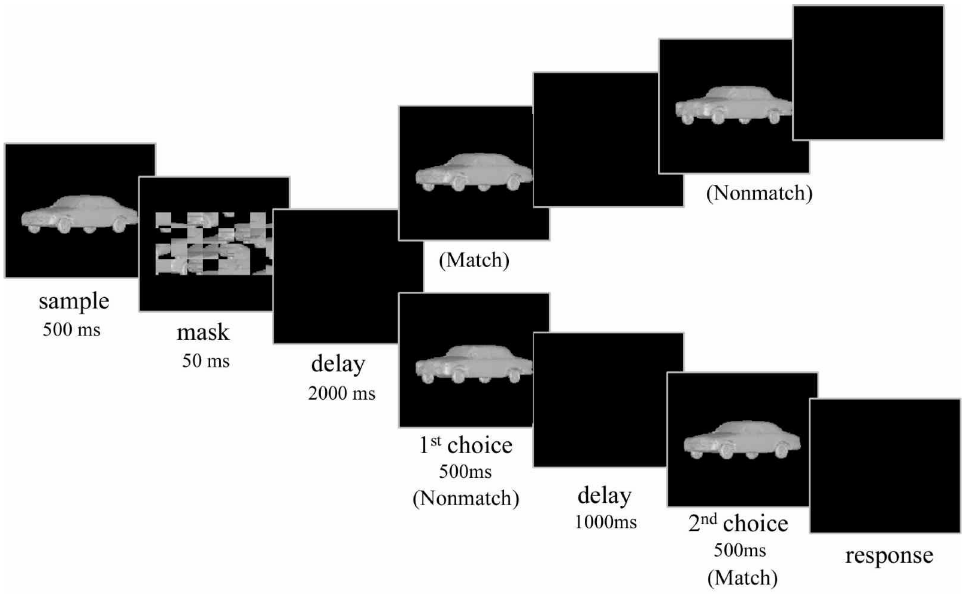

Figure 3: Training task for Experiments 1 and 2. Shown are an example each of the two different kinds of trials: match and nonmatch trials. In both, a sample car is followed by a mask and a delay. Then, in a match trial (upper branch), the match car appears as the first choice car, and a distractor car as the second choice car. For a nonmatch trial (lower branch), the order is reversed. The example objects shown here are from the 15 car system used in Experiment 2 (see section 4). Subjects had to make a response after the offset of the second choice car and received auditory feedback on the correctness of their response.

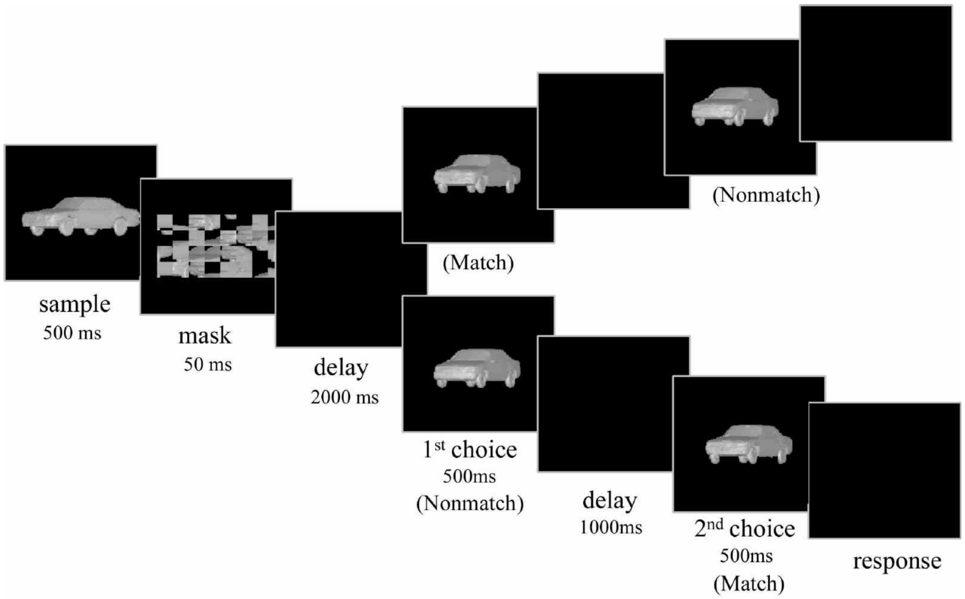

Figure 5: Testing task for Experiments 1 and 2. The task is identical to the training task from Fig. 3 except for the absence of feedback and the fact that the viewpoint choice cars were presented at could vary between trials (the example shows a viewpoint difference of $\Delta\varphi = 22.5°$, as used in Experiment 2).

2.2 Results

Subjects were 14 members of the MIT community that were paid for participating in the experiment plus the first author. Seven subjects and the first author were assigned to a "trained" group that received training sessions until the performance criterion as described above was reached, upon which they performed the testing task as described above. One subject, whose initial performance on the easiest ($d = 0.6$) training set was at chance, was excluded from the training group. For the remaining subjects, criterion was reached after one or two training sessions of one hour each (average 1.75 sessions). Another seven subjects, the "untrained" group, did not receive any training on the stimuli but were run only on the testing task.

As mentioned above, this comparison to an untrained group is essential: The visual system is very well able to perceive novel objects even without training, *i.e.*, there is a baseline discrimination performance for *any* novel object class. This is also expected for the car objects used in our study, as their shapes make them similar to real cars subjects have some visual experience with (in agreement with our objective to use a natural object class to investigate the learning of natural object classes). However, as the degree of similarity between match and nonmatch objects is continuously increased during training, subjects have to learn to perceive fine shape differences among the morphed cars used in the experiment. It is this learning component we are interested in, allowing us to investigate how class-specific training on one viewpoint transfers to other viewpoints.

Figure 6 shows the averaged performance of the subjects in the trained group on the test task. A repeated measures ANOVA (using SPSS 8.0 for Windows) with the factors of viewpoint and distance in morph space between match and nonmatch objects revealed highly significant main effects of both viewpoint difference and distance in morph space ($F(1,6) = 155.224$ and $F(1,6) = 21.305$, resp., $p < 0.005$) on recognition rate, with a non-significant interaction ($F(1,6) = .572$, $p > .4$) between the two factors. Interestingly, performance even for the 45° viewpoint difference is significantly above chance ($p < 0.001$ for both distances, t-test).

The performance of the untrained subjects is shown in Fig. 7. The ANOVA here again revealed significant main effects of viewpoint and distance (for main effect of distance, $F(1,6) = 7.814$, $p < 0.05$, for viewpoint $F(1,6) = 14.994$, $p < 0.01$, no significant interaction $p > .4$). Comparing the average recognition rates for the trained (Fig. 6) and untrained (Fig. 7) groups, it is apparent that recognition rates for the trained view are higher in the trained group than in the untraineed group whereas performance of the two groups seems to be equal for the $\Delta\varphi = 45°$ view. Examining the different performances in the two groups in more detail, t-tests (one-tailed, assuming that training improves performance) on the two different populations for the different conditions revealed that the performance of the trained group was significantly better than the untrained group for 0° viewpoint difference for both $d = 0.6$ and $d = 0.4$ ($p < 0.05$), while the difference was not significant for the 45° viewpoint difference ($p \geq 0.3$). Note that the observed performance differences are unlikely to be due to the untrained subjects' lower familiarity with the 2AFC paradigm, as performance differed only on a subset of conditions, namely those where sample and test objects were presented at the same viewpoint. As for the trained group, recognition in the untrained group for a viewpoint difference of 45° was significantly above chance ($p < 0.002$ for both distance levels).

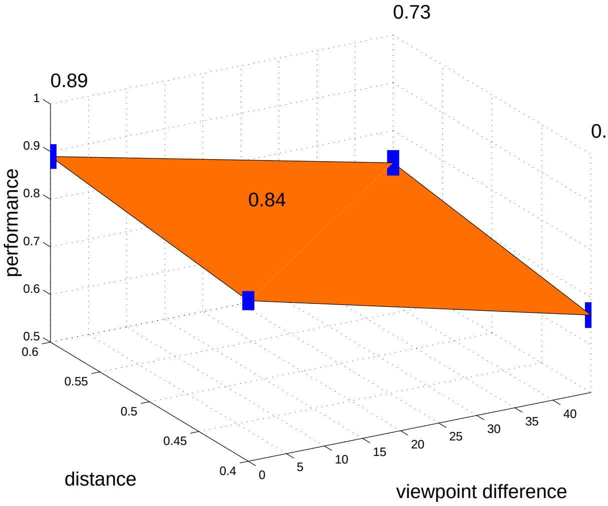

Figure 6: Average performance of the trained subjects ($N = 7$) on the test task of Experiment 1. The z-axis shows performance, x-axis viewpoint difference $\Delta\varphi$ between sample and test objects ($\Delta\varphi \in \{0°, 45°\}$), and y-axis morph space distance d between match and nonmatch objects ($d \in \{0.4, 0.6\}$). The height of the bars at the data points shows ± standard error of the mean. The numbers above the data points show the corresponding numerical scores.

Thus, the data show the following

1. For the trained as well as the untrained subject groups, recognition performance decreases with increasing target/distractor similarity and with viewpoint difference.

2. Both trained and untrained subjects perform above chance at the 45° view.

3. Training subjects with randomly chosen cars at the 0° view improves recognition of class members at the 0° view but does not affect recognition performance if the viewpoint difference between sample and test objects is 45°.

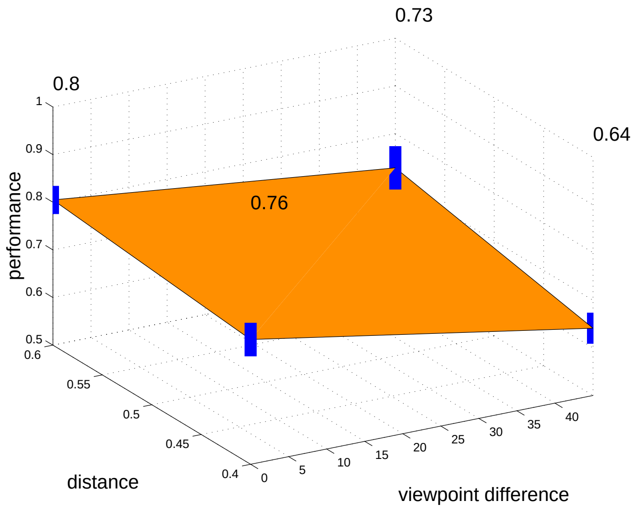

Figure 7: Average performance of untrained subjects ($N = 7$) on the test task of Experiment 1. Axis labeling as in Fig. 6.

2.3 Discussion

The results of Experiment 1 indicate that while there is a recognition benefit from training on the 0° view, it does not transfer to the 45° view. However, even for a viewpoint difference of 45°, recognition is still significantly above chance.

These results are especially interesting with respect to a recent study by Edelman [5] that, for a categorization task involving geon-based stimuli, reported two different performance regimes depending on the degree of viewpoint difference between sample and test object. He surmised that this might be the manifestation of two recognition mechanisms, one at work for small viewpoint differences and another one for larger ones.

The results of Experiment 1 suggest the following interpretation: While training at a single viewpoint does not transfer to a viewpoint difference of 45° between sample and test object, recognition in this case might rely on features that are robust to rotation (like the roof shape of the car) and which do not depend on object-class specific learning. Similar non-specific features can be used in the untrained group to perform recognition also for the *unrotated* viewpoint, but they are not sufficient to perform discrimination in fine detail, as evidenced by the lower recognition performance for the training view. Training lets subjects build a detailed class and viewpoint-specific representation that supplements the existing system: Subtle shape discriminations require sufficiently finely detailed features that are more susceptible to 3D rotation, whereas coarser comparisons can likely be performed also with cruder features or use more view-invariant representations optimized for different objects (see general discussion).

Over which range of viewpoints would we expect training at a single viewpoint to have an effect? To answer this question we performed simulations in HMAX presented in the next section.

3 Modeling: Representing Continuous Object Classes in HMAX

Our investigation of object recognition in continuous object classes is based on our recently presented HMAX model [27] that has been extensively tested on the representation of individual “paperclip” objects. After a brief review of HMAX, we shall demonstrate how the same model can easily be applied to the representation of natural object classes (the use of such a representation to perform object categorization is described in [28]).

3.1 The HMAX Model of Object Recognition in Cortex

Figure 8 shows a sketch of our model of object recognition in cortex [26, 27] that provides a theory of how view-tuned units (VTUs) can arise in a processing hierarchy from simple-cell like inputs. As discussed in [26, 27], the model accounts well for the complex visual task of invariant object recognition in clutter and is consistent with several recent physiological experiments in inferotemporal cortex. In the model, feature specificity and invariance are gradually built up through different mechanisms. Key to achieve invariance and robustness to clutter is a MAX-like response function of some model neurons which selects the maximum activity over all the afferents, while feature specificity is increased by a template match operation. By virtue of combining these two operations, an image is represented through a set of features which themselves carry no absolute position information but code the object through a combination of local feature arrangements. At the top level, view-tuned units (VTUs) respond to views of complex objects with invariance to scale and position changes.[¶] In all the simulations presented in this paper we used the “many feature” version of the model as described in [26, 27].

3.2 View-Dependent Object Recognition in Continuous Object Classes

As mentioned in the introduction, various studies have provided support that “natural” object classes, in particular faces [37, 38], are represented by a population of units broadly tuned to representatives of this object class. Other physiological studies have provided evidence that neuronal tuning in IT can be changed as a result of training [2, 13, 16, 19].

A population-based representational scheme is easily implemented in HMAX through a group of VTUs (the *stimulus space-coding units, SSCU*, which can also provide a basis for object categorization [28]) tuned to

[¶] To perform view-invariant recognition, VTUs tuned to different views of the same object can be combined, as demonstrated, *e.g.*, in [23].

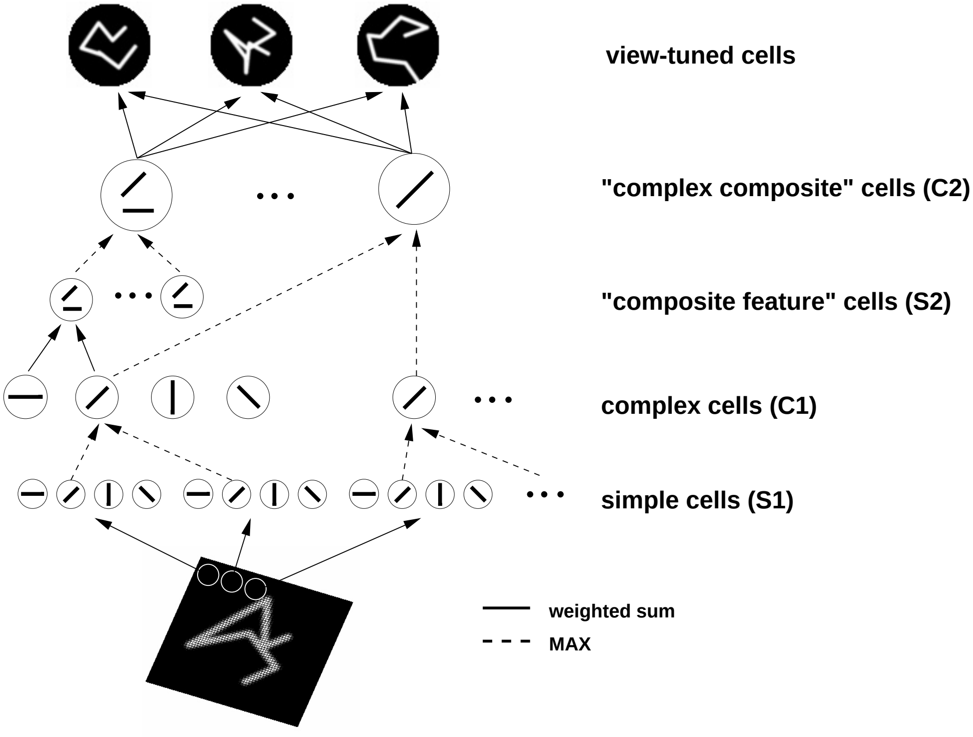

Figure 8: Our model of object recognition in cortex (from [26]). The model is an hierarchical extension of the classical paradigm [10] of building complex cells from simple cells. It consists of a hierarchy of layers with linear ("S" units in the notation of Fukushima [8], performing template matching, solid lines) and non-linear operations ("C" pooling units [8], performing a "MAX" operation, dashed lines). The non-linear MAX operation — which selects the maximum of the cell's inputs and uses it to drive the cell — is key to the model's properties and is quite different from the basically linear summation of inputs usually assumed for complex cells. These two types of operations respectively provide pattern specificity and invariance (to translation, by pooling over afferents tuned to different positions, and scale (not shown), by pooling over afferents tuned to different scales).

representatives of the object class.[‖] Discrimination between different objects proceeds by comparing the corresponding activation patterns over these units.

To investigate the properties of such a representation, we created a set of car objects using the eight car system. In particular, we created lines in morph space connecting each of the eight prototypes to all the other prototypes for a total of 28 lines through morph space, with each line divided into 10 intervals. This created a set of 260 unique cars. Each car was rendered from 13 viewpoints around the 225° training view (TV), spanning the range from 180° to 270° in steps of 7.5°, which yielded a total of 3380 images.

We then defined a set of SSCUs tuned to representatives of the car class. The representatives were chosen by performing k-means clustering on the set of 260 cars shown at the training view (results shown are for individual k-means runs — repeated runs tended to produce quantitatively similar results).

To examine the viewpoint-dependence of object recognition in the car class we then performed trials in which each of the TV cars was presented to the model (the "sample car"), causing an activation pattern $\mathbf{a}_{sample}$ over the SSCUs. Then a "match" and a "nonmatch" (distractor) object were chosen. The former was the same car shown from a different viewpoint $\varphi = 225°{+}\Delta\varphi,\ -45° \leq \Delta\varphi \leq 45°$ as described above, while the latter was a different car a distance d away from the sample car along the same morph line, shown at the same viewpoint φ as the match car. The two choice cars caused activation patterns $\mathbf{a}_{match}$ and $\mathbf{a}_{nonmatch}$, resp. Recognition of the rotated sample car was said to be successful if

$$||\mathbf{a}_{sample} - \mathbf{a}_{match}|| < ||\mathbf{a}_{sample} - \mathbf{a}_{nonmatch}|| \qquad (1)$$

using a (unweighted) Euclidean metric, *i.e.*, if the SSCU activation pattern caused by the sample object was more similar to the match object's activation pattern than to the activation pattern caused by the nonmatch object. This paradigm is equivalent to a two alternative-forced choice task and has the advantage that modeling of the decision process is straightforward. Recognition performance for each $(\Delta\varphi, d)$ combination was tested for all possible sample / distractor car pairs.

Figure 9 shows recognition performance as a function of d and $\Delta\varphi$ for a representation based on $n_{SSCU} = 16$ SSCUs (selected by k-means as described above), each with a tuning width of $\sigma = 0.2$ and $c = 256$ connections to the preceding C2 layer (*i.e.*, fully connected to all 256 C2 units [26]).

We see that the general trend observed in the experiment also holds in the simulations: Discrimination performance drops with increasing target / distractor similarity and increasing viewpoint difference between sample and choice objects. In particular, for the positive rotation direction investigated in Experiment 1 (and also Experiment 2, see below), performance reaches chance for rotations of 30°,[**] while it is still robustly above chance for viewpoint differences of 22.5° .

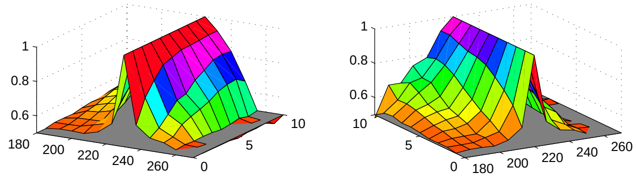

Figure 9: Recognition performance of the model on the eight car morph space. x-axis shows viewpoint φ of nonmatch object, y-axis match / nonmatch distance d (in steps along the morph line the sample object lies on) in morph space, and z-axis model discrimination performance for all (φ, d) stimulus pairs in the sample set. Model parameters were $n_{SSCU} = 16$, $\sigma = 0.2$, $c = 256$. The two subplots show the same graph from two different viewpoints to show positive rotations (*i.e.*, toward the front, so that the front of the car is turning towards the viewer, as used in the psychophysical experiments), left plot, and negative rotations (*i.e.*, towards the back, so that the side of the car faces the viewer), right plot.

In order to investigate how discrimination performance varies with the number of SSCUs in the representation, the tuning width of the individual SSCU and the number of afferents to each SSCU, we shall in the following plot the *average (one-sided) invariance range* as a function of these parameters, limiting ourselves to the positive rotations also used in the experiment. The average one-sided invariance range, $\bar{r}$, for a given set of model parameters and a given match / nonmatch distance d in morph (in steps along the morph line the sample object lies on) space is calculated by summing up the above-chance performance values, p'_i, for viewpoint difference $\Delta\varphi_i, p'_i = 2{*}(p_i{-}0.5)$ obtained from the raw performance scores p_i shown, *e.g.*, in Fig. 9. Then,

$$\bar{r} = \sum_{i=1}^{n-1}(p'_i - p'_{i+1})\Delta\varphi_i + p'_n\Delta\varphi_n\ , \qquad (2)$$

with $n = 5, \Delta\varphi_i = \{0°, 7.5°, 15°, 22.5°, 30°\}$. This definition assumes a monotonic drop in performance with increasing $\Delta\varphi$, *i.e.*, that if an object can be discriminated for a certain $\Delta\varphi$ it can also be discriminated for all $\Delta\varphi' < \Delta\varphi$. This condition was met in the great majority of cases.

Dependence of rotation invariance on tuning width of SSCUs. The dependence of the average rotation invariance on the tuning width of the SSCUs, *i.e.*, the σ of the Gaussian SSCUs, is shown in Fig. 10 (all other parameters as before). The invariance range seems to

[‖] Note that the receptive fields of the SSCUs do not have to respect class boundaries, as long as they adequately cover the input space [28].

[**] For a few parameter sets, performance at $\Delta\varphi = 30°$ was still slightly above chance.

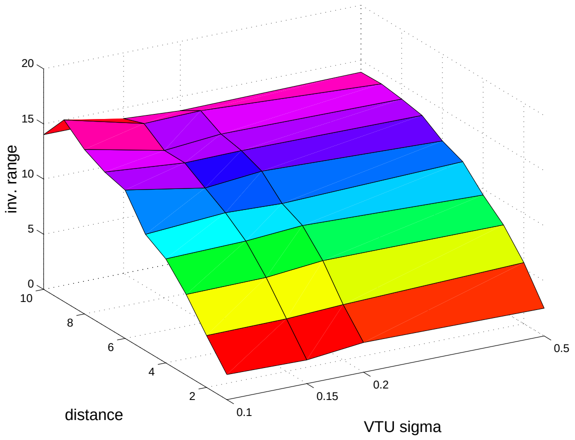

Figure 10: Dependence of average (one-sided) rotation invariance, $\bar{r}$, (z-axis) as a function of the tuning width, σ, of SSCUs (x-axis). y-axis in this and Figs. 11–13 shows distance in steps between match and nonmatch objects along the morph line the sample (match) object lies on. Other parameters as in Fig. 9.

be rather independent of the tuning width of the SSCUs. This is due to the high precision and dynamic range of the model neurons whose response function is Gaussian: Even if the stimulus rides on the tail of a sharply tuned Gaussian unit, the unit's response still depends monotonically on the match of the afferent activity to the unit's preferred stimulus. For a more realistic response function with a narrower dynamic range we would expect a stronger effect of tuning width, in particular a drop of performance as tuning becomes too narrow. Note that the average one-sided invariance range for cars is very comparable to that obtained for paperclip objects, which was on the order of $30°/2 = 15°$ [27].

Dependence on number of afferents to each SSCU. In the studies of recognition in clutter using HMAX [26] it was found that robustness to clutter can be increased in the model by having view-tuned units receive input only from a subset of units in the C2 layer, namely the n most strongly activated ones. The invariance range, on the other hand, was found to increase with the number of afferents. Figure 11 shows the dependence of the average invariance range on the number of strongest afferents to each SSCU (left plot) for a representation based on $n_{SSCU} = 16$, compared to a "grandmother" representation (right plot) where a dedicated "grandmother" unit was allocated for each sample stimulus and match and nonmatch objects were discriminated based on which of the two stimuli caused a greater excitation of the "grandmother" unit. This is identical to the representation used in the recognition experiments with paperclip objects [27]. Interestingly, while the invariance range shows a strong dependence on the number of afferents in the "grandmother" case, with invariance range asymptoting at about $c = 100$ (similar to what had been found for paperclips [27]), there seemed to be a much weaker dependence on the number of afferents in the population-based representation.

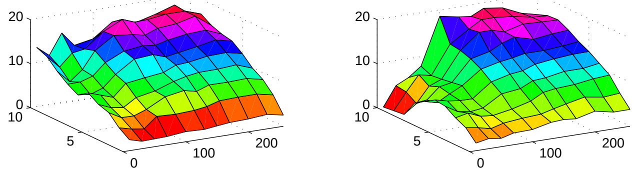

Figure 11: Dependence of invariance range on the number of afferents to each SSCU (x-axis), left plot, which were varied from having only the 10 most strongly activated C2 units for each SSCU feed into the respective SSCU to a fully connected network with 256 afferents. Other parameters as in Fig. 9. The right plot shows the average rotation invariance range for a "grandmother"-like representation where an individual neuron is allocated for each sample stimulus, and recognition performance is based just on this unit.

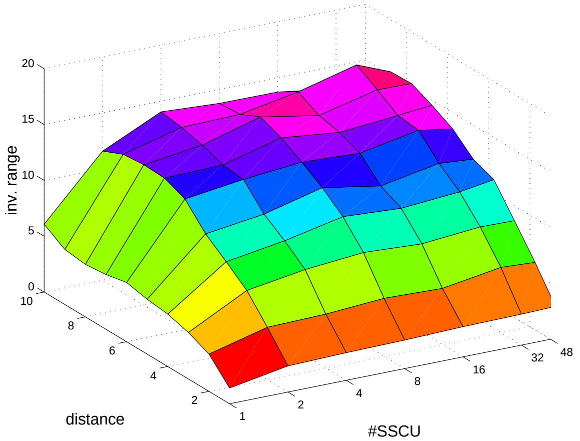

Figure 12: Dependence of invariance range on the number of SSCUs (x-axis). Other parameters as in Fig. 9.

Dependence on number of SSCUs. Figure 12 shows the average rotation invariance as a function of the number of SSCUs. While rotation invariance for a representation consisting of just one SSCU (the average) shows expectedly poor rotation invariance, the invariance is already sizeable for $n_{SSCU} = 2$ and grows only weakly with the number of SSCU for $n_{SSCU} > 2$. Thus, it may seem that a representation based on more than two units offers only marginal benefits. However, this picture changes dramatically if noise is added to the representation, which was studied in the next section.

Robustness to Noise. The above simulations all assumed completely deterministic model neurons where firing rates are quasi-continuous variables of very high precision. Real neurons, however, are likely to show a more stochastic response of limited precision to re-

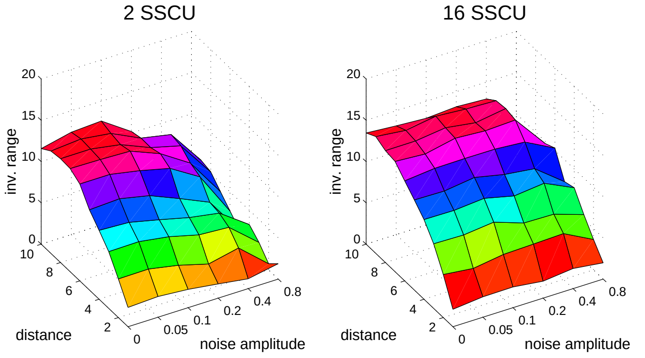

Figure 13: Effect of addition of noise to the SSCU representation for different numbers of SSCU in the representation. The x-axis shows the amplitude of the Gaussian noise that was added to each SSCU in the representation. The plot on the left shows the performance of a representation based on $n_{SSCU} = 2$, the right one for $n_{SSCU} = 16$.

peated presentation of the same stimulus. We can qualitatively simulate the implications of such a behavior in the model by adding noise to the model neurons and examining how much noise of a certain amplitude affects performance for the different model configurations.

Figure 13 shows how additive Gaussian noise of varying amplitude (which was just added to the activity pattern caused by the sample stimulus) affects invariance ranges for representations based on varying numbers of SSCUs. We see that while the performance for zero noise is similar, the representation based on $n_{SSCU} = 2$ is much less robust to noise than a representation based on $n_{SSCU} = 16$. Thus, increasing the number of units in a representation increases its robustness to noise (at least for the case of independent additive noise, as used here).

The “Other Class” Effect. An analog of the “Other Race” effect [14, 18] mentioned in the introduction can be modeled in quite a straightforward fashion, if we replace the SSCU representation tuned to cars with one tuned to a different object class. Here we chose six units tuned to prototypical cats and dogs (as used in separate physiological and modeling studies of object categorization [7, 28]), rendered as clay models and size-normalized, shown in Fig. 14. Figure 15 shows the “Other Class” effect obtained when using these six cat / dog SSCU to perform the car discrimination task from the previous sections: While performance in the no noise condition is only somewhat worse than with the specialized representation (left plot), even noise of very low amplitude reduces the performance to chance, as the cat / dog-specific SSCU respond only little to the car objects, thus making the activation pattern highly sensitive to noise. This underlines the influence of a specialized class representation on discrimination / recognition performance.

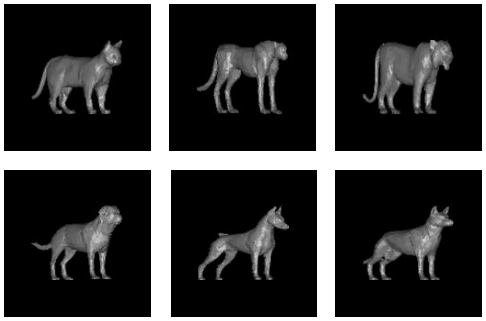

Figure 14: Cat / dog prototypes.

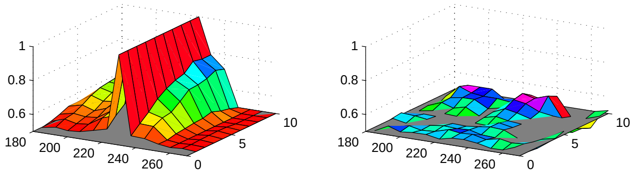

Figure 15: The “Other Class” effect with six SSCUs tuned to the cat / dog prototypes ($\sigma = 0.5$; for $\sigma = 0.2$, performance was even lower and numerical underflows occured; other parameters as in Fig. 9). Left plot shows no noise, right with noise amplitude of 0.05. Compare to Fig. 13.

Feature learning. The C2 features used in HMAX are based on combinations of difference-of-Gaussians (DOG) filters of different orientations that might not be optimal to perform object discrimination for the car class used in the experiments. Can performance be improved with a more specific feature set?

No learning algorithm for feature learning in the HMAX hierarchy has been presented so far. However, we can investigate the effect of a class-specific feature set in a two-layer version of HMAX [25] where S1 filters are not limited to DOGs but can take on arbitrary shapes, and C1 units pool (using the MAX function) over all S1 cells at different positions and scales tuned to the same feature, with C1 units feeding directly into VTUs, without S2 or C2 layers. Invariance properties of the two-layer version of HMAX using a set of 10 features consisting of bars and corners are comparable to the full model [25, 27].

We can obtain a feature set specific to the car object class by performing clustering on the set of image patches created by dividing each sample car into small 12×12 pixel patches. Figure 16 shows the features obtained by clustering the sample car image patch space (containing only those patches that had at least 10% of their pixels set) into 200 clusters using k-means.

Figure 17 shows the performance of the two-layer model using these features. In comparison to standard HMAX (Fig. 9), performance of the two-layer model is somewhat worse for positive rotation, with performance dropping to chance already for rotations of 22.5°. On the other hand, performance for negative rotations

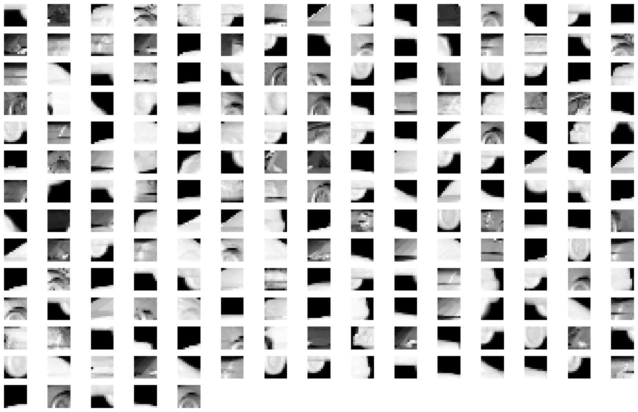

Figure 16: Car object class-specific features obtained by clustering the image patches of sample cars.

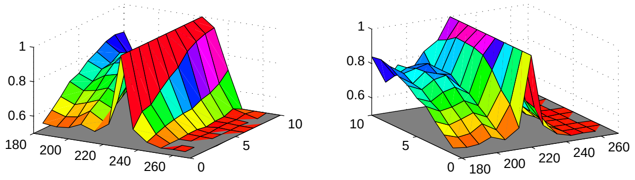

Figure 17: Performance of the two layer-model using the features shown in Fig. 16. Parameters were $n_{SSCU} = 16, \sigma = 0.2$, 200 afferents to each SSCU. Axes as in Fig. 9.

(*i.e.*, those that turn the side of the car towards the viewer) is somewhat better. Both effects could be due to the fact that many of the patches shown in Fig. 16 contain car parts likely to change under positive rotation, like the wheels, but which are more stable under negative rotation.

3.3 Discussion

The simulations point to several interesting properties of a population-based representation,

1. invariance ranges for a population-based representation where object identity is encoded by the distributed activity over several units (SSCUs) tuned to representatives of that class are comparable to a representation based on "grandmother" cells where recognition is based on a dedicated unit for each object to be recognized;

2. invariance ranges are already high for a representation based on a low number of SSCUs, but robustness to noise grows with the number of SSCUs in the representation;

3. even if each SSCU is only connected to a low number of afferents (the n most strongly activated) from the C2 layer, invariance based on the population representation is high.

The last point is especially intriguing, as it might point to a way to obtain robustness to clutter together with high invariance to rotation in depth, avoiding the tradeoff found for a "grandmother" representation [26]. Further, the simulations suggest an additional advantage of a population-based representation over a representation based on the C2 features directly: Suppose a certain object (*e.g.*, "my car") is to be remembered. If a car-specific representation has been learned it suffices to store the low-dimensional activation pattern over the SSCUs whereas in the absence of a specialized representation it will be necessary to store the activity pattern over a much higher number of C2 units to achieve comparable specificity.[††]

In the context of Experiment 1, the simulations suggest that the advantage of viewpoint- and class-specific training should only extend to roughly between 22.5° and 30° of viewpoint difference between training and testing viewpoint. It thus confirms our theory put forward in the discussion of Experiment 1 that performance there was due to the combination of a class- and viewpoint-specific representation and a pre-existing, less specific but more view-invariant representation. The class-specific representation is capable of fine shape discrimination but only over limited a range of viewpoints, while the more general one uses features that are less optimized for the novel object class but show greater tolerance to rotation. For small $\Delta\varphi$, the two representations can complement each other, while for larger viewpoint differences the unspecific features still allow recognition in some cases.

4 Experiment 2

Experiment 1 suggested that the view-invariance range derived from one object view extends less than 45° from the training view. The modeling work presented in the previous section predicted that an effect of training should only extend to between 22.5° and 30° of viewpoint difference. The purpose of Experiment 2 was to more finely investigate the degree of view-invariance and at the same time examine how the training effect observed in Experiment 1 carried over to a broader object class. The latter modification was chosen as the small size of the object class in Experiment 1 implied that discrimination hinged on a very limited number of features. In Experiment 2 we therefore increased the size of the class significantly (to 15 prototypes instead of 3) to make the discrimination task harder in the hope of increasing the learning effect. Further, we added an intermediate viewpoint difference, 22.5°, in the test task, as the simulations presented in the previous section suggested that the effect of training should start to drop off beyond this viewpoint difference.

[††] If there are also SSCUs tuned to objects from other classes, it would suffice to store the activation pattern over the most strongly activated SSCUs to achieve sufficient specificity, as simulations have indicated (not shown). Thus, it is not necessary for the SSCUs to carry class labels (cf. [28]).

4.1 Methods

4.1.1 Stimulus Generation

Stimuli in Experiment 2 were drawn from a morph space spanned by the 15 car prototypes shown in Fig. 18. Objects were rendered in the same way as for Experiment 1. Coefficient vectors in morph space were now generated by first randomly picking which coefficients should be different from zero with a probability of $p = 0.25$. Those coefficients were then set to random values between 0 and 1 picked from a uniform distribution and the whole coefficient vector subsequently sum-normalized to one. This algorithm was introduced to increase the diversity of the training stimuli as randomly setting all coefficients with subsequent normalization tended to produce a rather homogeneous set of objects visually similar to the average.

In test trials, distractors D_6 for the $d = 0.6$ trials were selected by picking a coefficient vector of the appropriate Euclidean distance from the sample stimulus. The vector was chosen by appropriately modifying an equal number of coefficients as were different from zero in the sample vector, observing the constraint that the (linear) sum over all coefficients had to stay constant. The distractor for the $d = 0.4$ trials was selected to lie on the line connecting the sample and D_6. Moreover, the same objects were chosen for the different $\Delta\varphi$ trials. This made performance comparisons in the different conditions easier as it decreased the effects of morph space anisotropy (that was due to the different visual similarities of the prototypes). For each $d \in \{0.4, 0.6\}, \Delta\varphi \in \{0°, 22.5°, 45°\}$ combination, subjects were tested on 30 match and 30 nonmatch trials for a total of 360 trials. All conditions were randomized.

4.1.2 Subject Training

Training for subjects in the trained group in Experiment 2 proceeded in a similar fashion as in Experiment 1. Subjects started out with match / nonmatch pairs a distance $d = 0.6$ apart in morph space and then performed blocks of 100 trials (50 match, 50 nonmatch), until performance on the block exceeded 80%. When their performance reached criterion, d was decreased by 0.1, down to a final $d_{final} = 0.4$. To make sure that subjects' performance was not specific to the training set, performance was tested on another set of $d = 0.4$ stimuli after they reached criterion. In all cases subjects' performance on the second $d = 0.4$ set was comparable to their performance on the first set.

4.2 Results

Subjects were 24 members of the MIT community that were paid for participating in the experiment, all of which were naive to the purpose of the experiment and had not participated in experiment 1. Seven subjects were randomly assigned to a "trained" group that received training sessions until the performance criterion as described above was reached. Six further subjects started to receive training, but did not complete it due to initial chance performance on the easiest ($d = 0.6$) training stimuli ($N = 2$), failure to reach criterion after five training sessions ($N = 1$, at which point the subject's performance had asymptoted at 75% on the $d = 0.4$ training set for the previous three sessions), or failure to come to training sessions ($N = 3$).

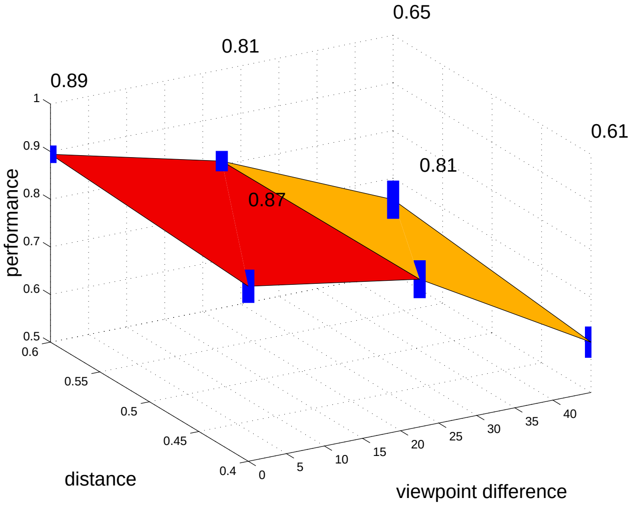

Figure 19: Average performance of trained subjects ($N = 5$) in Experiment 2. Axes labeling as in Fig. 6.

For one of the subjects that completed training, the data collection program malfunctioned during testing and the subject was excluded from further analysis. Another trained subject was not available for testing. 11 subjects did not receive any training on the stimuli but were only run on the testing task. One subject of that group whose performance was highly anomalous (at chance for 0° viewpoint difference, probably due to subject sleepiness) was excluded from further analysis. Another subject was excluded due to program malfunction during the experiment.

Training sessions already revealed that discrimination in the larger morph space was harder than in the three prototype space from Experiment 1: Subjects on average required four hours of training (range: 3–5 hours), more than twice as much as in Experiment 1.

Figure 19 shows the averaged performance of the subjects in the trained group on the test task. A repeated measures ANOVA with the factors of viewpoint and distance in morph space between match and nonmatch objects revealed a highly significant main effect of viewpoint difference ($F(2, 3) = 33.151$, $p \leq 0.005$) on recognition rate, but no significant effect of distance ($F(1, 4) = 3.259$, $p = .145$) and a non-significant interaction ($F(1, 6) = .492, p > .5$) between the two factors.

The averaged performance over the 9 untrained subjects is shown in Fig. 20. There are significant effects of both viewpoint difference ($F(2, 7) = 31.169$, $p \leq 0.001$)

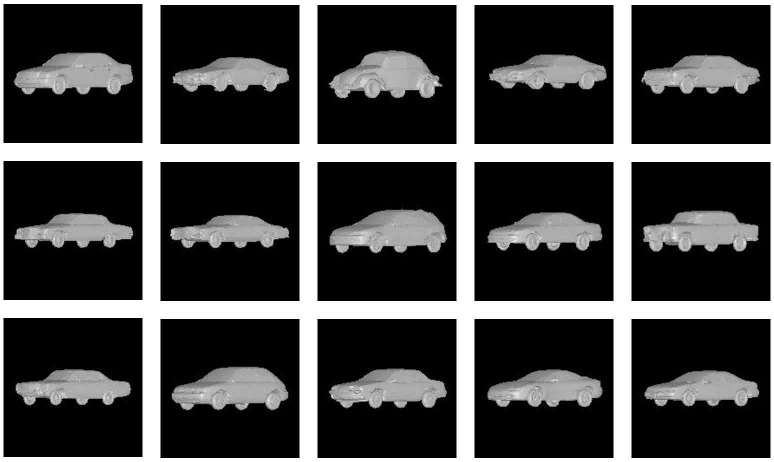

Figure 18: The 15 prototypes used in the 15 car system.

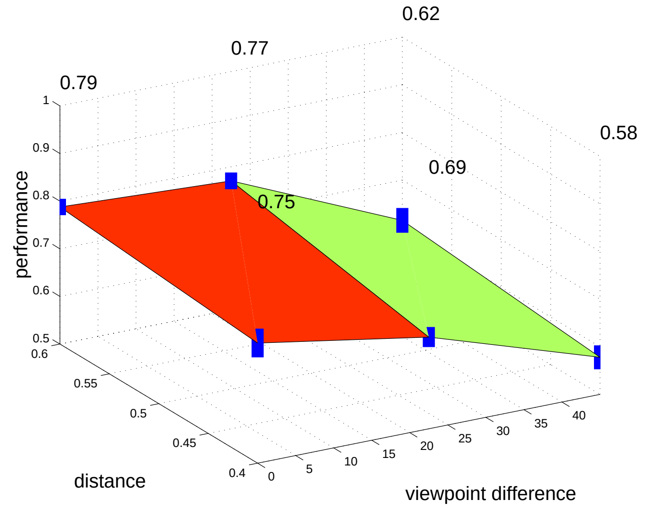

Figure 20: Average performance of untrained subjects ($N = 9$) in Experiment 2. Axes labeling as in Fig. 6.

and distance ($F(1, 8) = 22.478$, $p \leq 0.001$) with no significant interaction between the two ($F(2, 7) = .715$, $p > .5$).

Comparing the trained and untrained groups, t-tests show that recognition performance in the trained group is significantly ($p < 0.02$) higher at the training view for $d = 0.6$ and $d = 0.4$, and also for the $\Delta\varphi = 22.5°$ view and $d = 0.4$ ($p < 0.01$), but does not reach significance for $\Delta\varphi = 22.5°$ and $d = 0.6$ ($p = 0.11$). Recognition performance is not significantly different for the two groups at $\Delta\varphi = 45°$ for both distances ($p \geq 0.2$). For both groups and distances, performance at $\Delta\varphi = 45°$ is significantly above chance ($p < 0.02$, one-tailed t-tests).

4.3 The Model Revisited

What recognition performance would we expect from the model for the stimuli used in Experiment 2? To investigate this question, we used the training stimuli (400 cars from the $d = 0.6, 0.5$ and the two $d = 0.4$ training files) and performed k-means clustering on them to obtain a class-specific representation as subjects might have learned it as a result of training. To investigate the discrimination performance of this representation, we used a SSCU representation with the exact same parameters as in Fig. 9, *i.e.*, $n_{SSCU} = 16, \sigma = 0.2, c = 256$. We then evaluated the performance of this representation for the sample, match, nonmatch triples from the testing task as described in section 4.1.1. Performance at $\Delta\varphi = 45°$ was at chance as expected, but for $\Delta\varphi = 22.5°$, performance was 65% correct for $d = 0.6$ and 63% correct for $d = 0.4$, compatible with the results obtained for the eight car class (Fig. 9).

4.4 Discussion

Increasing the size of the object class to be learned increased task difficulty considerably as evidenced by the longer training times as compared to Experiment 1. As expected, this correlated with a more significant effect of training on recognition performance (even with a smaller group of trained subjects than in Experiment 1).

While the effect of training was highly significant for the training view, we only observed a significant training effect for $\Delta\varphi = 22.5°$ for $d = 0.4$, with the difference at $d = 0.6$ just failing to reach significance ($p = 0.11$). This effect could be interpreted in the "fine/specific" and "coarse/unspecific" dual framework of object recognition proposed in the context of Experiment 1 as indication that the benefits of the class- and viewpoint-specific representation learned in the training phase do not extend much farther than $\Delta\varphi = 22.5°$, as suggested by the simulations presented in section 3. The performance of the untrained subjects can be interpreted as indicating that the coarse class-unspecific representation performs roughly as well as the viewpoint-specific representation at $\Delta\varphi = 22.5°, d = 0.6$, *i.e.*, there is a balance between the class-

and viewpoint-specific and the coarse but more view-invariant representation, whereas for the finer shape discrimination required at $d = 0.4$ the specific representation still provides a significant advantage. Interestingly, the small effect of match / nonmatch distance in morph space for the trained group at $\Delta\varphi = 22.5°$ is paralleled in the model, where there is only a 2% performance difference for this viewpoint difference and the $d = 0.6$ (65%) and $d = 0.4$ (63%) conditions.

5 General Discussion

The psychophysical and modeling results presented in this paper point to an interesting answer to the initial question of object recognition in continuous object classes: Viewpoint-specific training builds a viewpoint-specific representation for that object class. While this representation supports fine shape discriminations for viewpoints close to the training view, its invariance range is rather limited. However, there is a less-specific, pre-existing object representation that cannot discriminate shapes as finely as the trained class-specific representation but shows greater tolerance to viewpoint changes. It is instructive to compare these observations to an earlier paper by Moses *et al.* [20] where it was found that generalization ability for changes in viewpoint and illumination was much greater for upright than for inverted faces, suggesting that prior experience with upright faces extended to the novel faces even though the novel faces had been encountered at one view only.

Based on our simulation results, we would expect a similar behavior, *i.e.,* limited invariance around the training view with high sensitivity to shape changes in addition to a coarser but more invariant system, also for other transformations, such as, for instance, varying illumination. It will be very interesting to test this hypothesis, by training a subject as presented in this paper but then varying, for instance, illumination angle, and to then compare trained and untrained groups and model performance. This would also allow us to make inferences about the invariance properties of the feature channels feeding into the SSCU (which determine the invariance range of the learned class- and viewpoint-specific representation).

Another interesting, more theoretical, question concerns the properties of the “pre-existing” representation: Can experience with rotated objects of a certain class provide greater viewpoint invariance albeit with coarse shape resolution also for novel objects belonging to a different class? Poggio and Vetter [24] (see also [35]) proposed the idea that class-specific view-invariant features could be learned from examples and then used to perform recognition of novel objects of the *same* class given just one view. Jones *et al.* [12] (see also [1]) presented a computational implementation of this proposal showing how class-specific learning could facilitate perceptual tasks. If such a mechanism transfers also to sufficiently similar members of a novel object class (for instance from real cars, which have been seen from many viewpoints, to the morphed cars), then it would provide a suitable candidate for the less-specific but more view-invariant representation found in this experiment. Some simulations along these lines have been performed in ([6], pp. 131), but the performance of such a scheme, for instance with respect to view-invariant recognition, was never tested. It will be very exciting to explore this issue in future work.

Acknowledgements

Thanks to Valerie Pires for running the subjects and for expert technical assistance, to Gadi Geiger and Pawan Sinha for stimulating discussions, advice on the experimental paradigm and data analysis, and for comments on an earlier version of this manuscript, and to Christian Shelton for the morphing system used to generate the stimuli.

References

[1] Beymer, D. and Poggio, T. (1996). Image representations for visual learning. *Science* **272**, 1905–1909.

[2] Booth, M. and Rolls, E. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex* **8**, 510–523.

[3] Bülthoff, H. and Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Nat. Acad. Sci. USA* **89**, 60–64.

[4] Bülthoff, H., Edelman, S., and Tarr, M. (1995). How are three-dimensional objects represented in the brain? *Cereb. Cortex* **3**, 247–260.

[5] Edelman, S. (1995). Class similarity and viewpoint invariance in the recognition of 3D objects. *Biol. Cyb.* **72**, 207–220.

[6] Edelman, S. (1999). *Representation and Recognition in Vision*. MIT Press, Cambridge, MA.

[7] Freedman, D., Riesenhuber, M., Shelton, C., Poggio, T., and Miller, E. (1999). Categorical representation of visual stimuli in the monkey prefrontal (PF) cortex. In *Soc. Neurosci. Abs.*, volume 29, 884.

[8] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cyb.* **36**, 193–202.

[9] Gauthier, I. and Tarr, M. (1997). Becoming a “Greeble” expert: exploring mechanisms for face recognition. *Vis. Res.* **37**, 1673–1682.

[10] Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Phys.* **160**, 106–154.

[11] Jones, M. and Poggio, T. (1996). Model-based matching by linear combinations of prototypes. AI Memo 1583, CBCL Paper 139, MIT AI Lab and CBCL, Cambridge, MA.

[12] Jones, M., Sinha, P., Vetter, T., and Poggio, T. (1997). Top-down learning of low-level visual tasks. *Curr. Biol.* **7**, 991–994.

[13] Kobatake, E., Wang, G., and Tanaka, K. (1998). Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J. Neurophys.* **80**, 324–330.

[14] Lindsay, D., Jack Jr., P., and Christian, M. (1991). Other-race face perception. *J. App. Psychol.* **76**, 587–589.

[15] Logothetis, N., Pauls, J., Bülthoff, H., and Poggio, T. (1994). View-dependent object recognition by monkeys. *Curr. Biol.* **4**, 401–414.

[16] Logothetis, N., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**, 552–563.

[17] Logothetis, N. and Sheinberg, D. (1996). Visual object recognition. *Ann. Rev. Neurosci.* **19**, 577–621.

[18] Malpass, R. and Kravitz, J. (1969). Recognition for faces of own and other race. *J. Pers. Soc. Psychol.* **13**, 330–334.

[19] Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* **335**, 817–820.

[20] Moses, Y., Edelman, S., and Ullman, S. (1993). Generalization to novel images in upright and inverted faces. Technical Report CS93-14, Weizmann Institute of Science, Israel.

[21] Newell, F. (1998). Stimulus context and view dependence in object recognition. *Perception* **27**, 47–68.

[22] Olshausen, B., Anderson, C., and van Essen, D. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* **13**, 4700–4719.

[23] Poggio, T. and Edelman, S. (1990). A network that learns to recognize 3D objects. *Nature* **343**, 263–266.

[24] Poggio, T. and Vetter, T. (1992). Recognition and structure from one 2D model view: Observations on prototypes, object classes and symmetries. AI Memo 1347, CBIP Paper 69, MIT AI Lab and CBIP, Cambridge, MA.

[25] Riesenhuber, M. and Poggio, T. (1998). Just one view: Invariances in inferotemporal cell tuning. In *Advances in Neural Information Processings Systems*, Jordan, M., Kearns, M., and Solla, S., editors, volume 10, 167–194 (MIT Press, Cambridge, MA).

[26] Riesenhuber, M. and Poggio, T. (1999). Are cortical models really bound by the "Binding Problem"? *Neuron* **24**, 87–93.

[27] Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025.

[28] Riesenhuber, M. and Poggio, T. (1999). A note on object class representation and categorical perception. AI Memo 1679, CBCL Paper 183, MIT AI Lab and CBCL, Cambridge, MA.

[29] Shelton, C. (1996). Three-dimensional correspondence. Master's thesis, MIT, Cambridge, MA.

[30] Tanaka, K. (1996). Inferotemporal cortex and object vision. *Ann. Rev. Neurosci.* **19**, 109–139.

[31] Tarr, M. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonom. Bull. & Rev.* **2**, 55–82.

[32] Tarr, M. (1999). News on views: pandemonium revisited. *Nat. Neurosci.* **2**, 932–935.

[33] Tarr, M. and Gauthier, I. (1998). Do viewpoint-dependent mechanisms generalize across members of a class? *Cognition* **67**, 73–110.

[34] Vetter, T. and Blanz, V. (1998). Estimating coloured 3D face models from single images: An example based approach. In *Proceedings of the European Conference on Computer Vison ECCV'98* (Freiburg, Germany).

[35] Vetter, T., Hurlbert, A., and Poggio, T. (1995). View-based models of 3D object recognition: invariance to imaging transformations. *Cereb. Cortex* **3**, 261–269.

[36] Wallis, G. and Rolls, E. (1997). A model of invariant object recognition in the visual system. *Prog. Neurobiol.* **51**, 167–194.

[37] Wang, G., Tanaka, K., and Tanifuji, M. (1996). Optical imaging of functional organization in the monkey inferotemporal cortex. *Science* **272**, 1665–1668.

[38] Young, M. and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science* **256**, 1327–1331.