Similarity-based viewspace interpolation and the categorization of 3D objects

Shimon Edelman
School of Cognitive and Computing Sciences
University of Sussex, Falmer, Brighton BN1 9QH, UK
shimone@cogs.susx.ac.uk

Sharon Duvdevani-Bar
Department of Applied Mathematics
Weizmann Institute of Science, Rehovot 76100, Israel
sharon@wisdom.weizmann.ac.il

Abstract

Visual objects can be represented by their similarities to a small number of reference shapes or prototypes. This method yields low-dimensional (and therefore computationally tractable) representations, which support both the recognition of familiar shapes and the categorization of novel ones. In this note, we show how such representations can be used in a variety of tasks involving novel objects: viewpoint-invariant recognition, recovery of a canonical view, estimation of pose, and prediction of an arbitrary view. The unifying principle in all these cases is the representation of the view space of the novel object as an interpolation of the view spaces of the reference shapes.

Representation by similarities to prototypes

To recognize a previously seen object, the visual system must overcome the variability in the object's appearance caused by factors such as illumination and pose. It is possible to counter the influence of these factors, by learning to interpolate between stored views of the target object, taken under representative combinations of viewing conditions (Ullman and Basri, 1991; Poggio and Edelman, 1990). Routine visual tasks, however, typically require not so much recognition as *categorization*, that is, making sense of objects not seen before. Despite persistent practical difficulties, theorists in computer vision and visual perception traditionally favor the structural route to categorization (Marr, 1982; Biederman, 1987), according to which forming a description of a novel shape in terms of its parts and their spatial relationships is a prerequisite to the ability to categorize it.

Recent developments in the field suggest that knowledge of instances of each of several representative categories can provide the necessary computational substrate for the categorization of their new instances, as well as for representation and processing of novel shapes, not belonging to any of the familiar categories. The representational scheme underlying this approach, according to which objects are encoded by their similarities to entire reference shapes, is computationally viable (Edelman and Duvdevani-Bar, 1997b), and is readily mapped onto the mechanisms of biological vision revealed by recent psychophysical and physiological studies (see (Edelman, 1997) for a discussion).

The concepts of similarity and categorization are central to our theory of representation, which builds upon R. N. Shepard's (1968) notion of *"second-order" isomorphism*. According to Shepard, a mapping between the world and a representational system should preserve similarities among objects, rather than attributes of individual objects (as in a "first-order" isomorphism). If the pattern of similarities (i.e., proximities in some internal feature space; cf. Shepard, 1987) among representations reflects faithfully the pattern of similarities among their target objects, the system can carry out categorization, by assigning similar representations to the same category. Although this "perceptual" construal of categorization may be limited in its scope (Medin et al., 1993), it carries with it some advantages (Edelman and Duvdevani-Bar, 1997b). First, similarity-based categorization implies a mathematically appealing theory of representation. Encoding objects by their similarities to a few reference shapes or prototypes results in a representation that is low-dimensional and therefore easy to learn. Second, such representation will be veridical, if the response of the similarity estimation mechanism is approximately constant over different viewing conditions, and if it falls off gradually and monotonically for shapes that are progressively dissimilar from the the optimal stimulus. Such a similarity estimation mechanism is readily available in the form of the standard building block of connectionist systems – a trainable function approximation module.

Although there is evidence that our version of similarity-based categorization can be practical in computer vision (Edelman and Duvdevani-Bar, 1997a), its full potential can only be realized following the development of a theory of the uses of similarity information for object recognition and categorization (Goldstone, 1994). In this note, we discuss certain computational aspects of this problem, based on the idea of interpolation of image data derived from groups of similar objects and ordered by viewpoint. We also describe an implementation of this idea and its testing on a small set of common 3D objects (Figure 1).

Interpolation of viewspaces

Consider the multidimensional space of measurements performed by a visual system upon the world (this can be the space of photoreceptor responses, or, in the case of an artificial system, the space of pixel values in the image provided by the camera). A scene such as a view of an object corresponds to a single point in the measurement space, and a smoothly changing scene (e.g., a sequence of views of an object rotating in front of the observer) — to a smooth manifold that we call the *viewspace* of the object. The dimensionality of the viewspace depends on the number of degrees of freedom of the object; a rigid object rotating around a fixed axis gives rise to a one-dimensional viewspace (see the curve labeled $\mathcal{V}_1$ in Figure 2).

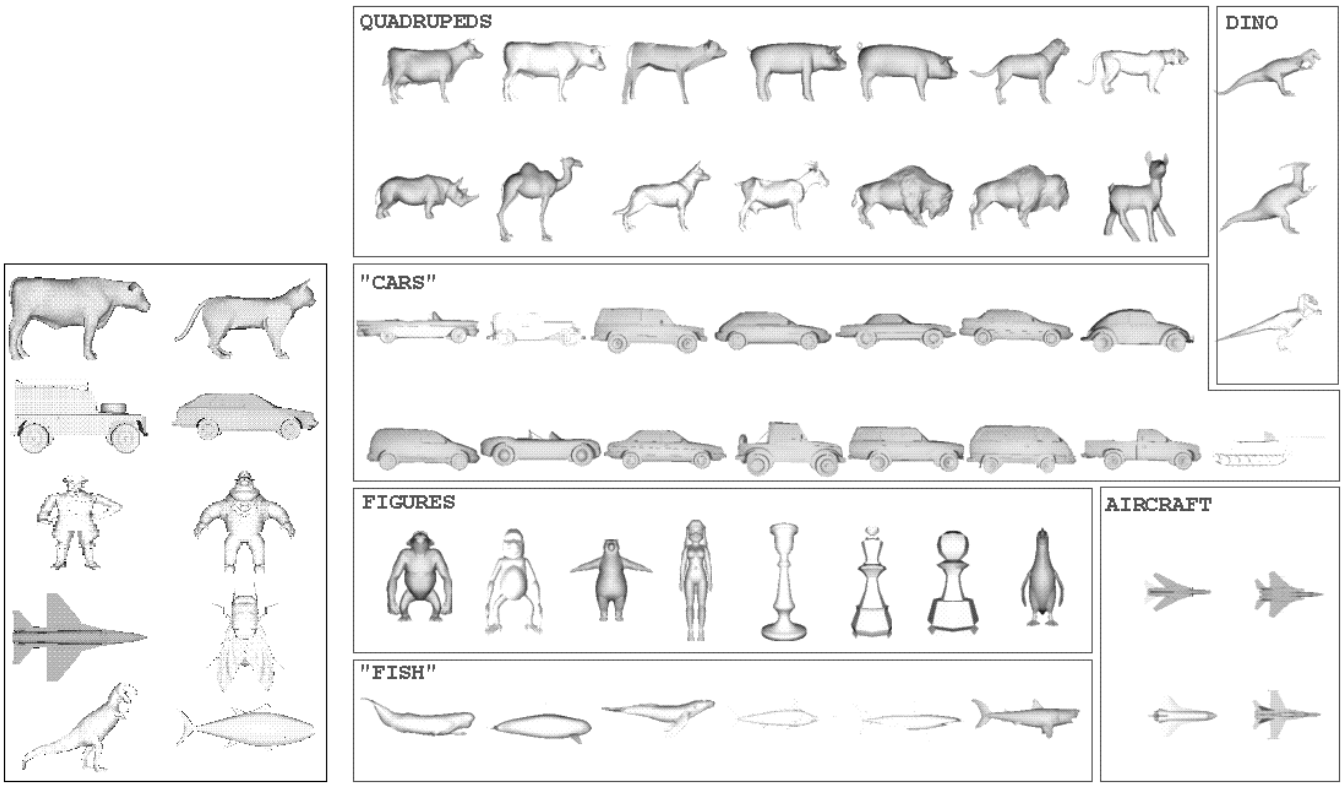

Figure 1: The 3D objects used in the present study. *Left:* the 10 reference shapes. *Right:* the 50 test shapes. All the objects are available from Viewpoint Datalabs (http://www.viewpoint.com).

Motivation

Given a transformation of an object, the structure of its viewspace is determined by the object's geometry.[1] The viewspaces of two nearly identical shapes will be very close to each other; a smooth deformation of the object will result in a concomitant smooth evolution of its viewspace. This observation can serve as the foundation for a principled treatment of novel objects. Specifically, a system that has internalized the viewspaces of a number of object classes can treat a view of a novel object intelligently, to the extent that it resembles the familiar objects (note that the concept of similarity is given here a concrete interpretation in terms of proximity in the measurement space). The computational mechanism whereby this can be done is *interpolation*.

Mechanism

The particular interpolation problem that arises here involves irregularly spaced data. Among the many methods developed for this case (Alfeld, 1989), the simplest one is inverse-distance weighting, due to D. Shepard (1968), in which the contribution of a known data point to the interpolated value at the test point is inversely proportional to the distance between the two. Unlike many other approaches to interpolation, Shepard's method does not require the solution of any large linear systems; its extension to many dimensions has been described in (Gordon and Wixom, 1978).

It should be noted that in our case the data "points" are actually entire manifolds – the viewspaces of the prototype objects. Accordingly, the viability of interpolation among viewspaces depends on the prior availability of a mechanism for dealing with viewspaces of individual objects. Because such a mechanism has been discussed extensively elsewhere (Edelman and Duvdevani-Bar, 1997b), we treat it here as a "black box" (cf. Figure 3) that can be trained to output a constant for different views of some target object (and a zero for views of other objects), or to estimate the pose of the target, or to transform one of the target views to another (e.g., to a "standard" or canonical view). All these possibilities are put to use in the next several sections.

Local viewpoint invariance for novel objects

Consider a system composed of k modules, each trained to output 1 for a number of representative views of some reference object. Note that the output of each such module can be interpreted as the similarity (i.e., proximity, or inverse distance) between the stimulus image and the module's target object. Now, suppose that the system is confronted with the i'th training object, which changes its shape smoothly ("morphs") into the j'th object. The output of the system, $\mathbf{x}$, will then undergo the following transformation: its i'th component, x_i,

[1] Certain features are common to viewspaces of all objects; e.g., the viewspace always closes upon itself as the object undergoes a complete rotation. Other features are peculiar to certain classes of objects; e.g., the viewspace of a rotationally symmetric object is a point; the viewspace of an object that possesses a mirror symmetry with respect to the axis of rotation crosses itself once.

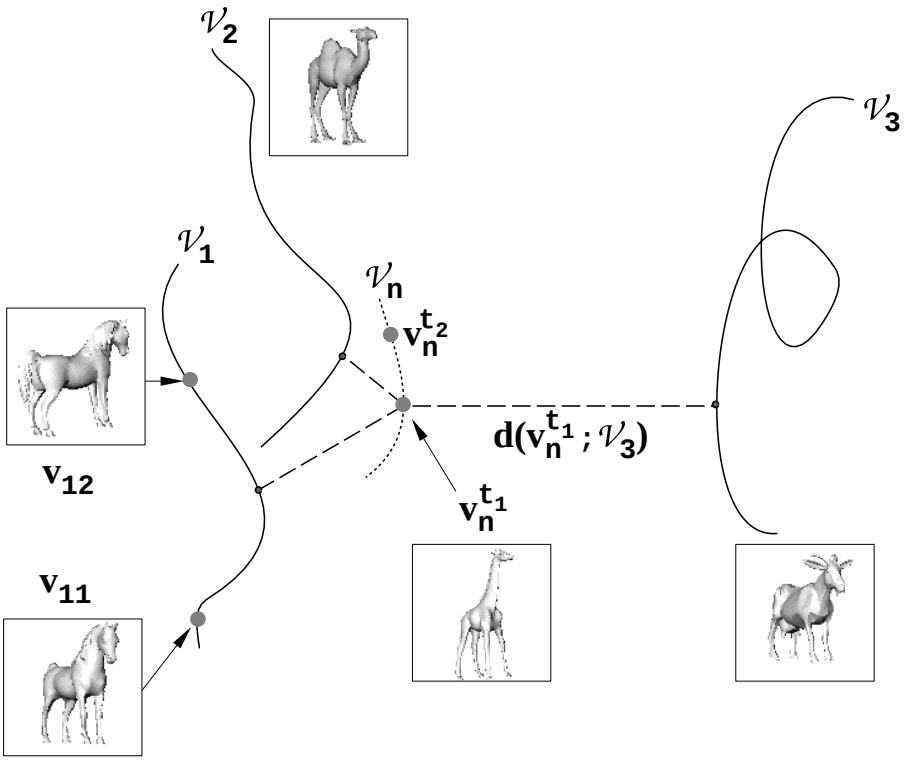

Figure 2: Interpolation of prototypical viewspaces. The change in the view (appearance) of an object unfamiliar to the system (in this example, `giraffe`) can be estimated by interpolating corresponding changes in the appearance of reference (prototype) objects (here, `horse`, `camel` and `goat`).

will change its initial value (≈ 1) to some $p < 1$, while the j'th component, x_j, will simultaneously change from some $q < 1$ to ≈ 1. Thus, the quantity $I \doteq \left(x_i + x_j\right)$ will remain approximately equal to 1 throughout this manipulation (insofar as the i'th and the j'th modules have been trained successfully).[2]

More generally, the output of the i'th module for a given test view $\mathbf{v}_n^t$ of a novel object, $x_i(\mathbf{v}_n^t)$, can serve as an indicator of the relevance of the i'th prototypical viewspace $\mathcal{V}_i$ to estimating the structure of the viewspace of the novel object $\mathcal{V}_n$. Consequently, the weight of $\mathcal{V}_i$ in determining the shape of $\mathcal{V}_n$ should be set to $x_i(\mathbf{v}_n^t)$.

We now apply this principle to the computation of a quantity Y that is intended to remain constant over changes in the test view $\mathbf{v}_n^t$ of a novel object. First, we compute the vector of responses of the k modules to a test view t_1; denote it by $\mathbf{w} = \mathbf{x}(\mathbf{v}_n^{t_1})$. Now, the estimate of Y for another test view t_2 is $Y\left(\mathbf{v}_n^{t_2}\right) = \mathbf{w}^T\mathbf{x}\left(\mathbf{v}_n^{t_2}\right)$, where T denotes the transpose. Note that the weights are pre-computed for a certain input, then used for other inputs (i.e., in other parts of the input space). Clearly, $Y(\mathbf{v}_n^{t_2})$ will remain approximately constant, as long as the test view $\mathbf{v}_n^{t_2}$ is not too far from the view $v_n^{t_1}$ used to estimate the weights $\mathbf{w}$, and as long as the novel object is not too different from at least some of the reference ones.

[2]This observation is related to an old method for achieving invariance with respect to a group of transformations by summing over the elements of the group (Pitts and McCulloch, 1965), and to more recent ideas of (Nosofsky, 1988).

The results of an evaluation of this approach to object constancy, which, it should be stressed, works for novel objects, are shown in Figure 4.

Recovery of a standard view for novel objects

We now introduce a variation on the theme of the preceding section, by training each object-specific module to output a *standard view* $\mathbf{v}_k^s$ of its respective object (Poggio and Edelman, 1990). The standard view of a novel object $\mathbf{v}_n^s$ can be estimated from a given test view $\mathbf{v}_n^t$ by applying Shepard's interpolation method: $\mathbf{v}_n^s = \sum_k \mathbf{w}^T\mathbf{v}_k^s$, where the weights $\mathbf{w}$ are the same as in the preceding section. Note that this method requires a set of modules trained, as previously, to output a constant, in addition to the modules trained to output a standard view.

The performance of a system trained to recover standard views of reference objects is illustrated in Figure 5. Note that the system was tested on 50 novel objects, in addition to the 10 reference ones. As one may expect, the performance of this method improves if the novel object belongs to the same category as the reference objects (Figure 5, bottom left).

Recovery of pose for novel objects

The information concerning the pose (i.e., orientation) of an object is contained in its image, provided that the object's shape is familiar, or that it can be related to some familiar shapes in a principled manner. We relied on this observation and on the method of Figure 3 to estimate the pose of novel

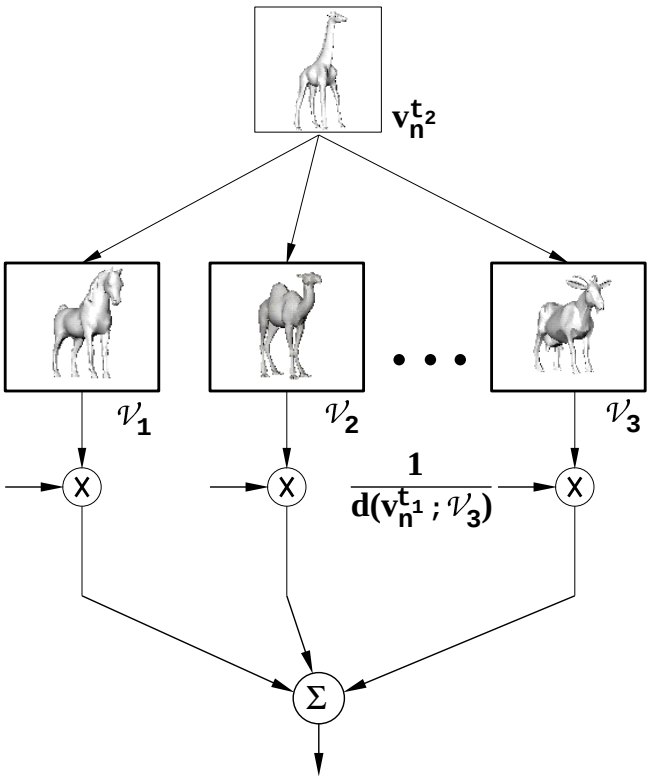

Figure 3: A mechanism for the interpolation of prototypical viewspaces. The inverse-distance weighting method of (Shepard, 1968a) is used to combine the outputs of a few "black boxes" — classifiers tuned to various prototypical or reference objects (Edelman and Duvdevani-Bar, 1997b). This scheme can estimate the viewspace of a novel object by interpolation of the familiar ones. As a result, it can support a range of tasks related to categorization, as described in the text.

objects, using a system trained to recover the pose of reference objects. The performance of this method is illustrated in Figure 6. As in the estimation of a standard view, the performance improves if the novel objects belong to the same category as the reference ones (Figure 6, bottom right).

Prediction of view for novel objects

The last experiment in this series examined the ability of the viewspace interpolation method to support the prediction of a novel view of a novel object, given its "standard" view. Intuitively, this corresponds to an attempt to guess what a novel object would look like from an unfamiliar viewpoint, a task that people are not very good at, if the objects are totally unfamiliar (Rock et al., 1989). The performance of our method in this experiment is illustrated in Figure 7.

Discussion

Viewspace interpolation on the basis of similarity, or inverse-distance weighting, can support a variety of visual recognition, categorization and prediction ("imagery") tasks. This method is related to a number of previously examined approaches. (Poggio and Edelman, 1990) demonstrated both pose and standard-view recovery for wireframe objects (their system was not tested on novel objects). (Lando and Edelman, 1995) averaged the viewpoint transformations of a number of objects to recover the standard view of a novel member of the same category (human faces). Likewise, (Beymer and Poggio, 1996) used the viewspace of one face to predict the appearance of unfamiliar views of another face. This approach is related to Basri's (1996) two-stage recognition algorithm, in which an input view is first associated with the most similar prototype (object class), then mapped into a standard view using a transformation specific for that class.[3]

We extend these earlier approaches, as follows: (1) we use interpolation among several viewspaces, rather than averaging or selection of the one nearest to the test view of the novel object; (2) we demonstrate the applicability of our method to a relatively wide range of shaded 3D shapes. Much work remains to be done in consolidating these results, exploring their computational underpinnings, and following up their psychophysical predictions. One implication for a cognitive theory of visual categorization can already be identified: categorization (i.e., making sense, in various ways) of a novel object does not seem to require the recovery of its 3D structure. Instead, it can rely on interpolation of cues from familiar objects, in proportions that are guided by their similarities to the novel one.

We conjecture, further, that the same principle can support

[3] All these methods, except (Lando and Edelman, 1995), rely on detailed correspondence information. In such methods, features in the input image and in the stored representations must be matched before any further processing can be done. The correspondence problem for objects such as those in Figure 1 remains at present largely unsolved.

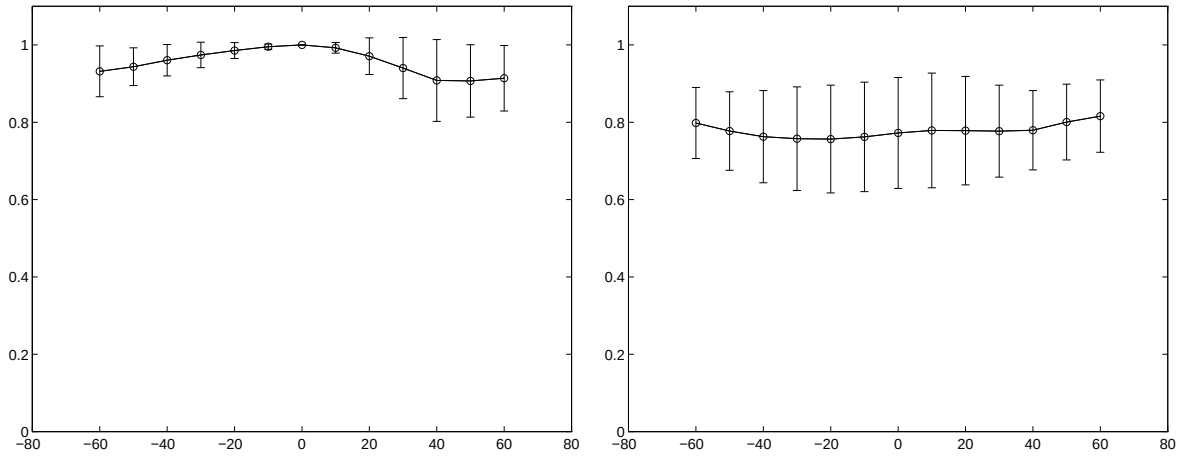

Figure 4: Local viewpoint invariance. The plots show the weighted sum of activities of 10 modules tuned to reference objects (Figure 1, left), evoked by different views of a test object; the data are the means and the standard errors computed over 50 test objects (Figure 1, right). The computation followed the interpolation method illustrated in Figure 3. Specifically, the vectors of the 10 module responses to the single initially available view of each of the test objects had been pre-computed, then used as the sets of weights in the generalization test stage. During testing, the system computed the weighted sum of the 10 module responses using different sets of weights in turn. *Left:* The output with the weights pre-computed for a standard (canonical) view of each test object. *Right:* The output with the weights computed for one of the test objects (`camel`), then used for the other 49 objects. Note that the output on the right is consistently less than on the left. One way to use this result for categorization is to let the weight set that yields the highest output sum determine to which object to attribute the test view.

dynamic (incremental) learning of object viewspaces. Given a stimulus, the similarity-based interpolation mechanism of Figure 3 can be used to decide whether or not it is a view of an object never seen before. If it is, viewspaces of familiar objects can be interpolated to estimate the viewspace of the novel one. Repeated exposure to views of the same object can then lead to a formation of a dedicated representation in long-term memory. Subsequently, storing additional views of a familiar object can serve to improve the performance of its dedicated module, and its contribution to the categorization of novel shapes.

References

Alfeld, P. (1989). Scattered data interpolation in three or more variables. In Lyche, T. and Schumaker, L., editors, *Mathematical Methods in Computer Aided Geometric Design*, pages 1–33. Academic Press, New York.

Basri, R. (1996). Recognition by prototypes. *International Journal of Computer Vision*, 19(147-168).

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

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

Edelman, S. (1997). Representation is representation of similarity. Behavioral and Brain Sciences, to appear.

Edelman, S. and Duvdevani-Bar, S. (1997a). A model of visual recognition and categorization. *Phil. Trans. R. Soc. Lond. (B)*, 352(1358):1191–1202.

Edelman, S. and Duvdevani-Bar, S. (1997b). Similarity, connectionism, and the problem of representation in vision. *Neural Computation*, 9:701–720.

Goldstone, R. L. (1994). The role of similarity in categorization: providing a groundwork. *Cognition*, 52:125–157.

Gordon, W. J. and Wixom, J. A. (1978). Shepard's method of 'Metric Interpolation' to bivariate and multivariate interpolation. *Mathematics of Computation*, 32:253–264.

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

Marr, D. (1982). *Vision*. W. H. Freeman, San Francisco, CA.

Medin, D. L., Goldstone, R. L., and Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100:254–278.

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.

Pitts, W. and McCulloch, W. S. (1965). How we know universals: the perception of auditory and visual forms. In *Embodiments of mind*, pages 46–66. MIT Press, Cambridge, MA.

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

Rock, I., Wheeler, D., and Tudor, L. (1989). Can we imagine how objects look from other viewpoints? *Cognitive Psychology*, 21:185–210.

Shepard, D. (1968a). A two-dimensional interpolation function for irregularly spaced data. In *Proc. 23rd National Conference ACM*, pages 517–524. ACM.

Shepard, R. N. (1968b). Cognitive psychology: A review of the book by U. Neisser. *Amer. J. Psychol.*, 81:285–289.

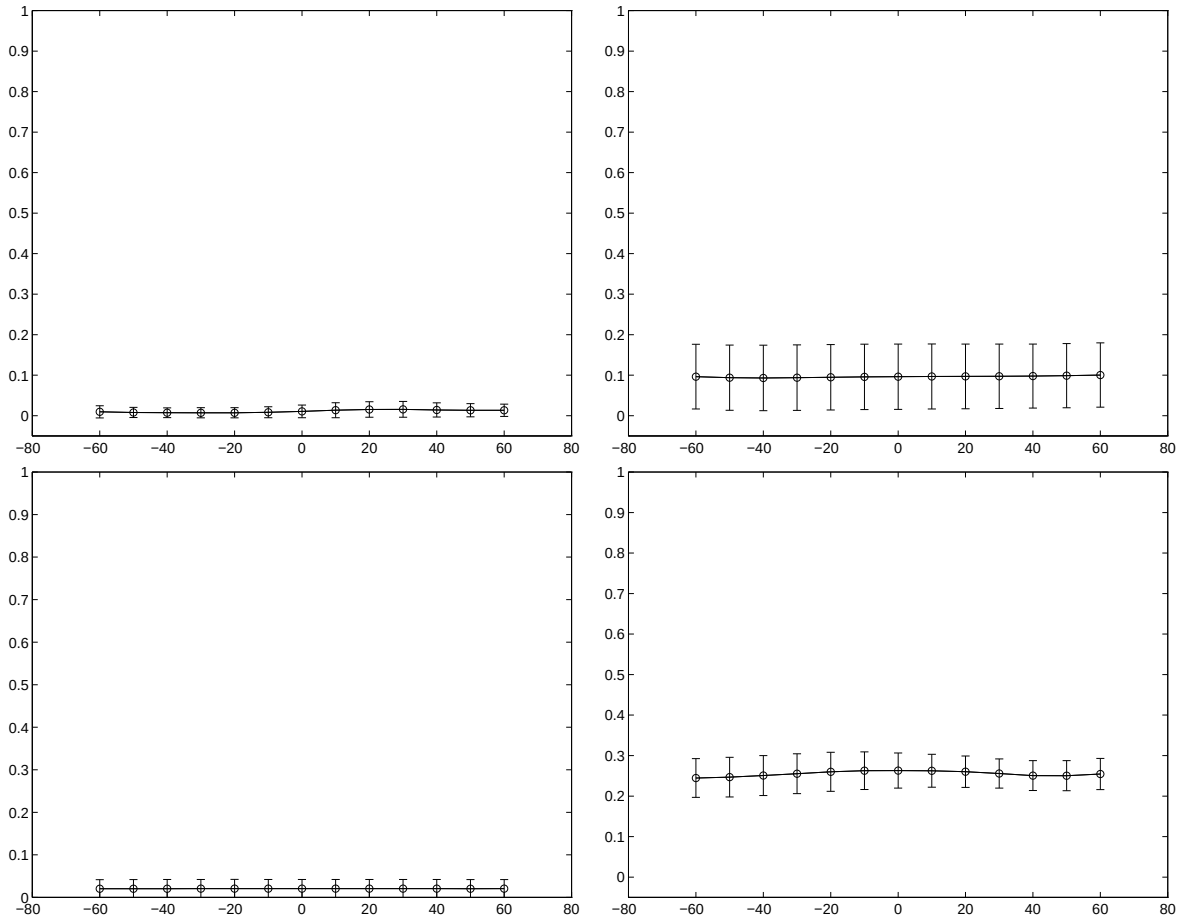

Figure 5: Recovery of a standard view. Each of the 10 modules was trained to output the standard view of one of the reference shapes. Then, for 13 test views (spaced at 10° around the canonical view) of each of the test objects, the *distance* between the estimated standard view of that object and its true standard view, $d = \cos\left(\hat{\mathbf{v}}^s, \mathbf{v}^s\right)$, was calculated and plotted against the angular distance between the test and the standard view. *Top left:* performance on views of the 10 training objects. *Top right:* performance on the 50 test objects. *Bottom left:* performance on views of the 20 objects of the CARS category. Here, the results are significantly better, because the novel objects belong to the same category as the reference objects. *Bottom right:* distance between the recovered view and the standard view of a "wrong" object (randomly chosen out of the 50 test objects); the distances in this control plot are consistently larger than in the other plots in this figure.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323.

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

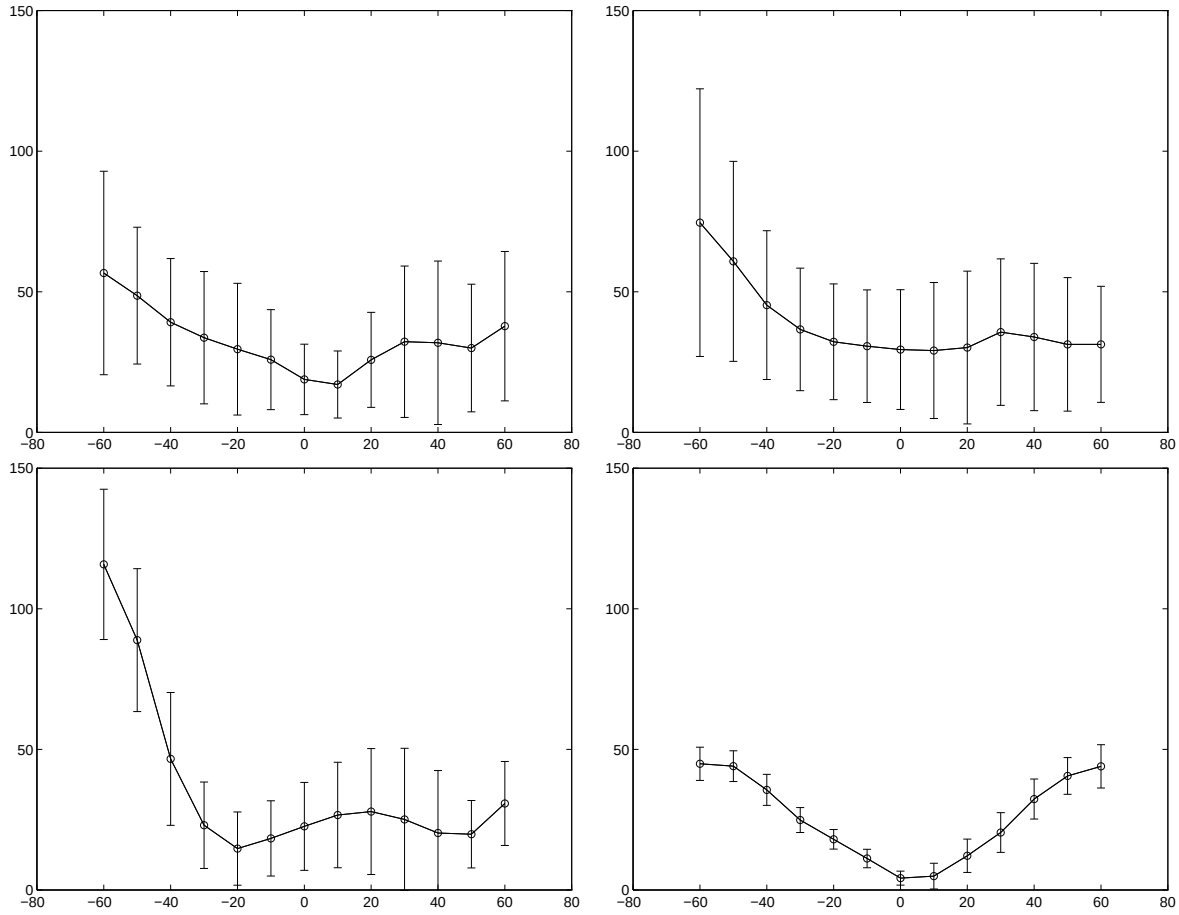

Figure 6: Recovery of pose. The system was trained to estimate the orientation of an object from its view. The plot shows the difference between the estimated pose and the true one, for several values of pose (ranging between $-60°$ and $60°$, $10°$ apart). *Top left:* results for the 10 training objects. *Top right:* results for the 50 test objects. *Bottom left:* results for the 20 objects of the CARS category. *Bottom right:* recovery of the pose of the 20 CARS, based on the outputs of the two CAR modules. As in Figure 5, performance is better for novel objects that belong to the same category as the reference ones.

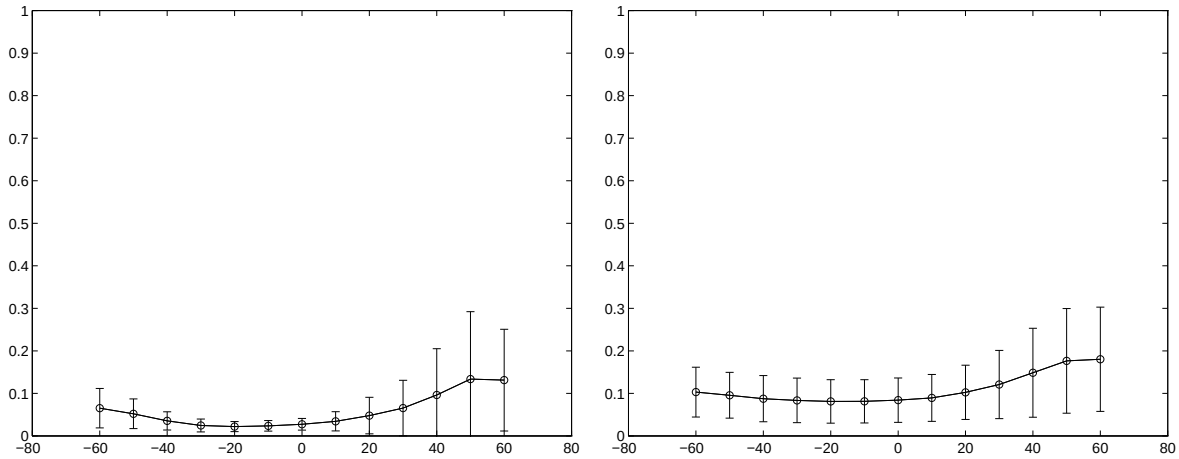

Figure 7: Prediction of transformed views. The plot shows the cosine distance between predicted and true views of an object, for several pose values, ranging between $-60°$ and $60°$ at $10°$ intervals. *Left:* the distance averaged over the 10 reference objects. *Right:* the distance averaged over the 50 test objects.