# The Role of Canonical Views in 3-D Object Recognition: Psychophysical and Computational Approaches

**Takafumi Sasaoka** Kanazawa Institute of Technology
**Toshio Inui** Kyoto University

Many researchers have reported that 3-D objects are recognized more readily from certain perspectives (*canonical views*; e.g. Palmer et al., 1981). If object recognition is assumed to be a process of matching an input image with object representations stored in the brain, it is possible that canonical views are stored as object representations. In this study, we conducted psychophysical, recognition experiments using computer-generated novel 3-D objects called paper-clip objects. The results demonstrated the presence of canonical views that were recognized best, and with high consistency, by most of participants. We trained GRBF networks (Poggio & Edelman, 1990) to recognize the objects used in the psychophysical experiments. We found that the representations corresponding to the psychological canonical views were learned in the hidden layer in the case of three hidden units. Moreover, these units have a larger generalization range than other units. It is concluded that these representations enabled the generalization of more views using less memory resources.

Keywords: canonical views, GRBF network, view-generalization.

## Introduction

The shapes of 3-D objects vary almost infinitely depending on the viewpoint of the observer. In spite of this, however, we can recognize objects independently of our perspective. The mechanism of view invariance is still a matter of debate. To clarify this issue, we focused on the presence of 'canonical view', reported in a number of studies (e.g. Palmer et al., 1981; Blantz et al., 1999).

It is known that recognition is best facilitated when 3-D objects are seen in a certain perspective, which has been termed the canonical view. Palmer et al. (1981) conducted several cognitive tasks using common objects, such as rating the goodness of different perspectives and analyzing the perspective that participants imagine first. Their results were highly correlated and corresponded across participants. They operationally defined 'canonicalness' as a variable to explain these results and also reported that RT in a naming task decreased monotonically according to the canonicalness of a view. If object recognition is assumed to be a process of matching an input with representations of the object stored in the brain, then the characteristics of canonical views should be related to object representations in the brain. In this study, we attempted to clarify the role of canonical views in object recognition through psychophysical experiments and network simulations.

## Psychophysical Experiments

We conducted a recognition experiment and a canonical view selection task.

### Stimuli

We used computer-generated novel objects called paper-clip objects (Bülthoff & Edelman, 1992) as stimuli. By using such objects, we can investigate the properties of canonical views, while avoiding the effects of past experiences with the objects. All objects were constructed by randomly generating eight points in a unit cube with the same length and connecting these points sequentially with sticks (dodecagonal prisms) rendered with shading, with the hidden surfaces removed.

### Recognition experiment

The recognition experiment consisted of two phases; a training phase and a test phase. In the training phase, a target object that was rotating horizontally at the speed of 10sec/rotation was presented for 30sec. The participants were required to memorize this object. In the test phase, a static test view was presented. The participants indicated whether it was the target or a distractor by pressing a response key. Although the test view was presented until a response was made, participants were instructed to respond as quickly and as accurately as possible. The target view presented in the test phase was one of 36 perspectives of the object sampled in 10º increments around the vertical axis starting from a perspective of 0º, which was arbitrarily chosen for each target. After the response, the next test view was presented. There were 144 trials and 12 different target objects. Seven participants were assigned to each target.

### Canonical view selection task

After the recognition experiment, the target object was presented again rotating horizontally. Participants were asked to select one view which was most typical for the object. They stopped the rotation and adjusted it to the most typical view by pressing the keys of a keyboard. No time limit for selection was set.

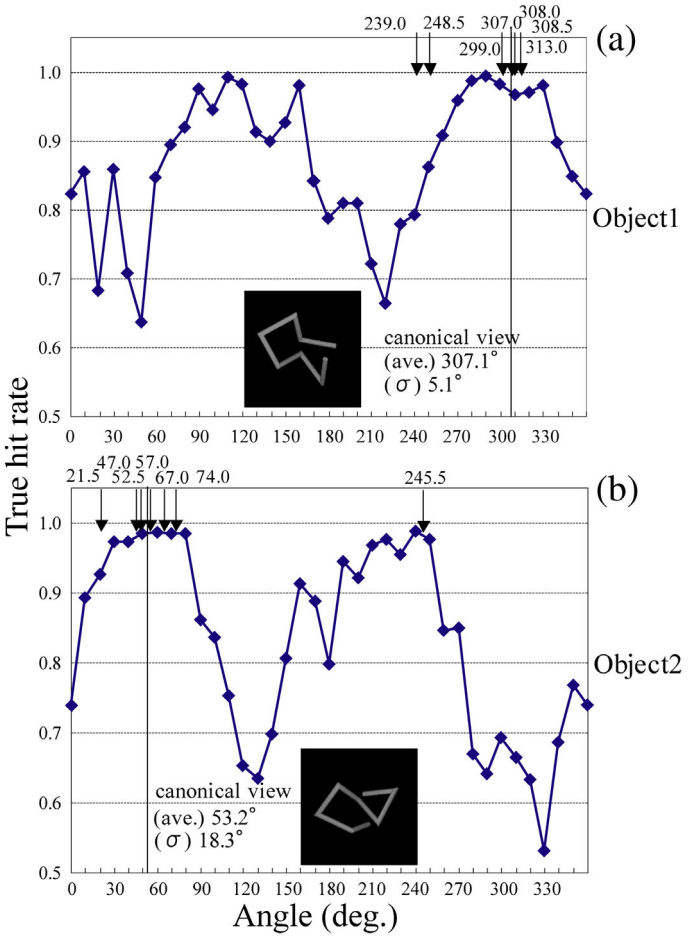


Figure 1 Part of results of the recognition experiment: arrows on the top of each plot represent the views selected by participants. The canonical view of each target is shown imposed on each figure.

## Results

Figure 1 shows the results for two targets, which are typical for all targets. Recognition performance varied view-dependently, in spite the fact that all views were presented in the training phase. The arrows on the top of each plot represent the views selected by participants in the canonical view selection task. The arrows were concentrated approximately in the proximity of views with high true hit rates and they were organized into one or two clusters. In the case of a single cluster, we calculated its mean, and when there were two clusters, as with the two examples in Figure 1, we calculated the mean of the larger cluster and defined it as the canonical view. In Figure 1, we have shown two canonical views that have all the typical characteristics of canonical views of objects used in the experiment: clearly visible sticks and the length of the long axis being longer than the others. Moreover, these views were recognized more accurately than other views. Therefore, these results are consistent with Palmer et al's findings.

## Simulation

To what extent is the canonical view explicable by geometric factors? To clarify this issue, we conducted a simulation experiment using a version of GRBF networks (Generalized Radial Basis Function networks; Poggio and Edelman, 1990).

The scheme of radial basis functions was originally applied to approximate multivariate functions. Poggio and Edelman (1990) regarded object recognition task as a type of the interpolation of the multivariate function of feature points in a perspective. They trained the network to output the standard view of a target when the target view was input. During training, a set of views was learned in centers of Gaussian basis functions in the hidden layer. If a sufficient number of units were in the hidden layer, the network could output the standard view when any view of the target was input. However, Poggio and Edelman (1990) did not refer to the representations that were acquired in hidden units. In our simulation, we investigated whether the representations in the hidden units corresponded to the results of the psychophysical experiment.

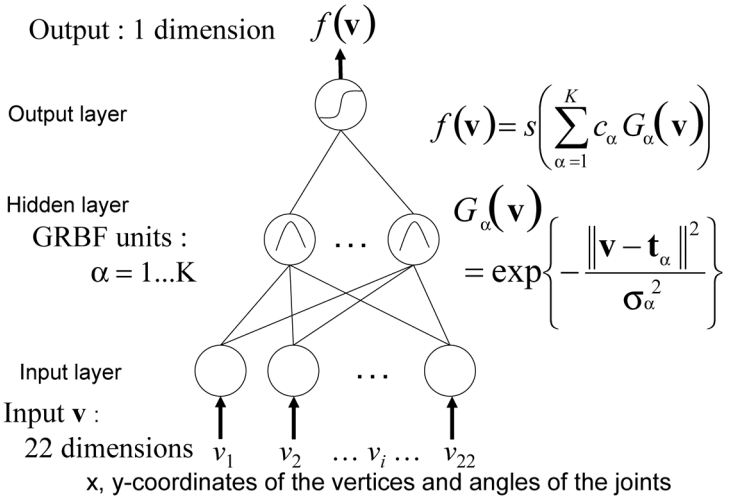


Figure 2 The architecture of the model used in our simulation

## Architecture of the network

Our network consisted of three layers (Fig. 2). The first layer had 22 input units. In this network, a view of an object was represented by a 22 dimensional vector $\mathbf{v} = (v_1, v_2, \cdots, v_{22})$ consisting of x, y-coordinates of vertices (16 dimensions) and angles formed by two adjacent sticks in the picture plane (6 dimensions). In the hidden layer, when a view $\mathbf{v}$ is input, the output of a GRBF unit can be written as

$$G_\alpha(\mathbf{v}) = \exp\left\{-\frac{\|\mathbf{v}-\mathbf{t}_\alpha\|^2}{\sigma_\alpha^2}\right\} \quad (1)$$

where $\alpha$ is an index of GRBF units, and $\mathbf{t}_\alpha$ is the center vector of the $\alpha$th unit. Finally, the weighted sum of outputs

of GRBF units is input into an output unit. The unit has a sigmoidal output function, so that it takes a value from 0 to 1. Hence the output of the network is given by

$$f(\mathbf{v}) = s\left(\sum_{\alpha=1}^{K} c_\alpha G_\alpha(\mathbf{v})\right) \quad (2)$$

where s (·) represents a sigmoidal function and $c_\alpha$ is the weight for the connection between the αth GRBF unit and the output unit.

In the training phase, the network was trained to output 1 when the view of the target was input. The training set consisted of 36 views of the target, the same as the set used in the test phase of the recognition experiment. Weights and centers were initialized by random values before the training. During the training, these parameters were updated to minimize the error function according to the gradient descent method until the error was less than a certain small value ($1.0*10^{-5}$). After the training, the network was tested by the testing set. The testing set consisted of 360 views of the target and 10 distractors sampled randomly from distractors used in the test phase of the recognition experiment.

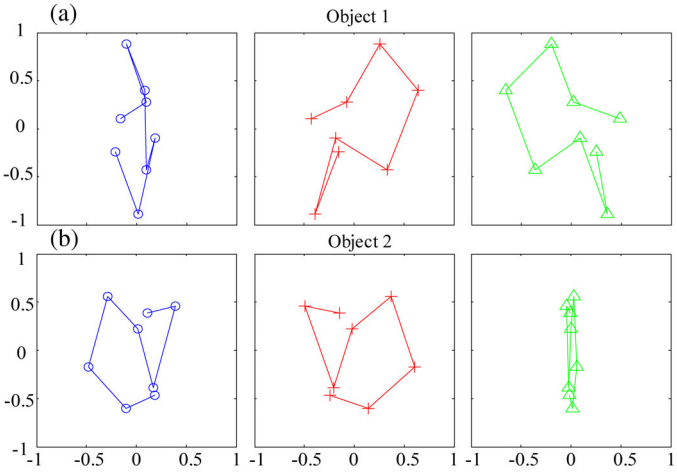


Figure 3. Representations in the hidden units: (a) Object 1 and (b) Object 2

## Results

The number of hidden units (*K*) was varied in the range from 2 to 16. In the case of *K* = 3, we obtained representations with properties corresponding to the results of the psychophysical experiment. Figure 3 shows an example of representations acquired by the hidden units. To simplify, these are illustrated by using only elements of coordinates of the center vector (16 dimensions). Figure 4 shows the activation profiles of each unit. Although there were some variations in the results, depending on the initial values, the results were robustly similar. The representations acquired by hidden units could be classified into two classes; 'canonical view units' and 'non-canonical view units'. Comparing Figure 3 with Figure 1, it is clear that the representation of the unit shown in the right column of Figure 3 (a) is similar to the canonical view for Object 1 in Figure 1 (a). Similarly, the unit shown in the left column of Figure 3 (b) suggests that the network learned a similar representation to the canonical view of Object 2 in Figure 1.

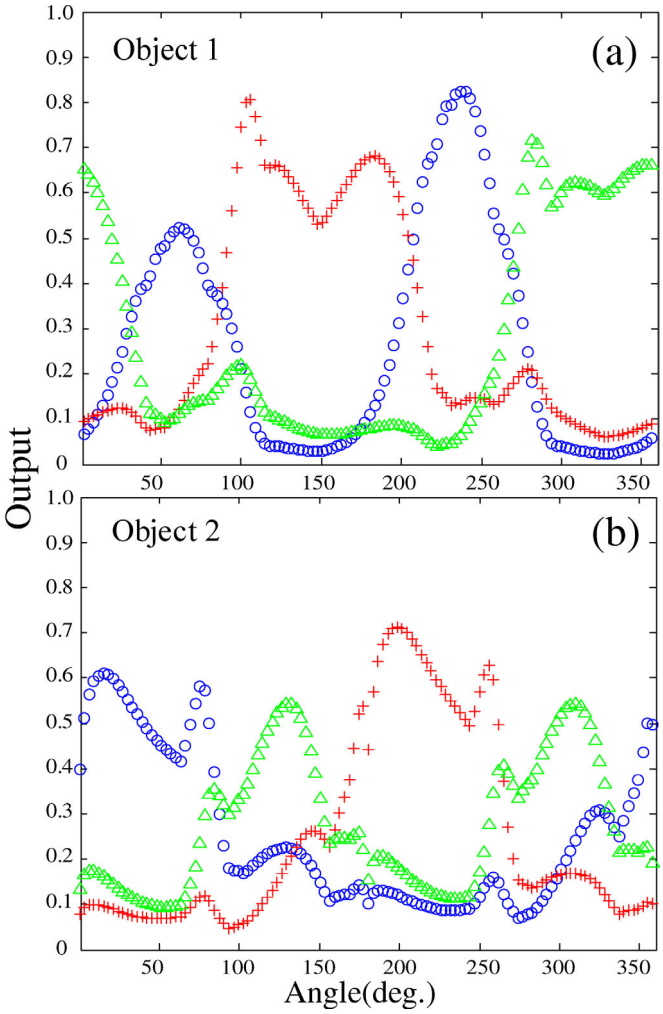


Figure 4. Profiles of outputs of each hidden unit: (a) Object 1 and (b) Object 2. Each plot corresponds to the plot with the same symbol in Figure 3.

From Figure 4, it is clear that these units tuned to the proximity of the canonical view. There were also units tuning to the view 180º opposed to the canonical view, i.e., the middle column of Figure 3 (a) and Figure 3 (b). These views are mirror-symmetrical to the canonical view, for which true hit rates were as high as for the canonical view. Activation profiles of these units tuning to the canonical view and its mirror-symmetrical view were very similar to the recognition performance (see Figures 1 and 4).

On the other hand, the units tuning to non-canonical views i.e., Figure 3 (a) left and Figure 3 (b) right, activated with two peaks that are separated by 180º (see Figure 4). As can be seen in Figure 3, the width of these representations was narrower than the actual views. For Object 2, the width of the representation was collapsed. However, the angles were near the actual values (not shown).

In the case of two objects, although the units tuning to the canonical view appeared, their activation was similar to non-canonical view units. In the case of $K = 4$, however, canonical view units appeared for these objects.

Generalization range of a unit can be defined by the sigma of the Gaussian fitted to an activation function of each unit. We averaged sigmas of canonical view units and non-canonical view units over all targets. The mean sigma value of canonical view units was 43.9º. In contrast, that of non-canonical view units was 27.9º.

We defined the discrimination performance of the network by the error when target views were input, $E_\alpha$ ($=1-f(\mathbf{v}_{target})$), with when distractor views were input, $E_\beta$ ($=1-f(\mathbf{v}_{distractor})$). We compared the mean $E_\alpha$ with the mean $E_\beta$ across all distractors. The mean $E_\beta$ was much larger than the mean $E_\alpha$ (about $2.5*10^4$ times larger even in the smallest case). However, there were a few cases in which $E_\alpha$ for some views was larger than the $E_\beta$ for views of some distractors.

## Discussion

Results of both the psychophysical experiment and the simulation study indicate that representations corresponding to the psychological canonical views were learned in the hidden units of the GRBF networks. This suggests that participants in the recognition experiment may have stored representations similar to these views. Moreover, the generalization range of these units was about 10º larger than that in other units. This indicates that these views are 'non-accidental views', that are robust against horizontal rotation. Logothetis et al. (1995), in their electrophysiological study using paper clips consisting of the same number sticks as ours, have reported that the generalization range of view-tuned neurons in the IT (Inferior Temporal cortex) of Macaques was about 30º. Compared to this, it seems that canonical view units have a relatively large generalization range. This suggests that storing such representations enables generalization of a larger number of views. However, it has been pointed out that paper-clip objects are too special to explain common object recognition (e.g. Biederman, 2000). Our results suggest that canonical views are explicable by geometric factors such as robustness against rotation. Therefore, the results may also be applicable to common objects. The canonical views in our results were the perspectives in which the 3-D structure of the object could be clearly perceived. Such structural information may play some role even in the recognition of paper-clip objects.

For two targets, however, four hidden units were needed to obtain the same results as those of other targets. In the canonical view selection task, the selected views for these objects had a larger spread than for other objects. These results may be explicable if the change in the width depending on the horizontal rotation was relatively small for these objects. It might be necessary to take the complexity of the shape of objects into consideration in the quantitative analysis of these effects.

Non-canonical view units learned a representation in which the x-coordinates gathered on the y-axis. Representations of angles, however, were near the actual values. These units activated with two peaks that were 180º apart. Between these views, angles of the joints were identical. In contrast, the sign of the x-coordinates was reversed. These results indicate that such units tuned to the angles of the view rather than to the coordinates, probably due to that we used only the horizontal rotation. Under ecological viewing conditions, the most common rotation is the horizontal rotation (Lawson, 1999). In this sense, our choice of horizontal rotation was reasonable. Different rotations should be considered in future investigations.

One problem with this study is that the scheme and object representations used in our simulation were too simple to explain the complete mechanism of object recognition. In spite of this, our simulation results were able to predict the psychological canonical view of objects by taking into consideration simple variables, such as the visible length of the long axis of objects.

## Conclusions

In this study, both the psychophysical and the computational analysis suggested that the canonical view is the most economical representation for object recognition. These results suggest that the canonical view plays a key role in the representation of objects within the brain.

## References

Biederman, I. (2000). Recognizing depth-rotated objects: A review of recent research and theory. *Spatial Vision*, 13, 241-253.

Blanz, V., Tarr, M.J., & Bülthoff, H.H. (1999). What object attributes determine canonical views? *Perception*, 28, 575-599.

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

Lawson, R. (1999). Achieving visual object constancy across plane rotation and depth rotation. *Acta Psychologica*, 102, 221-245

Logothetis, N.K., Pauls, J., & Poggio, T. (1995). Shape representation in the Inferior Temporal cortex of monkeys, *Current Biology*, 5, 552-563

Palmer, S.E., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. *In* Long, J. & Baddeley, A.(ed.), *Attention and Performance*, Vol.9, Hillsdale, NJ: Erlbaum, 135-151.

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