In: Proceedings of the Seventh International Joint Conference on Artificial Intelligence, Vancouver, B.C. Canada, 1981

A PARALLEL COMPUTATION THAT ASSIGNS CANONICAL OBJECT-BASED FRAMES OF REFERENCE

Geoffrey E. Hinton

MRC Applied Psychology Unit
Cambridge, England

ABSTRACT

A viewpoint-independent description of the shape of an object can be generated by imposing a canonical frame of reference on the object and describing the spatial dispositions of the parts relative to this object-based frame. When a familiar object is in an unusual orientation, the deciding factor in the choice of the canonical object-based frame may be the fact that relative to this frame the object has a familiar shape description. This may suggest that we first hypothesise an object-based frame and then test the resultant shape description for familiarity. However, it is possible to organise the interactions between units in a parallel network so that the pattern of activity in the network simultaneously converges on a representation of the shape and a representation of the object-based frame of reference. The connections in the network are determined by the constraints inherent in the image formation process.

I INTRODUCTION

People can recognise a familiar spatial structure from a novel viewpoint. There is considerable evidence that they do this by imposing a canonical, object-based frame of reference and generating a description of the spatial structure relative to the assigned frame [1 2 3]. A tilted square, for example, can still be seen as a square because relative to the tilted frame that is imposed on it, it still has two vertical edges on either side and horizontal edges at top and bottom.

The central problem in using object-based frames is to devise a way of assigning the appropriate frame to a perceived object. This is not an easy task, even if sources of information like stereo, shape-from-shading, or optical flow have yielded the precise 3-D structure of the object relative to the viewer-centered frame of reference. Heuristics like planes of bilateral symmetry, gross elongation, and the gravitational or contextual vertical can help to suggest candidate object-based frames, but the final choice between the alternatives often depends on which object-based frame gives rise to a familiar shape description. An upside down table, for example, is seen as just that, because by seeing it as upside down we can see it as a familiar shape. In this case, there is clearly nothing in the image to indicate that an upside-down frame should be assigned.

This paper describes a way of using parallel hardware to implement a cooperative computation in which the process of choosing an object-based frame and the generation of a description relative to that frame occur simultaneously, with each influencing the other.

II FRAMES, FEATURES, AND MAPPINGS

Following Marr [4], I shall assume that early visual processing yields a representation of a scene in terms of 3-D features relative to a 3-D frame of reference defined by the retina (or camera). These "retina-based" features cannot be used directly for object recognition because they depend not only on the shape of the object but also on the spatial relationship between the retina and the object. Choosing an object-based frame of reference is equivalent to choosing a mapping from 3-D retina-based features to 3-D object-based features. This mapping compensates for the relationship between the retina and the object and thus yields features that are independent of this relationship. These features constitute a shape description that can be used for object recognition.

If we think in terms of a parallel system in which each feature corresponds to a specific hardware unit which is active when the feature is present, then a particular choice of an object-based frame must be capable of pairing each retina-based unit with a corresponding object-based unit, so that activity in the one can cause activity in the other. Fig. 1 shows such an arrangement for a simplified 2-D domain. There are many "channels" emanating from each retina-based unit. A choice of an object-based frame corresponds to activation of a particular "mapping" unit. An active mapping unit opens one channel from each retina-based unit to the object-based unit that is appropriate given that mapping. Thus, once the mapping has been selected, a pattern of active retina-based units

will cause a pattern of active object-based units, and this later pattern will constitute a description of the object's shape.

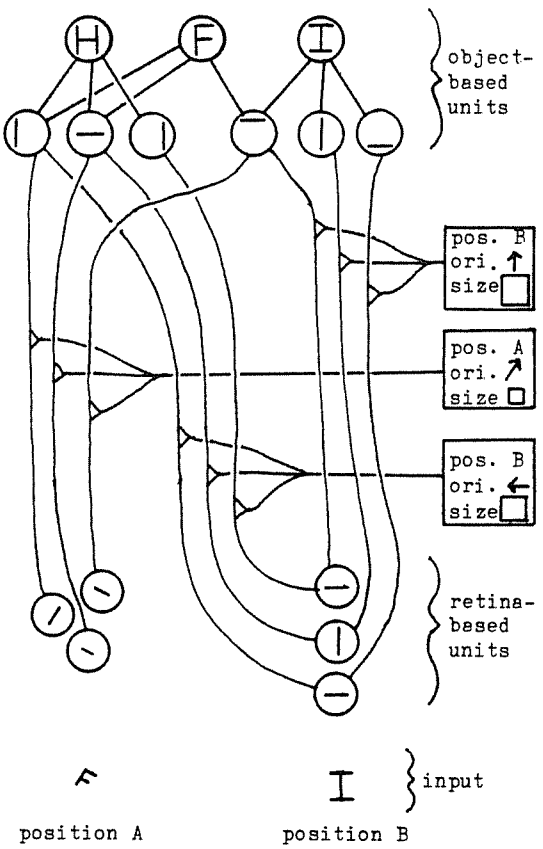

Figure 1. This shows how each mapping unit (large box on right) controls a set of channels from the retina-based units (bottom) to the object-based units (top). Only a few of the mapping units and channels are shown. The topmost units respond to combinations of active object-based units. The meaning of the triangular symbols is explained in the text.

III TOP-DOWN SELECTION OF A MAPPING

The meaning of the triangular symbol in fig. 1 is quite complex. It stands for two rules:

1. Multiply the activity level in the retina-based unit by the activity level in the mapping unit and send the product to the object-based unit.

2. Multiply the activity level in the retina-based unit by the activity level in the object-based unit and send the product to the mapping unit.

The first rule is what causes the appropriate mapping when a single mapping unit is fully active (activity varies from 0 to 1). The second rule selects a particular mapping when the activity of the object-based units is determined top-down. Fach possible pairing of an object-based feature with a retina-based feature of the same general type will send activity to a particular mapping. If the object is present in the input, the "correct" pairings will all agree on the same mapping, whereas the "cross-pairings" will not agree with one another. So the best way of instantiating the object in the scene will be indicated by the mapping with the greatest total input. This way of using an explicit representation of mapping space to accumulate evidence for a particular mapping is also used by Ballard [5].

The structure of the connectivity in this network captures the mathematical structure of a three-dimensional spatial relationship. There are constraints on the ways in which features relative to one frame can be paired with features relative to another frame if the pairings are to correspond to a single 3-D relationship between the two frames. Only certain sets of pairings are mathematically consistent, and each of these sets corresponds to a set of channels that are all connected to the same mapping unit.

IV SIMULTANEOUS SELECTION OF SHAPE AND MAPPING

I shall assume that the retina-based representation is rich enough to allow a roughly correct segmentation of a "figure" and also to allow a number of plausible object-based frames to be selected for the candidate figure using bottom-up heuristics. The evidence in favour of an object-based frame provides enough input to the relevant mapping unit to give it a low initial level of activity. The pattern of activity in the retina-based units is then mapped through all the candidate mappings simultaneously, but because the mapping units are not fully active, each mapping is attenuated (see rule 1 above) and so there are many slightly active object-based units. Among these object-based features there will be subsets which can be recognised as partial descriptions of familiar shapes.

The precise details of how shape descriptions are recognised as familiar is not central to the model. In Fig. 1 this mechanism is grossly simplified and shown as a set of shape-recognition units. All we need to assume is some kind of positive feedback mechanism which provides top-down support for familiar combinations of object-based features. So, among the many slightly active object-based units, certain subsets will receive more top-down

support than others. Once this happens, there is a runaway process. The pairings of above-average object-based features with retina-based features cause above-average support for the relevant mapping. This, in turn, causes less attenuation in the mapping from retina-based to object-based features. The more active a mapping unit gets, the more it contributes to its corresponding shape description, and vice versa. In a few iterations the system converges on a particular mapping and a particular shape description, and the inhibition between the mapping units (and also, optionally, between the shape units) causes the unsuccessful candidate mappings to be completely suppressed. Other things being equal, canonical frames are preferred to non-canonical ones, because of the top-down support for familiar combinations of object-based features.

V THE N^2 PROBLEM

If there are N retina-based units each of which can activate any of N object-based units, N^2 channels are required. This number can be reduced in various ways. First, N itself can be dramatically reduced before the mapping by distributed encoding of the individual features and mappings as patterns of activity in many different hardware units. Each unit represents a connected region in the space of possible features. A specific feature is coded by activity in all the units whose regions contain it. The number of effectively different features can then greatly exceed the number of units. This type of encoding is especially effective when the number of features present at any one time is much smaller than the number of possible discriminable features. There is not space here to include a formal treatment of the efficiency and limitations of this type of encoding.

Second, if the mapping is performed in K sequentially linked stages, the fan-out from a unit in one layer to the unit in the layer above need only be $\sqrt[K]{N}$, and since there are K layers only $K.N.\sqrt[K]{N}$ connections are needed. However, iterations take K times as long and different mappings inevitably become confused during the settling phase when many mappings are partially active. If, for example, there is one mapping to handle translation followed by another to handle rotation, it is impossible to allow just the mappings T_1R_1 and T_2R_2 to occur simultaneously, where T and R denote the translational and rotational constituents of a full mapping. The retina-based features will be mapped through both T_1 and T_2 and the resulting intermediate features will be mapped through both R_1 and R_2, and so the full mappings T_1R_2 and T_2R_1 will also have occurred.

VI CONCLUSION

I have outlined a way of using parallel hardware to implement a flexible mapping system that can perform a parallel search for a canonical, object-based reference frame. This is an important advance from the pandemonium or perceptron type of model because it handles the effect of viewpoint on the image in a principled way. There are, of course, other sources of variation in the image of an object, such as non-rigid transformations of the object itself, but there is no reason to lump these together with the effects of viewpoint. They are separate problems with their own separate structure. Finally, I should point out a major and controversial implicit assumption of the model. Object-based frames must be assigned one at a time, so that they can all share the same mapping apparatus.

REFERENCES

[1] Rock, I. Orientation and form. New York: Academic Press, 1973.

[2] Marr, D. & Nishihara, H. K. Representation and recognition of the spatial organisation of three-dimensional shapes. Proc. Roy. Soc. Series B, 1978, 200, 269-294.

[3] Hinton, G. E. Some demonstrations of the effects of structural descriptions in mental imagery. Cognitive Science, 1979, 3, 231-250.

[4] Marr, D. Representing visual information. In A. R. Hanson & E. M. Riseman (Eds.), Computer Vision Systems. New York: Academic Press, 1978.

[5] Ballard, D. H. Generalising the Hough transform to detect arbitrary shapes. Computer Science Department TR-55, 1979, University of Rochester.