NEURAL NETWORKS AND COMPUTATIONAL THEORY - SOLVING THE RIGHT PROBLEM

D C. PLAUT
plaut@cmu.edu

Follow this and additional works at: http://repository.cmu.edu/psychology

Published In
BEHAVIORAL AND BRAIN SCIENCES, 12, 3, 411-412.
Neural networks and computational theory: Solving the right problem

David C. Plaut
Computer Science Department
Carnegie Mellon University


J. J. Gibson's fundamental contribution to the study of perception was to shift emphasis away from physiology towards an analysis of the visual information that is available to behaving organisms. David Marr reexpressed this insight as the need, in understanding any information processing system, for a computational theory that describes what is being computed and why it is appropriate for the purposes of the task at hand, independent of how this computation is implemented. This perspective underlies much of the recent progress in both natural and machine vision research (Marr, 1982; Horn, 1986; Kanade, 1987; Richards, 1988).

In contrast, Strong and Whitehead (hereafter S&W) state at the outset of their article that neural network researchers "take human neurophysiology rather than overt behavior as their starting point" (p. 3). This statement foreshadows both the strengths and limitations of the model of human visual attention that they present. In particular, while S&W make important contributions to understanding the possible role of various neurophysiological mechanisms, the lack of a sufficiently general computational theory of high-level vision renders their proposed solution inadequate.

As S&W point out, the fundamental difficulty in high-level vision is handling spatial information appropriately. When presented with two visually identical objects, people can both recognize each object (ignoring their spatial variation) and discriminate between them (relying on their spatial variation). The central "solution" to this paradox presented by S&W consists of (1) representing objects as sets of non-spatial features, and (2) associating each of these (foveal) feature sets with the pattern of locations of extrafoveal stimuli to disambiguate otherwise identical objects. Unfortunately, both of these suggestions are flawed.

Representing objects with non-spatial features. The identity of an object can critically depend on purely spatial information. Let us distinguish between "global" spatial information (i.e. variation due to changes in viewpoint) and "local" spatial information (i.e. relationships between the parts of a single object). The difference between a square and a diamond (see Figure 1), or between a "p" and a "d" in lower case, depends solely on the relationship between the object and some larger frame like the page or environmental upright (Rock, 1973). Any object representation based on non-spatial features would be unable to distinguish between members of these pairs. Although people can be biased towards either interpretation, they have no trouble distinguishing between them. In addition, the spatial relations among the parts of an object provide important information during recognition, such as in Palmer's (1975a) "fruitface" (see Figure 2). Hence in achieving viewpoint-invariant recognition, spatial information cannot simply be discarded (e.g. by taking the logical OR of all spatially-indexed features of a certain type, as S&W suggest). Rather, global and local spatial information must be decoupled and represented separately (Pinker, 1985).
One way of accomplishing this is to redescribe the retinotopic feature information relative the the appropriate object-centered reference frame (Rock, 1973; Palmer, 1977; Marr & Nishihara, 1978; Hinton, 1981a). Recognition is invariant over viewpoint because the same object-centered description is derived for any instance of an object; the effects of viewpoint in each particular instance is "absorbed" by the reference frame. Since the object description is composed of spatially-indexed object-centered features, it retains the spatial relations among the parts of the object (enabling "fruitface" to be recognized), while the same retinotopic features can be interpreted as a square or diamond depending on the assumed relationship of the object to the retina (i.e. what object-centered frame is assigned). Hinton & Lang (1985) show how a parallel network that, under normal viewing conditions, correctly recognizes objects by assigning the appropriate object-centered frame, will produce illusory conjunctions when its retinotopic input is masked before the network has time to settle.

In Section 2.1, S&W argue against using spatially-indexed features for object recognition because "the problem of illusory conjunctions recurs at the level of scene integration" (p. 7). While it is true that an additional mechanism is still needed to prevent false bindings between objects and locations, this approach solves the retinotopic problem of feature integration so that the separate problem of associating objects with locations can be accomplished relative to a more stable reference frame (as described below).

**Using the pattern of extrafoveal locations as a spatial tag.** Any approach that derives a viewpoint-invariant representation of an object during recognition, regardless of whether it uses non-spatial or object-centered features, is faced with the problem of how multiple instances of the same object are represented and distinguished. Since these instances must differ in their environmental location, a natural approach is to "tag" each object description with additional information that is unique to each such location.

S&W propose to use as a spatial tag the arrangement of locations of extrafoveal stimuli that are visible when the object is fixated. While this information is relatively easy to derive and is useful for controlling eye movements, it is not sufficiently stable or unique to meet the requirements of human high-level vision. People are able to use their knowledge of the spatial arrangements of objects in the environment during tracking, navigation, and spatial reasoning. These abilities require a representation of object location that is stable over movement of other objects or the observer. The representation proposed by S&W fails on both of these accounts. Furthermore, it is incapable of distinguishing among regularly spaced objects. These limitations are not merely the result of simplifications that can be resolved by elaborations of the current approach, but rather they arise out of inadequate assumptions about the role of spatial information in high-level vision.

A more satisfactory spatial "tag" is an explicit representation of the actual environmental location of the object. Clearly this information is unaffected by object or observer motion, and would only need to be updated if the object itself moved (Hinton & Parsons, 1988). It also provides a natural way to relate the spatial tags of neighboring objects, which is important for using knowledge of previously recognized objects to bias the subsequent recognition of nearby objects (Palmer, 1975b). The location of the object in the environment can be computed by combining the relation of the object to the retina (i.e. its assigned object-centered frame) with the relation of the retina to the environment (i.e. the current direction of gaze) (Hinton, 1981b). S&W argue against the use of the environmental location of an object as its spatial tag based on the difficulty of combining and calibrating eye, head, and body position in calculating the environmental direction of gaze. Some of this difficulty can be alleviated by explicitly representing locations in intermediate reference frames, such as head-centered (Feldman, 1985) and body-centered (Mountcastle et al., 1975). In fact, contrary to S&W's claim that eye position is not
involved in computing spatial tags, Andersen, Essick & Siegal (1985) found that the response properties of neurons in area 7a of the macaque (which is critically involved in spatial vision) are well predicted by a model that assumes that this area combines a retinal stimulus location with eye position information to yield a representation of the location of the stimulus in head-centered coordinates.

In closing, let me caution against the unjustified claims of the superiority of neural network research over other approaches that S&W espouse in their opening paragraph. The success of AI, and particularly computer vision, has been far from "marginal" (p. 3). In fact, the abilities of conventional object recognition systems (e.g. Ikeuchi, 1987; Lowe, 1987) far outstrip those of any existing neural network. As a computational formalism, neural networks have considerable potential for providing a natural expression of a wide range of neurobiological and psychological phenomena. Yet understanding how people recognize and spatially reason about objects in their environment is much too hard a problem to succumb to a simple change in formalism. We must develop a comprehensive computational theory that describes the appropriate computations for accomplishing the purposes of high-level vision, based on the constraints inherent in the problem and the nature of the information that is available to a behaving observer. Without such a theory, the problems that our models solve are unlikely to be the ones that we ourselves do.

References


**Figures**

**Figure 1:** Square/diamond ambiguity (after Rock (1973)).

**Figure 2:** ``Fruitface'' (after Palmer (1975a)).