RELATIONS BETWEEN SCHEMATA-BASED COMPUTATIONAL VISION AND ASPECTS OF VISUAL ATTENTION

by
Roger A. Browse
Department of Computer Science
University of British Columbia
Vancouver B.C. Canada

I. INTRODUCTION

This paper explores relations between aspects of visual attention and the operations of schemata-based computational vision systems. These relations are shown to suggest the requirement for methods which operate towards interpretation without model invocation. A specific mechanism is described which permits interpretation based interaction between information from different resolution levels, but does not rely on model invocation. This mechanism is then used in the examination of some related perceptual phenomena, permitting a more computational view of their operations.

II. SCHEMATA-BASED VISION SYSTEMS

An issue of interest to both cognitive psychology and artificial intelligence is the question of how knowledge of a domain of objects can be applied towards visual interpretation. Schemata-based knowledge organizations (Rumelhart and Ortony, 1976; Neisser, 1976) are now being used to address this issue (Freuder, 1976; Havens, 1978; Havens and Mackworth, 1980; Browse, 1980). One distinctive feature of a schemata-based interpretation is the organization of its domain knowledge along "natural" lines. The knowledge is object centered and relies on familiar structuring mechanisms such as component and instance hierarchies.

A domain of knowledge structured in this way is conducive to a recursive cuing mechanism (Havens, 1978): basic image elements act as cues for simple scene objects, which in turn act as cues for more complex objects, etc.

> For example; in the domain of line drawings of human-like body forms (Browse, 1980; 1981), a certain configuration of lines may cue a "hand", which in turn cues "arm", which cues "body".

At each level of this hierarchy, objects are described as being composed of simpler objects.

The occurrence of the objects which are required in the description may not be enough, however, to confirm the existence of the more complex object. There are also relations which must be valid among the components. This distinction will be referred to as the distinction between having found the required elements and having met the required relations.

> For example, all the required elements may exist to make up an "arm": the "hand", the "upper-arm", and the "lower-arm", but a number of required relations must also hold. The elements must be connected in a certain way, and the angles between the elements must be within certain bounds.

While it is difficult to be certain of the presence of an object on the basis of the required elements only, we shall see that there are special situations in which this information is very valuable. These situations rely on a capability of grouping image elements.

During the interpretation process, any element X in the image will have associated with it a set of model possibilities (or labels). This set is simply the set of all objects which are described using X as a required element. In the absence of a means of grouping elements, the interpretation process may deal with the discovery of an element by taking the course of model invocation (or testing). This operation involves selecting one or more of the model possibilities, and testing for their existence by locating the other required elements, and determining the validity of their required relations.

The model invocation approach can provide a dynamic determination of whether the processing procedes top-down or bottom-up (see Havens, 1978). As well it can provide a means of iterative refinement of interpretation and segmentation (see Mackworth, 1978). The operation of model invocation can, however, be costly because it is exhaustive search over the model possibilities.

On some occasions it may be feasible to delete some of the model possibilities without actually invoking them. This is possible whenever uniform constraining relations can be devised over a type of image element.

> For example, if we know that certain lines must be a part of the same object, then the model possibility sets for those lines can be intersected.

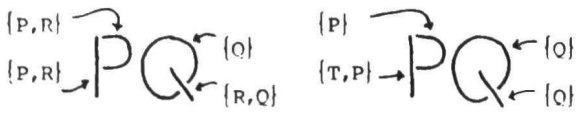
Waltz (1972) has shown that such a uniform constraint may be formulated for the interpretation of line drawings of the blocks-world. The result was that subsequent backtrack search was seldom required. Mackworth (1977) has provided a generalization of the use of such network consistency methods in artificial intelligence problems.

III. FEATURE INTEGRATION AND MODEL INVOCATION

Recently in cognitive psychology there have emerged theoretical ideas about visual attention which relate to the notions of model invocation and operations on sets of model possibilities. The Feature Integration Theory of attention (Treisman and Gelade, 1980) proposes that individual image features are detected rapidly and in parallel, but, in order that an object be identified as consisting of two or more separate features, locations must be processed serially with focal attention. If this is prevented, illusory conjuctions may be formed (Treisman and Schmidt, 1981). Thus, in human vision, the application of focal attention is required by model invocation.

There is an increased expense which accompanies the application of focal attention. Treisman, Sykes, and Gelade (1977) have shown that the amount of time required to detect objects made up of a conjunction of features increases linearly with the display size, but that display size does not have such a great effect on the detection of objects which can be defined without consideration of the relations among features. In computational terms, these objects can be identified only by the examination of model possibility sets to determine the presence of required elements, whereas conjunction objects require the establishment of the relation of common spatial location between features.

To make the relation more clear, consider an example taken from experiment IV of Treisman and Gelade (1980):



(a) target "R" (b) target "T"

Figure 1: two search conditions depicted as features which compose the letters, with their sets of model possibilities attached.

The task was to detect the presence of one of the targets "R" or "I" in a visual field of "P"s and "Q"s. Feature integration theory predicts that the search time will increase linearly with display size for the "R" target, and will increase less for the "T" target. This prediction was found to be correct. The difference between the two conditions is shown in figure 1. For the target "R", the required elements (features) are available in two different ways: either by the presence of an "R" or by adjacent "P" and "Q". In this target condition, the relations among the required elements must be examined. Computationally, this means invoking the model for an "R" each time its required elements are present. The target "T" can be detected by only examining the model possibilities for the primitives because the required features can only be present if the target is present.

There are indications that it is computationally more expensive to invoke models than to use consistency methods (Waltz, 1972; Mackworth, 1977). The proposed relation between model invocation and feature integration adds to the justification for the search for computational methods which can operate towards interpretation at the pre-invocation level.

IV. FILTERING ACROSS RESOLUTION LEVELS

Following the clues provided in the previous sections, Browse (1981) has devised a method which permits the interaction between information obtained at different levels of resolution. This method operates towards interpretation, but before model invocation.

A schemata-based representation for the knowledge of the body-form structure has been developed. This knowledge is specified in terms of image primitives attainable at two different levels of resolution in the image (lines and blobs). Areas of the image for which the correspondence of image primitives across levels is known are areas in which the following uniform constraining relations may be applied:

1. For any blob which must have an integral interpretation, the corresponding lines must all have a common interpretation model. Thus the sets of model possibilities for each of the lines may be reduced to their intersection.
2. The ultimate interpretation must be the same for both levels of resolution (at least instance-hierarchy related), hence the possibility sets may be intersected across levels.

Figure 2a depicts two lines which are known to correspond to a specific blob because of their image hierarchy structure. Also depicted is a model possibility set for each element. By applying rule

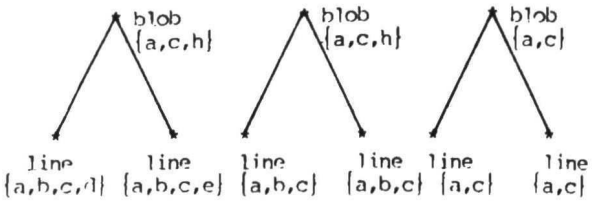


Figure 2: Three stages in the application of consistency across resolution levels.

(1) we eliminate {d,e}, and by applying rule (2) we arrive at the final set of possible models {a,c} as shown in figure 2c. See Browse (1981) for an example of the operation of these methods in the body-form domain.

The correspondences being utilized by these methods are only available in the limited area of the image which has been processed at the highest level of resolution (fovea), and the ultimate usefulness of such operations will be influenced by the appropriateness of the selection of these locations by the program (Browse, 1980). It also remains to establish computational advantages in the order of processing the local and global information (see Navon, 1977; Kinchla and Wolfe, 1979).

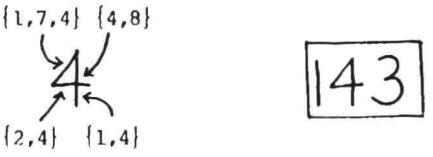
V. GROUP PROCESSING AND MODEL INVOCATION

Kahneman and Henik (1977) have formulated a "group-processing" model of the application of attention which is similar to the application of constraining relation (1) of the previous section. Their model proposes a pre-attentive grouping operation which selects large scale objects for subsequent analysis. The experiments which demonstrate the validity of this model employ displays such as that of figure 3.

(a) (b)

Figure 3. Group processing digit detection display

One of the two displays such as shown in figure 3, is presented briefly and the task is to detect a specified target digit. The results show that groups are processed separately, but that processing is almost uniform within groups.



(a) (b)

Figure 4. Features available at two resolutions.

Assume that high resolution feature information is availa[illegible], and that for each such feature, a set of model possibilities is established (as shown in figure 4a). Also assume the availabilty of coarse level

information which gives the identification of larger objects (figure 4b).

Consider the following interpretation of these results: In the first stage, the global objects are detected, as are the high resolution features specifying their model possibility sets. These sets can only be assigned, however, to the established objects, as depicted in figure 5.

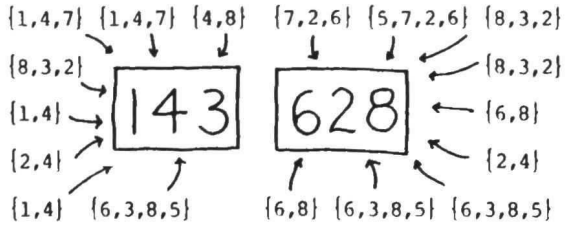


Figure 5. low resolution objects detected and model possibilities assigned to high resolution features, which are roughly located.

At this point, there are obviously too many features associated with the object for it to be a single digit, so a subsequent breakdown of objects takes place. In that this second phase is a higher resolution, it can only take place over a smaller area, so one of the two main objects is selected for more detailed examination (see figure 6).

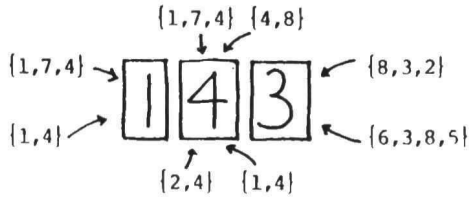


Figure 6. Features assigned to objects detected at a finer level of resolution, for one of the low level objects.

A second examination of the possibility sets reveals that the required elements are available for only one digit in each of the defined positions, and hence their identities can be established in parallel, without serial application of attention to each of the specific locations.

Feature integration theory proposes that object identification may take place in parallel, based on features alone, or serially based on conjunctions of features when necessary. The group processing results indicate intermediate steps at which features are assigned to objects detected at low resolution, and once this assignment is complete, some model possibilities may be discarded by using the constraining relation (1) from the previous section. The location of objects to which these features are attached will become more refined if necessary, to the point of either allowing object identification through confirmation of the presence of the required elements alone, or if necessary by considering the relations among features.

The identity of features may be determined over a wide visual field, but without specific location. Location may become more specific through attachment to low resolution image elements, but only over a more restricted visual field. Finally, the actual location may be determined to permit feature integration. This final locating action operates over a small area of the visual field, and therefore requires serial application if more than one location is to be searched.

VI. SUMMARY

By adopting a schemata-based approach, computational vision domains may be structured so as to use the component hierarchy as a mechanism for cuing the models to be invoked. The complete examination of the relations required by models can be computationally expensive, and for human vision, presupposes the application of foveal attention.

A mechanism has been described which permits the interaction between information from different levels of resolution by eliminating model possibilities and imposing groupings over high resolution features. This mechanism has been shown to be useful in describing the relation between group processing phenomena and Feature Integration Theory.

VII. REFERENCES

Browse, R.A.
1980. Mediation between central and peripheral processes: useful knowledge structures. PROC CSCSI-3, Victoria, B.C. pp.166-171.

Browse, R.A.
1981. Interpretation-based interaction between levels of resolution. submitted to IJCAI-6, Vancouver B.C., Aug 1981.

Freuder, E.C.
1976. A computer system for visual recognition using active knowledge, Ph.D. thesis, AI-TR-345, M.I.T., Cambridge, Mass., 1976.

Havens, W.S.
1978. A procedural model for machine perception. Ph.D. thesis, Thechnical report TR-78-3, Department of Computer Science, University of British Columbia.

Havens, W.S. and Mackworth, A.K.
1980. Schemata-based understanding of hand-drawn sketch maps. PROC CSCSI-3, Victoria, B.C. pp.172-178.

Kahneman, D. and Henik, A.
1977. Effects of visual grouping on immediate recall and selective attention. in S. Dornic (ed.) Attention and Performance VI Hillsdale, N.J. Lawrence Erlbaum. pp.307-332.

Kinchla, R.A. and Wolfe, J.M.
1979. The order of visual processing: "top-down", "bottom-up", or "middle-out". Perception and Psychophysics, 1979,25, pp. 225-231.

Mackworth, A.K.
1977. Consistency in networks of relations. Artificial Intelligence 8, pp.99-118.

Mackworth, A.K.
1978. Vision research strategy: black magic, metaphors, mechanisms, mini-worlds, and maps. in A.R Hanson and E.M. Riseman (eds.),Computer Vision Systems, Academic Press, 1978. pp.53-59.

Navon, D.
1977. Forest before trees: the precedence of global features in visual perception. Cognitive Psychology, 1977,,9. pp. 353-383.

Rumelhart, D.E. and Ortony, A.
1976. The representation of knowledge in memory. Center for Huamn Information Processing report 55, University of California, San Diego. Jan 1976.

Treisman, A., Seykes, M., and Gelade, G.
1977. Selective attention and stimulus integration. in S. Dornic (ed.) Attention and Performance VI Hillsdale, N.J. Lawrence Erlbaum. pp.333-361.

Treisman, A. and Gelade, G.
1980. A feature integration theory of attention. Cognitive Psychology 12 pp.97-136.

Treisman, A. and Schmidt, H.
1981. Illusory conjunctions in the perception of objects. to appear.

Waltz, D.L.
1972. Generating semantic descriptions from drawings of scenes with shadows. Technical report MAC AI-TR-271, M.I.T., Cambridge, Mass.