コンピュータビジョン 69－3
（1990．11．22）

物体の認識と学習：反マー的アプローチの提案

徐 剛、 辻 三郎
大阪大学 基礎工学部 制御工学科

マーの視覚計算理論では、物体認識は物体中心座標系の３次元モデル表現に基づいて行なわれる。しかし、人間は画像座標系の２次元パターン表現に基づいて物体認識を行なっている証拠がある。さらに、２次元パターン表現の方が物体の学習ができる。本論文では、まずマーの視覚計算理論の是非について述べたあと、物体の認識と学習に関する２次元パターン表現の優位性を論じる。最後に２次元パターン表現に基づいた学習する動的ビジョンシステム(SLAVS, Self-Learning Active Vision System)の構想について紹介する。

Object Recognition and Learning: A Proposed Anti-Marrist Approach

Gang Xu and Saburo Tsuji
Department of Control Engineering
Osaka University, Toyonaka, Osaka 560, Japan
xu@ce.osaka-u.ac.jp

ABSTRACT

In Marr's vision theory, object recognition begins where the object-centered 3D model representation completes. On the contrary to his claim, there is evidence to say that in human vision object recognition is not done between object-centered 3D models, but between 2D retinotopic patterns of the image and the memory. More importantly, based on this representation, objects can be learned. This paper first comments on Marr's computational approach to vision, and then discusses the superiority of 2D viewer-centered representation with respect to object learning and recognition. Finally we present the blueprint of a Self-Learning Active Vision System (SLAVS), which includes segmentation and gaze control, recognition, learning, and spatial reasoning and feedback modules. The main feature of this system is that each time a 2D pattern is successfully matched, it is learned and organized into the corresponding object model. Thus the system has its history and arrow of time.

1. INTRODUCTION

As defined by David Marr, and many others as well, vision is a process of knowing what is where by looking [Marr, 1982]. It seems legitimate to say that the "where" question is relatively simpler. We have many sensing modules to compute the distance, the depth and the surface orientation. However, to answer the "what" question, known as the problem of object recognition, has proven to be very difficult. More than 20 years of research on computer vision has not but revealed the hardness of the problem. The main reason for the difficulty attributed by many people was the lack of 3D information lost in the process of projection. In this situation, Marr's computational vision paradigm received overwhelming welcome in our computer vision community. In the past ten or fifteen years, we have built a lot of theories and machines to compute 3D information from one or two or a sequence of images. We have shape from stereo, shape from motion, shape from texture, shape from shading, shape from contour, shape from focus, so on and so on [Barrow & Tenenbaum, 1981; Grimson, 1981; Horn & Schunk, 1981; Ikeuchi & Horn, 1981; Krotkov, 1987; Ullman, 1979; Witkin, 1981; Xu & Tsuji, 1987]. Because there are so many, we call it totally as shape from X. The unstated hope behind all these efforts is that if we can recover all lost 3D information, then the hard things will become easier.

Another direction of efforts is to obtain object-centered representations of objects. Viewer-centered 3D information as computed by the shape-from-X modules is dependent on viewing direction and thus is claimed to be unsuitable for representing 3D objects. Marr calls the 3D information registered in the image frame as 2.5D Sketch and argues that we need an object-centered 3D Model representation, based on which recognition can be done regardless of viewing direction [Marr, 1982]. Marr and Nishihara has proposed the Generalized Cylinders [Marr & Nishihara, 1978], and the others have proposed the Perimetric Models [Pentland, 1987].

Recently, a new vision paradigm, Active Vision paradigm [Aloimonos & Badyopadhyay, 1987; Bajcsy, 1988; Ballard, 1989] emerged, and research within this framework has since been very active. Though different people emphasize different aspects, I summarize its advantages as: (1) more 3D information or more accurate 3D information; (2) changing ill-posed problems (of shape-from-X) to well-posed problems; (3) direct acquisition of object-centered representation; and (4) exploratory intelligent control. The first three are extensions along the previous research directions, but the last one is an important aspect of vision, and perception at large, which we had previously neglected.

All of these challenges to the problem have contributed to our knowledge of this most knowledgeable sense of humans. Indeed, one function of our vision is the recovery of 3D information from 2D projections. However, the "what"

problem is still there. Suppose that we are given very accurate range data from a laser ranging system, can we do recognition? We consider that to solve the "what" problem, one must first answer the question: what is "what"? That is, one must first determine what representation one uses to describe 3D objects. Secondly, one must know how to build these descriptions for each object. We believe that the best way to do that is learning. Therefore the representation for 3D objects must also be proper for learning object models. By bringing learning into consideration, we not only solve the problem of knowledge acquisition, but it also provides a strong constraint on what representation one chooses for describing 3D objects.

2. 3D OBJECT-CENTERED REPRESENTATION vs. 2D VIEWER-CENTERED REPRESENTATION

There are generally two classes of representation for describing 3D objects: object-centered 3D models, and a set of viewer-centered 2D retinotopic patterns. Most of the current object recognition systems are based on the object-centered model representation. Variations of this class of representation are Generalized Cylinders and Superquadric Models. Objects are decribed as parts and their spatial relationships in the form of nodes and arcs. Recognition takes place between a description built from the image in a bottom-up manner and the model, or between the image and an appearance prediction of the model. Although these systems work, the problem is that all objects cannot be described by this class of representation [Brooks, 1981; Pentland, 1987]. One cannot imagine the generalized cylinder representation can be used for face recognition.

The other class of representation is a set of retinotopic 2D patterns. The reasoning is that if we know all its possible 2D appearances, then recognition can be done by matching the image against all the 2D appearances stored in the memory. Compared with the object-centered model representation, this class is more general. The difficulty with it is that the number of possible appearances is very large. One effort is to describe the objects by only representative views, for example, "Aspects", defined as topologically equivalent classes of appearances with which ranges of view points are associated [Koenderink & van Doorn, 1979; Rosenfeld, 1987; Weschler, 1990]. Recognition is then reduced to 2D matching between topological equivalents in the image and memory. Application of this representation to artificial machine parts is successful [Ikeuchi & Kanade, 1989]. However, again the problem with the aspect representation is that it is hard to define aspects for natural objects. Noise and occlusion make "aspects" of a complex object subtle and unstable.

There seems to be enough psycological evidence to say that in human vision the matching is 2-dimensional. To cite one example, recall the famous experiment by Shepard and Metzler (1971), which discovered that the time

taken to identify two unfamiliar objects that differ from each other by a rotation is proportional to the physical angle of rotation. One can thus conclude that a "mental rotation" is actually being performed to bring the first image into correspondence with the second, requiring longer time if the angle is larger. What is implicit in this conclusion is that we are doing 2D matching, not 3D matching. The reason is straightforward: if the objects were described in a 3D model as a series of oriented blocks, then the time taken to identify the two codes would be invariant with the angle of rotation. This strongly suggests that objects be described in the 2D viewer-centered representation, not in the 3D object-centered representation. However, this does not mean that 3D information is useless. Without it mental rotation is impossible. 3D information does not necessarily mean object-centered. Viewer-centered 3D information, called 2.5D sketch, is also a representation based on which mental rotation can be directly performed. On the other hand, if we are familiar with the object, i.e., if we have learned the object's appearances, then recognition can be done without mental rotation. This is the reason why in Shepard and Metzler's experiment they must use the objects with which the subjects are unfamiliar in order to investigate mental rotation.

With respect to learning, it is evident that the 2D viewer-centered representation is more convenient than the 3D object-centered representation. Actually, the canonical object-centered representation approach advocated by Marr and Nishihara (1978) tries to compute everything in one step. There is no room for learning. On the other hand, 2D viewer-centered representation approach tries to do things little by little, and need virtually no pre-computation to build the model. To take it in an extremely naive way, what one needs to do is just to record every pattern he/she sees, and compare every new view with the stored old patterns, if there is enough memory to tolerate such an uneconomic usage. The problem, then, is how our memory can "organize" the patterns, to reduce the abundant information involved.

We have so far used the term "pattern" for the 2D viewer-centered representation. What we mean here is only shape pattern, not including patterns of color, texture, or motion, etc., although incorporating all of them would definitely improve the overall performance. We believe, however, that the shape information is the most fundamental. To make it more specific, we consider that an edge image is equivalent to a shape pattern, because major shape information is all included in the edge image. Marr called it "Primal Sketch". However, we will call it "2D Sketch" in correspondence with 2.5D Sketch and 2.1D Sketch (see 3. 1, [Nitzberg and Mumford, 1990]), which all emphasize the viewer-centered 2-dimensionality.

3. BLUEPRINT OF THE SELF-LEARNING ACTIVE VISION SYSTEM

We have drawn a blueprint for a self-learning active vision system (SLAVS) based on the 2D Sketch representation, which is shown as a block diagram in Fig. 1.

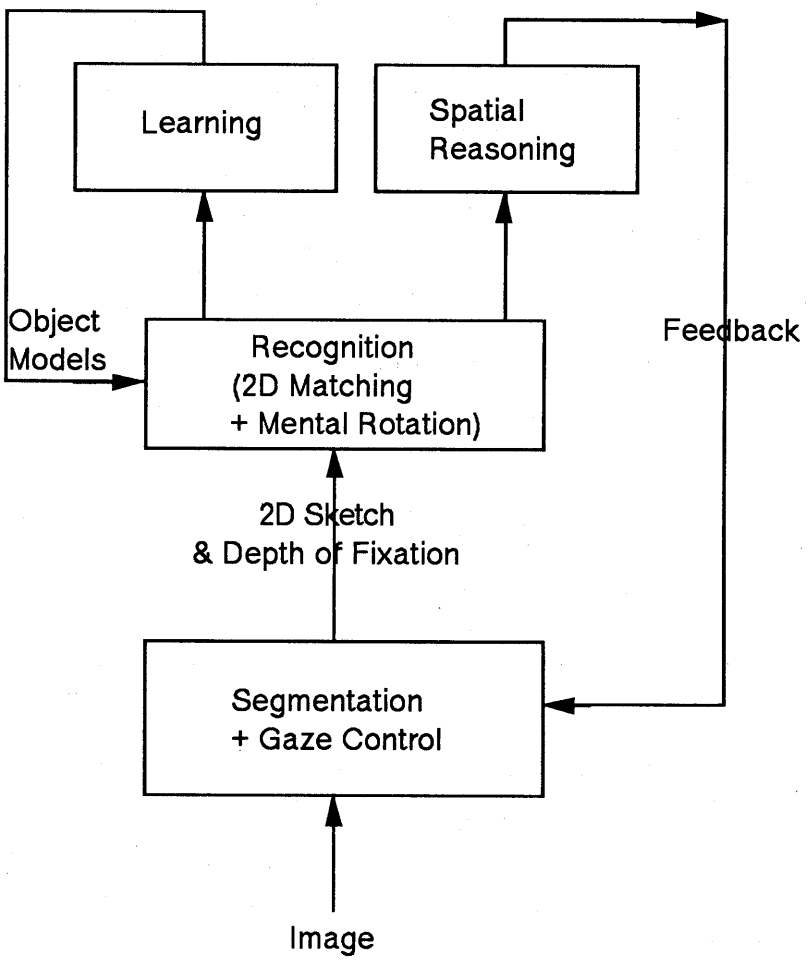

Fig. 1 Block Diagram of our Self-Learning Active Vision System (SLAVS)

There are 4 subsystems in SLAVS: segmentation and gaze control, recognition, spatial reasoning and learning. The system has two characteristics. The first is that each time an object is recognized, the object model is updated. Thus the whole system has its history and evolves as it experiences. The more the model learns, the less the recognition makes effort. The second is that the process is a cyclic loop of bottom-up and top-down. There is virtually no boundary between perception and cognition. There are three basic operations that the system can perform: matching an image part against models; searching for an object's instances in the visual field; judging whether two image parts are the same object. Complex tasks can be programmed as

combinations of these three basic operations. The system receives commands from a human operator.

3.1 Segmentation and Gaze Control

For similicity of explanation, we first introduce the lowest level of processing, the segmentation and gaze control subsystem. This stage first filters the image to a 2D Sketch, the edge image, which contains major shape information. The 2.1D Sketch is an image segmented into regions of different depths [Nitzberg & Mumford, 1990], which likely correspond to different objects. Gaze control chooses one from among the regions based on the saliency measures or feedback control from the spatial reasoning subsystem. It is universally accepted that perfect edge detection and segmentation of image into objects with only low-level constraints is generally impossible. In this system we try to solve the problem together with recognition. If the chosen region corresponds, entirely or partially, to an object, then the segmentation is affirmed. If not, then gaze control chooses another segmentation for recognition.

3.2 Recognition

The recognition subsystem is a 2D matcher plus a "mental rotater". Matching should be of a certain degree of flexibility and fault-tolerance, and should also be able to report partial correspondence resulting from occlusion. We consider that the connectionist, or neural approach is the most promising one to meet these standards. As claimed by many connectionists, the connectionist model is effective at avoiding the search problems that accompany serial computational architectures, and can assess the goodness of correspondence between the image part and all the patterns stored in the memory, finding the best match at the same time.

Mental rotation takes place when there is no direct matching between the current image pattern and the patterns stored in the model. The mental rotator synthesizes patterns virtually viewed from new angles based on the local 3D shape information of the current image part, and tries to match them against the model. But the system cannot always activate mental rotation, because this takes much greater time, and is not guaranteed to succeed. It should be activated only when the system is commanded to check whether or not an image part is the same as another image part or a particular object model.

To do 2D matching, one needs to know two more things: scale and position. We consider that the scale can be determined from the depth of fixation if an active camera system with the automatic focusing and gaze control mechanism is used. The scale is inversely proportional to the depth of fixation, with the pattern unchanging by assuming othorgraphic projection. If

the visual input is a static picture, then what is available is relative scale. If a building is samller than a man, then it is a toy, not a real building, no matter how real it looks. Position can be determined as the centroid of the region. But if there is occlusion, then search is likely required to bring the two patterns superimposed.

One important thing not to forget is that there are not only many viewing angles relative to the object, but still for each viewing angle there is also a freedom of rotation. It is known in psycology that humans cannot recognize a pattern if it is rotated by a certain angle from its usual appearances. This implies that only when the relative "posture" between the eyes and the objects is constant, can we recognize the objects. Fortunately, both ourselves and the objects have restricted relations with the gravitational direction. Thus the possible ways that objects appear are also limited. What we need to do is to keep the camera's vertical axis to be always within the vertical planes respect to the gravitational frame.

3.3 Learning

The learning ability is the main progress of this system. One worry about the 2D representation of 3D objects is that the number of views can be explosive if one tries to model all objects in this way. We think that the views should be "organized" rather than only being stacked. The connectionist model shows convincingly the capability to deal with this problem. Many patterns can be stored in one unit. The matrix of synapses provides a space in which information can be compactly represented. What interests us more is that it has the ability of "interpolation", relieving us from the burden of learning the 2D sketches viewed from every direction. To cite one example from [Kohenon, Oja & Lehtio, 1989], their model can identify faces viewed from angles different from those used before for learning.

3.4 Spatial Reasoning and Feedback

SLAVS is an active system. The segmentation and gaze control subsystem receives control commands from the spatial reasoning subsystem. Its main task is to decide whether or not to continue the visual search, and if yes, where to focus next image part. The memory should not only include 2D sketches associated with the objects, but also spatial relationships among them, which provide the basis for the Spatial Reasoning subsystem to produce feedback commands. If the visual task of recognizing some image part or of locating the object in the image is finished, then the system should stop and wait for another new command. If to finish the task needs more recognition, then the problem is which image part should be checked next. It is rare that the whole object can be recognized in one glance. Usually a part of the object is first recognized, which then generates a prediction of what should appear

where. This prediction is fed back to the low-level processing subsystem, starting a new cycle.

4. SOME UNKNOWNS

What we have described above is the current blueprint of the SLAVS system. Many of the issues involved have to be further detailed and revised in the process of implementation. Still, there are some issues we are not sure of at this stage.

(1) Size constancy Human vision exhibits surprisingly perfect size constancy. We have indicated the necessity of compensating for the scale change by the depth of fixation. But at which scale or depth of fixation should we learn the patterns?

(2) 0th level analysis The computational complexity issues of a vision system should be seriously examined if it is to display human-like, real-time performance [Tsotsos, 1987]. We have not yet checked whether a visual task can be done within the 100 step limit in our proposed system.

(3) Learning spatial relations Spatial relationships among the objects in the memory should also be learned, but we are yet to single out the way. Again, the problem is representation, 2D or 3D, viewer-centered or world-centered?

5. CONCLUDING REMARKS

In this paper we have made criticism and comments on Marr's computational vision paradigm. It was then argued that representation of objects is crucial to recognition and learning, and that sets of 2D viewer-centered sketches are a better representation than object-centered 3D models. We then proposed a self-learning active vision system (SLAVS), which incorporates the segmentation and gaze control, recognition, learning and spatial reasoning modules. We have also indicated the unknowns and ambiguities in the proposal for future investigation and improvement. Implementation of this system is to be started soon.

ACKNOWLEDGEMENTS

The authors would like to thank Dana Ballard, Ruzena Bajcsy, Daniel Lee, Hiromi Tanaka, Masahiko Yachida, Seiji Yamada and Matthew Barth for their suggestions and help.

REFERENCES

Aloimonos, J. and Badyopadhyay, A. (1987) "Active vision", *Proc. 1st IEEE Int. Conf. on Computer Vision*, pp. 35-54

Bajcsy, R. (1988) "Active perception", *Proc. of the IEEE*, Vol.76, No.8, pp.996-1105
Ballard, D. (1989) "Reference frames for animate vision", *Proc. of 11th Int. Joint Conf. on Artif. Intel.*, pp.1635-1641
Barrow, H.G. and Tenenbaum, J.M. (1981) "Interpreting line drawings as three-dimensional surfaces", *Artificial Intelligence*, Vol.17, pp.75-116
Brooks, R. (1981) Symbolic reasoning among 3D models and 2D images", *Artificial Intelligence*, Vol.17, pp.285-348
Grimson, W.E.L. (1981) *From Images to Surfaces: A Computational Study of the HUman Early Visual System*, MIT Press
Horn, B.K.P. and Schunk, B.G. (1981) "Determining optical flow", *Artificial Intelligence*, Vol.17, pp.185-203
Ikeuchi, K. and Horn, B.K.P. (1981) "Numerical shape from shading and occluding boundaries", *Artificial Intelligence*, Vol.17, pp.141-184
Ikeuchi, K. and Kanade, T. (1989) Modeling sensors: Toward automatic generation of object recognition program, *Computer Vision, Graphics and Image Processing*, Vol.48, No.1, pp.50-79
Koenderink, J. J. and van Doorn, A. J. (1979) "The internal representation of solid shape with respect to vision", *Biological Cybernectics*, 32, pp.211-216
Kohonen, T., Oja, E. and Lehtio, P. (1989) "Storage and processing of informationin distributed associative memory systems", *Parallel Models of Associative Memory*, ed by Hinton, G. and Anderson, J., Lawrence Erlbaum Associates, pp. 125-167
Krotkov, E. (1987) "Focusing", *Int. Journal of Computer Vision*, Vol.1, pp.223-237
Marr, D. and Nishihara, H. (1978) "Representations and recognition of the spatial organization of three-dimensional shapes", *Proc. R. Soc. London*, Vol. 200, pp.269-294
Marr, D. (1982) *Vision*, Freeman
Nitzberg, M. and Mumford, D. (1990) "The 2.1D sketch", *Proc. of 3rd IEEE Int. Conf. on Computer Vision*, to appear.
Pentland, A. P. (1987) "Recognition by parts", *Proc. 1st IEEE Int. Conf. on Computer Vision*, pp. 612-620
Rosenfeld, A. (1987) Recognizing unexpected objects: A proposed approach, *Proc. of DARPA Workshop on Image Understanding*, pp.620-627
Shepard, R.N. and Metzler, J. (1971) "Mental rotation of three-dimensional objects", *Science*, Vol.171, pp.701-703
Tsotsos, J. (1987) "A 'complexity level' analysis of vision", *Proc. 1st IEEE Int. Conf. on Computer Vision*, pp. 346-355
Ullman, S. (1979) *The Interpretation of Visual Motion*, MIT Press
Wechsler, H. (1990) *Computational Vision*, Academic Press
Witkin, A. P. (1981) "Recovering surface shape and orientaion from texture", *Artificial Intelligence*, Vol.17, pp.17-45
Xu, G. and Tsuji, S. (1987) "Inferring surfaces from boundaries", *Proc. 1st IEEE Int. Conf. on Computer Vision*, pp. 716-720