CVGIP: IMAGE UNDERSTANDING
Vol. 53, No. 1, Jan, pp. 112–117, 1991

DIALOGUE

Ignorance, Myopia, and Naiveté in Computer Vision Systems

RAMESH C. JAIN

Artificial Intelligence Laboratory, Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan 48109

AND

THOMAS O. BINFORD

Robotics Laboratory, Stanford University, Palo Alto, California 94395

After a period of tremendous excitement and enthusiasm, many industrial people and researchers are disenchanted with computer vision, and others are certainly much less enthusiastic about it. The time has come to regroup. To restore the upward trend of our field, critical introspection followed by serious corrective action is required. Active researchers in computer vision can make it a balanced science that can be applied in many disparate areas by following research approaches used in most of the successful applied scientific fields. Our aim in this paper is to provoke discussion and actions that may lead to corrections in our favorite research field. © 1991 Academic Press, Inc.

1. INTRODUCTION

A computer vision system recovers useful information from one or more images of a scene. By *scene* we mean the physical environment that is of interest. An *image* is the output of a sensor used to *see* the environment. It is important to realize that in most applications an image is a lower-dimensional projection of a scene. The information desired is application-dependent and must be recovered from images.

Clearly, computer vision methods have enormous potential applications. Defense, industry, medicine, and several other fields saw this potential and tried to harness it. Computer vision research and companies had a boom period in the first half of the 1980s. Unfortunately, now things are not that rosy. Many industrial people and academic researchers are disenchanted with computer vision, and others are certainly much less enthusiastic. In other application areas also, the progress, if any, has been extremely slow. To change the current downward trend of our field, critical analysis is required. We believe that computer vision can be turned around: from a field in shambles to a respectable and balanced science.

In the early days of computer vision research, it was believed that the major bottleneck in solving problems in vision was the computing power and image acquisition facilities. This was particularly true for industrial applications. Binary vision was used in industrial applications partly because of the computational cost and partly because the methodology was understood. There was not a clear path to go beyond binary vision and connected components. The last decade has seen significant growth in the computing power available to vision researchers. In 1980, there were only a few places that could afford to have a lab with good image acquisition capability. Now anyone can buy image acquisition hardware and a computer to build a powerful vision workstation.

How has the availability of the equipment affected the research culture? It would appear that the availability of much needed laboratory facilities would encourage facility-starved researchers to over-experiment and develop a rich experimental sub-field. However, it is intriguing that as the availability of laboratory facilities improved, the interest of computer vision researchers in experimenting with their techniques decreased. Rather, the common practice of researchers presenting and peers accepting unsubstantiated claims has continued. This is tolerable in the infancy of a field; mature scientific disciplines are expected to develop experimental methodologies, comparative evaluation techniques, and theory that is based on realistic assumptions. This has not yet happened in computer vision. We still accept subjective quality of the output as judged by the author of a paper. We have not yet developed objective evaluation methods. In fact, a good question to ask ourselves is: Why have vision researchers become more infatuated with techniques that are not tested in laboratories, as the availability of laboratory facilities improved? We, like most other researchers, are aware of the importance of having both theory

1049-9660/91 $3.00

and experimentation as the two equally important components of a science and engineering field. The importance of theory cannot be overemphasized. But at the same time, a discipline without experimentation is not scientific. Without adequate experimental methods, there is no way to rigorously substantiate new ideas and to evaluate different approaches. Moreover, it was expected that computer vision techniques would be applied to solve problems in many fields. How can one develop techniques that will be applied in disparate fields without developing experimental aspects of the field?

This paper presents some thoughts on what are the basic problems in computer vision and how the research in computer vision has, like that in other fields in their infancy, suffered from the *drunk man under the lamppost* and the *Emperor's new clothes* syndromes. There is no implication that the authors are not guilty of many things mentioned here. The idea is critical introspection of our field, not to blame a particular person or group.

Rather than analyzing the problems to be solved, and then developing appropriate tools, we usually apply our favorite tools that we learned in graduate school (possibly studying an entirely different discipline). These tools include mathematical approaches, data structures, and other techniques. In general, we tried to apply a tool that we learned in some other discipline to computer vision, without ever really analyzing whether it suited our needs. Many times, we took a fancy to a tool because a friend told us (or we read about it in a popular magazine) how good it is in some application. We got excited about the tool and decided to apply it to our current vision problem or the next problem that we encountered. Our approach has been mostly tool-driven. Thus, like the drunk man, we are looking for solutions to our problems using well-developed tools. It would be nice if these tools were relevant to our problems and applications. To find the applicability of tools, and to calibrate them, we will have to experiment with carefully developed experimental methods. If we do not do that, then we will remain under the comfortable well-illuminated lamppost, looking for keys that we lost miles away in the dark. As is clear to every person, to find keys, we will have to illuminate the area where we lost them; even searching in the dark will be more productive than searching under the lamppost.

Every new field has its fads and fashions. Thomas Kuhn suggests that science progresses through a succession of paradigms. At a given time, a paradigm is popular and at that time the situation is like the Emperor's new clothes. Most researchers believe in the paradigm and do not dare to challenge it. Even when people feel that there is something wrong with what everybody is seeing, they think that maybe what they see is not right and hence they say that they are seeing the same thing that everybody else is supposed to see according to well-known experts. It requires an innocent child to have the naivete to see and say that the Emperor is naked (there is something wrong with what experts want everybody to believe). Computer vision is a very fertile field for this kind of thing to happen. We have researchers from computer science, psychophysics, neurophysiology, cognitive science, electrical engineering, and mathematics applying their tools to solve problems in computer vision. If a paper authored by a well-known researcher at a prestigious laboratory presents a solid mathematical treatment of a supposedly important problem in vision using very sophisticated approaches, possibly based on advanced or obscure mathematics, a reviewer trained in computer science, who never heard of those approaches, has either a tendency to believe in it or the common sense to keep quiet. Similarly, if a psychophysicist presents his new exciting theory about human vision, a computer vision researcher may get enough impressed to design his next system based on that theory, which may be out of fashion in psychophysics by the time data structures for the implementation of the system are designed. In most scientific disciplines, the real progress, as pointed out by Kuhn, takes place when an estimated paradigm is overthrown. We have been very slow in overthrowing paradigms that took us to a dead end.

The most exciting aspect of computer vision is that it draws from many disparate disciplines and can be applied to so many areas. The importance of vision, and perception, in our life suggests that computer vision systems have the potential to help us in many fields. This may be a curse, if we do not learn how to work in this exciting field.

2. IGNORANCE, MYOPIA, AND NAIVETE

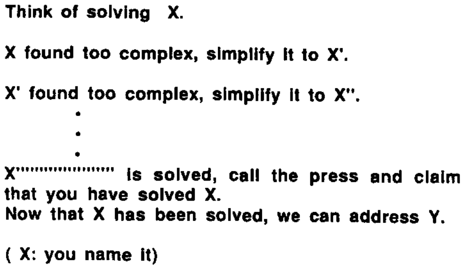
The three most serious limitations of computer vision systems are a grossly insufficient and inadequate use of knowledge, the inability to see far enough, both spatially and temporally, and the lack of experimental tradition.

A perception system has to rely on knowledge of many types. Unfortunately, computer vision research has been predominantly concerned with development of operators. Representation of knowledge, reasoning, uncertainty management in combining outputs of different operators, and creating the knowledge base to characterize the performance of operators have been mostly ignored. Even the systems that emphasized knowledge mostly used only expert systems like shallow reasoning. If we want vision systems to be intelligent and powerful, we will have to remove their ignorance.

Another major problem with vision systems has been the development of techniques that are myopic, both spatially and temporally. Many techniques for image analysis make assumptions that are true only locally; they fail to consider the fact that most images contain many surfaces and hence most local assumptions are violated when one goes from one surface to another. Such operators can only be applied if segmentation has been done before applying these operators. Also, because perception is dominated by reasoning rather than operators, the misplaced emphasis on local operators results in losing the perspective of the problem.

In dynamic vision, myopia is more explicit. Most research in this field has addressed the so-called structure from motion problem. Here the emphasis has been on recovering structure using a minimal number of points in a minimal number of frames. One would expect that in a dynamic environment, the emphasis should be on recovering robust information, exploiting the availability of a sequence that allows the luxury to wait until an appropriate time instant to recover the information. Interestingly, even most optical flow approaches rely on just a few frames. Some researchers advocate use of multiresolution operators for edge detection and do not hesitate to apply 51 × 51 operators, but try to recover all structure information in three or fewer frames.

The third problem, already discussed in the Introduction, is the absence of experimental computer vision. This lack of experimental aspect has resulted in researchers never thinking hard to define what problem they are really addressing. A common research methodology is shown in Fig. 1. As pointed out in the figure, by the time a problem is defined and solved, it has at best a tenuous relationship with a problem in computer vision. Most established disciplines solve this problem by rigorous standards for experiments reported in published literature. It is no surprise that each discipline has evolved appropriate evaluation methods for the performance of the techniques and processes. If there are no set evaluation methodologies, researchers have a tendency not to define their problem. This affects theory also. Researchers doing theoretical work never specify their assumptions explicitly and hence there is no need to justify whether the theory presented by them is really related to computer vision. It is common to see a paper that starts to address a problem in computer vision and justifies the proposed techniques based on whatever (little) the author knows about the psychophysics or neurophysiology. That justification reminds us of Martin A. Gardner's definition of pseudo-science.



Think of solving X.

X found too complex, simplify it to X'.

X' found too complex, simplify it to X''.

•
•
•

X'''''''''''''''''''' is solved, call the press and claim
that you have solved X.
Now that X has been solved, we can address Y.

(X: you name it)

FIG. 1. The research approach commonly used in many new fields, including computer vision.

3. SEGMENTATION

One of the first operations that a computer vision system must perform is the separation of objects from the background. This operation, commonly called segmentation, can be performed by using either the similarities or the dissimilarities of certain properties of points in an image. Clearly, models of possible objects can help in segmentation, but in most cases it is not known what objects are in the scene. Thus segmentation usually begins without use of model knowledge, but can be enhanced by using such knowledge.

3.1. *Importance of Segmentation*

After a very active early period in segmentation research, the problems in segmentation have usually been ignored. Most of the early approaches were based on simplistic models of intensity characteristics of surfaces of objects. Such approaches could only succeed in images that had simple objects without much intensity variation. Knowledge-based approaches tried to rely either mostly on very general knowledge of objects or only on superficial image knowledge. There have been very few efforts to use three-dimensional models of objects and knowledge of image formation in refining segmentation. Due to the limitations of segmentation approaches, not much success was accomplished. We believe, however, that researchers were addressing one of the most important problems in computer vision.

Much research in the last decade has addressed several problems, such as the shape from *X* techniques, that implicitly either assume that the segmentation problem has been solved and their techniques are applicable to each region in an image or assume that there is no segmentation problem. Both these assumptions are serious mistakes. In the first assumption, the problem of integrating different modules finally is going to be very difficult. The second assumption of no segmentation problem is fatal. This assumption essentially means that the whole image is one surface. Images that satisfy this assumption are rare and uninteresting. One can develop a plethora of techniques that are irrelevant for most computer vision tasks, if they rely on such an assumption.

Examples of techniques that have difficulties in the presence of discontinuities are optimization techniques, such as regularization, that use a performance measure which is the sum of a function at each point of the image. These mathematically elegant and rigorous techniques have very limited use with real images because they cannot handle discontinuities and rely on optimization criteria that are not very meaningful in most situations due to the restrictive assumptions made for the formulation of a mathematically tractable problem. Usually, these techniques are based on smoothness of some local property. We know that most useful information in images is near boundaries of regions, precisely the location where the smoothness assumption is not valid. No wonder these techniques have not been applied to real images, even of constrained scenes. The success of these techniques is demonstrated, if at all, using only pathological cases like objects sprinkled with random dots.

3.2. *Control Structure*

It is clear that information about possible objects can help segmentation. Most efforts in this direction relied on weak general purpose operators in the early stages. To make matters worse, the behavior of those operators was not understood or known. It is difficult to design a high-level system that relies on little understood, poorly performing low-level operators. Many researchers are emphasizing either multiresolution operators or computation of multiple intrinsic property images from an image and then combination of these outputs to recover desirable information. The idea of creating multiple images in which each image emphasizes a property, maybe at a particular scale, will be effective if we know how to combine those images. Currently the criterion used in generating such images seems to be more useful for human visualization than for automatic recovery of information.

It is important not only to design more powerful early vision tools for signal-to-symbol transformation, but to understand their behavior also. Knowledge-based systems can only use tools they know. We have done little in the direction of developing approaches to characterize our operators. Our current understanding of low-level operators does not even appear to be adequate for implementing an expert system like shallow reasoning approach. If we want computer vision systems capable of deep reasoning, then we will have to go farther than just developing an operator. We will have to analyze the behavior of the operator and develop techniques to represent this knowledge in programs explicitly so that the system can use it.

An example of misplaced emphasis is the research related to scale space. Much has been said about properties of edges in scale space, the feasibility of reconstruction of the signal from its scale space representation, but not much has been done to analyze the factors that result in dislocation of edges, creation of false zero-crossings, and disappearance of zero-crossings. Very few efforts have been made to develop approaches that can reason using scale space and higher knowledge about objects to detect and localize edges in images.

4. KNOWLEDGE

Interpretation of an image, in fact any perceptual process, uses knowledge and reasoning at every stage. This knowledge could be about entities in an image and the domain of application. Conventionally, in early processing, models are very general, like edges modeled as intensity steps; the later processes use explicit object models. Image interpretation requires knowledge about

- Environment, including illumination,
- Objects, their geometry, and other properties,
- Relations among different representations,
- Sensing process (image formation), and
- Behavior of operators for different processing tasks.

All systems use models either explicitly or implicitly to represent knowledge about all of the above. Clearly, efficacy and flexibility of a system depend on the models used. The more general the models, the more flexible the system. The major problem with general models is the computational power required to implement and use them. Specialized models, on the other hand, result in relatively rigid systems that are computationally efficient. Vision researchers realized these issues early in the game. All debates about top-down versus bottom-up control structures were related to these issues. Of course, in addition to computational efficiency many other issues should be considered in determining what models to use in a system and how to use them. Let us consider only the computational aspects first.

We can draw a parallel here with search and knowledge on one hand and general and specific models on the other. Search is very important, but without adequate use of knowledge, it is computationally impractical. These are the basic issues in artificial intelligence. In fact, the best explanation of why complex systems have to abandon the goal of perfection is well advocated by Simon in his principle of *satisficing*. In vision, by using general models we may provide flexibility, but at the cost of Herculean computation. Clearly, we need to select models that provide maximum flexibility at a reasonable computational cost.

Some of the popular general models used in early vision are

- step edges,
- smoothness of . . .
- rigidity of objects.

All these models are justifiable in limited contexts, but their popularity is more due to their generality and mathematical tractability than their suitability in computer vision. (Remember our drunk man!) Consider step edge models. What we are really interested in is the boundary between two regions. The intensity variations on the two sides of the boundary are almost always smooth. The step edge model is inadequate on both counts: it tries to capture discontinuity at a point, and it fails to capture the smoothness of intensity variations across the boundary. Why is it popular then? Because it is easy to model both mathematically and computationally. What is interesting is that we are reluctant to explore other models or other approaches to boundary detection, though research efforts during the last decade have shown us that edge detection using the step model is inadequate. What is more interesting is that we are willing to develop one more edge detector, but we do not want to develop objective and quantitative methods to evaluate the performance of an edge detector. About three decades of research on edge detection has produced N image detectors without a solid approach to evaluate the performance. In most disciplines, researchers evaluate the performance of a technique by a controlled set of experiments and specify the performance in clear objective terms. In edge detection, practically no efforts were even made to define objective measures. We still evaluate the performance of an edge detector by looking at the results. Interestingly, we define an edge model at a point and then complain that boundaries have gaps; our model is for a point but we judge its performance on curves!

The popularity of smoothness and rigidity assumptions has similar problems. Before one can really apply a smoothness assumption, an image must be segmented. The smoothness assumption is valid for a surface. Careless application of it in images without segmentation can only result in frustration. Application of regularization to any image without segmentation clearly shows the problems with these assumptions.

Similarly, the rigidity assumption has been used extensively in formulating structure from motion. Clearly, this assumption will be very useful after an object has been partitioned into its rigid components. The problem with both these assumptions is that they are *locally* correct assumptions and can be used if we know the areas of images where they should be applied. Unfortunately, the major problem in vision is to find those areas.

4.1. *Representations*

The goal of a perception system, whether biological or machine, is to create a model of the real world and to use this model for interacting with the real world. We can neither experience nor measure the physical world directly. A perception system must create a model of the world, or environment at a given time instant, using its past experience and the currently sensed information. In fact, it is this model of the environment and the world that allows interactions among perception, cognition, planning, and action components of an intelligent agent.

In building the environment model, a system uses its knowledge of the objects, knowledge of sensors, and general knowledge of the domain in which the system is to function, the location of the sensor. All these knowledge sources are very different from each other in the nature of their information representation and the amount of information. Moreover, as shown in Fig. 2, the information must be represented at several levels for implementation of efficient and robust recovery processes. It is clear that representation of information is one of the most important issues in a perception system.

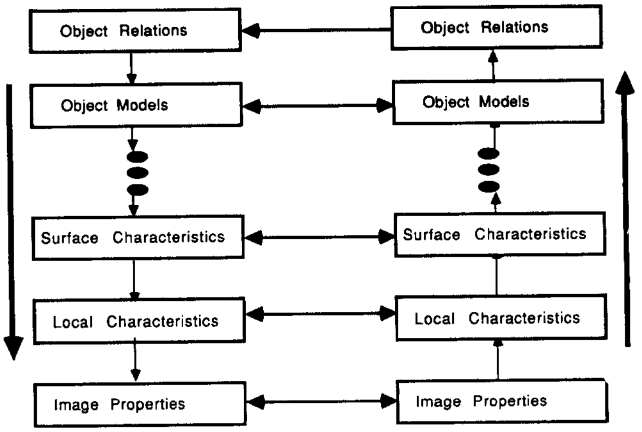


FIG. 2. A computer vision system applies knowledge at every level in its information recovery.

At the object level, we want representations that are very rich. These representations should be rich enough to represent millions of three-dimensional objects that may not differ much from each other. This representation should allow very subtle differences among objects as well as subpart hierarchies and gross representations. Approaches based on representing objects using only a few features or representing an object using features in multiple views may not be suitable. Feature-based approaches are too abstract to allow representing subtle differences and multiple-view-based approaches do not allow nature connections among the same views of an object.

One purpose of perception is to represent past experiences compactly and retrieve them efficiently, when needed. To allow compact and flexible representation, object models should be structured. Structured models may use parametric representation of components. Indexing schemes must be used for efficient retrieval. Moreover, at any given instant, the knowledge of the world as constructed by the system is incomplete, inexact, and uncertain. The representation should allow this incompleteness and uncertainty.

5. GOOD TRENDS

Fortunately, the last few years have shown some good trends in computer vision research. If we can strengthen those and start filling in other voids, computer vision may become not just an exciting field but also a very useful one.

The availability of range cameras encouraged many researchers to start addressing problems in surface characterization using the explicit depth information. Though many early approaches were direct extensions of the work in other areas, now researchers are studying differential-geometry-based approaches to understand surface characteristics better and use them in segmentation. Another encouraging trend in this direction is more attention to the use of geometric models in computer vision. The last few years have seen many new efforts start emphasizing the role of geometric models in object recognition and inspection. Though explicit three-dimensional reasoning using geometric models is still not very common, there is a trend in this direction.

The idea of integrating information from multiple sensors or multiple operators seems to be slowly maturing. After struggling with simple-minded approaches trying to combine information in image space, researchers are now starting to build incremental models in 3-D. Some intrinsic surface characteristics are finally getting careful attention. The last few years have seen some activity in understanding color; activity in understanding other images is also increasing.

The first few years of research in dynamic vision saw papers addressing problems related to many aspects of dynamic vision. Later, for almost one complete decade, most effort has been on myopic problems, such as recovering structure using a minimum number of frames or determining optical flow using two frames. Many other techniques in this area were based on first and second derivatives of optical flow and thus required third derivatives of intensity values, leading to methods useless for real images. Some recent work on image sequence processing shows that many seemingly complex problems can be solved by using appropriate techniques borrowed from systems engineering that do not appear powerful when considered locally. In general, more attention is being given to the analysis of long sequences of images to solve problems in dynamic vision.

Systems are now being considered to acquire images based on what needs to be done next. Though these systems do not do much reasoning yet, this is a step in the right direction. Similarly, the last few years have seen some attention to qualitative vision. The approaches based on qualitative reasoning are a good step in analyzing phenomena that can only be captured at a qualitative level.

6. CONCLUSIONS

Let us make computer vision a real discipline. This can be done by emphasizing theory, experimental aspects, and applications of computer vision. We have been guilty of neglecting the experimental aspects and applications in our field. Let us now bring balance into our field by systematically developing experimental computer vision as a discipline. Our journals should accept papers that report objective evaluation of operators and discuss a system to solve a real problem. For theoretical as well as experimental research, the authors should clearly state their assumptions and justify their claims based on the results that can be obtained under those assumptions. Vague justifications, such as subjective evaluations of images or justifications based on a psychophysical model, should not be allowed. An author should know where his contribution is and should substantiate tools using that discipline. It should be clearly pointed out to authors, when necessary, that one can be as ad hoc using mathematics as without. Both math hacking and computer hacking have a place in computer vision, but they are equally ad hoc.

To encourage experimental aspects, sharing of images and programs should be encouraged, even facilitated, by journals.

We find computer vision to be as challenging and as intriguing as we found it in our earliest days in this field, and we still have the strong belief that computer vision will soon be applied in many exciting applications. Let us do it.