COMPUTER VISION, GRAPHICS, AND IMAGE PROCESSING 36, 387–391 (1986)

Anything You Can Do, I Can Do Better (No You Can't)...*

KEITH PRICE

Powell Hall MC-0273, Intelligent Systems Group, University of Southern California, Los Angeles, California 90089-0273

Received February 3, 1986; revised March 5, 1986

Computer vision suffers from an overload of written information but a dearth of good evaluations and comparisons. This paper discusses why some of the problems arise and offers some guidelines we should all follow. © 1986 Academic Press, Inc.

INTRODUCTION

Many of the comments in this paper apply to any scientific domain and are not unique to computer vision, but some other research domains have well-defined methods for evaluating research (e.g., in medical research, does it help the patient?). I make several assumptions about the nature of computer vision research: it has a lot of small pieces; by itself each piece is meaningless (by itself computer vision is meaningless); no one has solved all the problems; many real problems are solvable with current techniques; a perfect system will never exist (human vision is not perfect); and people normally see what they "want" to see rather than what is there. I will assume that everyone is honest, but I also assume that people work under some pressure for results or publications and do not want any additional non-research work. First I will present a fictional story to illustrate the nature of the problem, followed by an outline of the problems that arise. Then I will give the results of an informal survey of published papers to see how they contribute to the problem. The paper will close with some suggestions for improving the current situation.

A SHORT STORY

A graduate student at a small eastern technical school determines that an operator, called the Homer Operator (HO for short), can be used to determine stereo disparities. She writes her thesis, which, as usual, contains many irrelevant details, and publishes several papers with all the details that seem relevant. The program, as all graduate programs, follows none of the standards for documented, maintainable code (thus it is never changed by the local users). Meanwhile, a professor at a smaller eastern university has been working on the stereo problem and wishes to compare his results with the new HO. He reads the papers, but the new implementation does not work. He reads the thesis and still his program fails. Finally he tries to get the programs, but is told that they are hard to find, understand, run, read, etc. (besides he does not have the same computer system). After these failures the professor abandons all attempts at real comparisons.

*Thanks to Ken Sloan for the title and for discussions which led to changes from the original version.

0734-189X/86 $3.00

At a large west coast university, another professor tells a new graduate student to reimplement the algorithm described in the original thesis and papers. Disparities, which seen reasonable, are generated by the program, and the student proceeds with research in motion, forgetting stereo. Eventually, another student tries the programs on completely new data and the program fails to produce meaningful results. This student, being adept at symbolic computation, and having nothing else to do, discovers that the original algorithm works properly only under extremely specific conditions, which were never explicitly discussed, but which often occur in practice. Her mathematical analysis results in the Milton Operator (MO) which works under all conditions, but requires effectively infinite time. Her work is hailed by theoreticians in the field, but the experimentalists still want a means to compute the stereo results accurately.

WHAT ARE THE PROBLEMS?

This story introduces some of the problems encountered in using the research of others. (The over-stated and under-practiced phrase "Stand on other's shoulders, not on other's toes" comes to mind.)

* How do you evaluate the work of others when you do not have their programs? You can look at their results in the papers, but that leaves many questions unanswered.

* How do you use the algorithms of someone when you do not have the other person? When asked by our contracting agencies what our plans are for transfer of technology, we (university folk in general) reply: companies should hire our graduates (not our graduate students).

* What does it mean when a reimplementation does not work? Who failed, the algorithm or the implementation?

* How do you compare results? Mine works 80% of the time by some measure, yours 80% by another measure, and they seem to agree 40% of the time.

* How do you control the tuning of algorithms? If the choice of a single threshold can greatly change the entire result, is the solution in the method for choosing the threshold or in the algorithm?

* How do you present (or extract from a presentation) the central idea? Many papers contain an important idea that is lost in the reporting of irrelevant results. Often the author may not be aware of the ultimately important idea that is the most important part of the paper.

These problems have a variety of solutions, most of which depend on people doing things right.

WHAT ARE SOME OF THE CAUSES?

Many of the problems are common to other scientific diciplines so that solutions utilized by others would be applicable. A major share of the failure to adequately compare and replicate results must lie with the author of the original paper—too often the paper does not provide enough guidance to the reader to evaluate and use the results. Likewise, readers (including reviewers) are very quick to dismiss an approach which builds on work with which they are not familiar, leading authors to

duplicate old papers in their new papers thus destroying the impact of the new work.

To explore how much current papers contribute to the problem, I undertook an informal survey of published papers (from *IEEE Trans. Pattern Anal. Mach. Intell.*, for July 1984 through July 1985—7 issues). I considered papers in computer vision, image analysis, and scene analysis (approximately 42 total papers).

Twenty of the papers presented results on either 1 or 0 images. Fifteen show 2–4 sets of results and 4 papers show 5–7. Three papers mention the application of the method on a large number of images (16, or 50 and more). Very few of the results are on pictures which also appear in other papers. (Stereo and segmentation papers not included in the survey have used common images in the presented results.)

A small portion of the papers (11) address a well-defined, clearly stated problem (this classification is hard to quantify and would vary according to the person classifying). About the same number (13) address a problem without consideration of whether it applies to anything, or make "unreasonable" assumptions.

These results are significant for several reasons. Computer vision must address real problems with real input data if it is to survive as a research area. Modelling human vision is a valid pursuit, but it will always be difficult to ignore the "high-level" aspects of human vision in any analysis of the "low-level" human visual system. Therefore I am uneasy about papers which claim to solve a computer vision problem and show no proof, or claim to solve a problem only if certain information is provided. These papers provide no way to judge their contributions and no way to compare alternative methods or to confirm new implementations.

One area where "non-result" papers are valid (and are lacking) is in a detailed analysis and comparison of various approaches to the same type of problem. This would lead to a better understanding of the real strengths of certain approaches and ultimately to improved algorithms and systems.

SOLUTIONS WHICH DO NOT SOLVE THE PROBLEM

One "obvious" solution is to require precise standard for everyone to follow. Such standards detract from research and ultimately may destroy innovative research. A few de facto standards do exist (sometimes conflicting standards for the same thing, but still a few alternatives) because of the sharing of basic programs and data. This is good and must be encouraged, but should not lead to the creation of closed cliques (clubs? cartels?) of researchers.

Publication delays are already so long that added burdens on reviewers would not help. But, reviewers must be aware of other work in the field and should require adequate proof in the paper. A set of test images has been discussed, but the implementation is difficult because of data formats, and the costs. It is always easier to call and ask for an individual image, in whatever format, than to rely only on some standard database.

WHAT SHOULD WE REALLY DO?

We ("we" is used to emphasize that researchers must impose the same standards on themselves as on others) must be both producers and consumers of research results; we should apply the same standards when reading and writing. Do not make demands you do not wish to keep. This section suggests a few steps to make life easier on all researchers.

When trying to solve a problem, formulate it in a way that it can be solved. The paper should also state the problem being solved and describe how that problem is solved by the reported method. For example, consider the problem of edge detection. (I personally do not create edge detectors, but I like to use them for certain problems.) Relying on the duplication of human performance may not produce good results for a given application, since people see subtle (even non-existent) edges in some cases and ignore strong ones in others. Finding ideal step edges is easy, but few such ideal steps occur in an image. Clearly, finding a reasonable subset of edges is easy—a lot of papers report on that. The hard problem is determining the actual advantage of one method over another, since this depends on the ultimate usage. There are many ways to compare edge detectors (e.g., step edges with added noise), but it is only when the edges are used that it matters.

Do not try to solve a simple problem with an overly complex method. (This is my favorite topic, but it really should take only a paragraph.) Consider several hard problems with trivial solutions—reading prices on grocery items (my butcher shop has scales that print bar-coded price stickers for each purchase); reading addresses on letters (again bar-type codes are now being used); reading numbers on railroad cars (again bar codes); medical imaging (not really trivial, but the problems change faster than the image-based solutions because of chemistry and technology). When the researcher develops a complex solution to a simple problem, the solution itself becomes more important than the original problem. Someone else may see only the original problem and will not understand why the large complex solution was chosen, and thus not see anything important in the solution.

Present lots of results. I was taught (by Raj Reddy) that a thesis requires results on at least six different images. Statistical studies use training sets and test sets to control this problem. We cannot produce 100 test samples when each sample requires hours of computation for generating the results. It is hard for six "natural" images to all have the same obscure property that makes everything work right. When a large number of images is tried, the limits of the program become clearer and parameters are harder to tune for specific images (general solutions are used). Beware of any research that has only one image in the results. There is a reason—it is difficult to replicate and difficult to run. Any realistic (commercial) system must run on thousands of images before it is acceptable.

From a purely scientific point of view, the fault in replication from the published papers may lie with the original author as much as the second researcher. Too many papers are vague and incomplete when presenting the exact problem and the exact solution.[1] Exactness does not imply length; it requires attention to the real problem being solved. The author must be allowed to build upon previous work and the reimplementer must be willing to read such works. It is not acceptable for a researcher to complain that something could not be done because of the failure to develop the underlying system. Unfortunately, the underlying system may require more effort than the new algorithm being tested. In any science, the necessary laboratory equipment is assumed; here it is a large software base which is assumed in any report of research.

[1]Bill Thompson, in a personal communication comments, "Good ideas are replicatable without knowing the details! (Corollary: If it depends on the details, it is probably not a good idea.)"

Universities have a need to train new graduate students in how to do computer vision research. The ideal training for a new student is the analysis and implementation of some known system. Give a graduate student several papers and a computer account, and soon there will be a new program. (This was my training, and I have seen new students prosper in this mode. We tried this to a limited extent in our lab last summer, and overall it was successful—in most cases we were converting programs from one system to another rather than implementing the work of outside researchers.) This will give the student some feel for what research is really like, and will provide other researchers with a means of comparing results. Not all students will be successful in this; often they will get bogged down in trivial details and will not complete the original project.

CONCLUDING COMMENTS

Computer vision (indeed most of computer science) is primarily concerned with creating and studying "artificial" systems (unlike biology) which studies natural systems). The creating and studying tend to be separate and unrelated efforts—one group "creates" another group "studies," but they do not use the same model of what to do. This is not helped by the fact that created methods rarely survive long enough for any extensive study, even by their creator.

Research in computer vision has suffered from a lack of building on past work. Only through real effort can programs be shared among a large group of researchers. Researchers should make the effort to obtain implementations of other researchers' systems so that we can better understand the limitations of our own work.