THE CHALLENGE OF GENERAL MACHINE VISION[1]

Theo Pavlidis
Computer Science Dept.
Stony Brook University
Stony Brook, NY 11794, U.S.A.

1. Introduction

The last two decades have seen several successful applications of Machine Vision and that has raised hopes that we may be able to solve the general Machine Vision problem in the near future. In my opinion this is not likely to happen. There are three types of difficulties that impede progress in general machine vision that I will address in this paper.

- The complexity of human vision: Bottom-up and top-down processes are tightly interwoven and we have no good models for dealing with that.
- The fact that perceptual similarity is not the same as mathematical similarity.
- The illusion of progress by relying on "proofs by example" that are not always valid.

Successful applications have been made possible by removing such difficulties because of the special nature of each particular problem and I will list several such examples. The lists are by no means complete. Listing all successful machine vision applications is well beyond the scope of the paper. However, all applications I am aware of are problem specific.

Note: The paper is accompanied by an Appendix (http://theopavlidis.com/MachineVision/Appendix.pdf) that contains all color illustrations.

2. The Complexity of Human Vision

I believe that machine vision researchers have grossly underestimated the difficulty of the general problem ignoring the evidence from Psychobiology and Neuroscience. More than 20 years ago Bela Julesz wrote "In real-life situations, bottom-up and top-down processes are interwoven in intricate ways," and "progress in psychobiology is ... hampered ... by our inability to find the proper levels of complexity for describing mental phenomena" [1]. V.S. Ramachandran wrote "Perceptions emerge as a result of reverberations of signals between different levels of the sensory hierarchy, indeed across different senses" and criticized the view that "sensory processing involves a one-way cascade of information (processing)" [2].

[1] To appear in the January 2014 issue of the journal of *Signal, Image and Video Processing*.

Figure 1 illustrates the complexity of the human (and animal) visual processing. The parts inside the red contour have received little attention in the machine vision literature. Humans (and animals) bring context knowledge in all their perceptual tasks, especially in vision. We have expectations of what we are about to see! The adaptive significance of such ability cannot be overstated. The existence of optical illusions confirms the importance of such expectations. The block labeled "Prior Knowledge" refers to knowing what we are about to see; for example, that the image is an MRI scan of the human brain from the top. In my opinion the only machine vision problems solved are those where prior knowledge drives the feature extraction and we have no need for the hypothesis generation/hypothesis testing part.

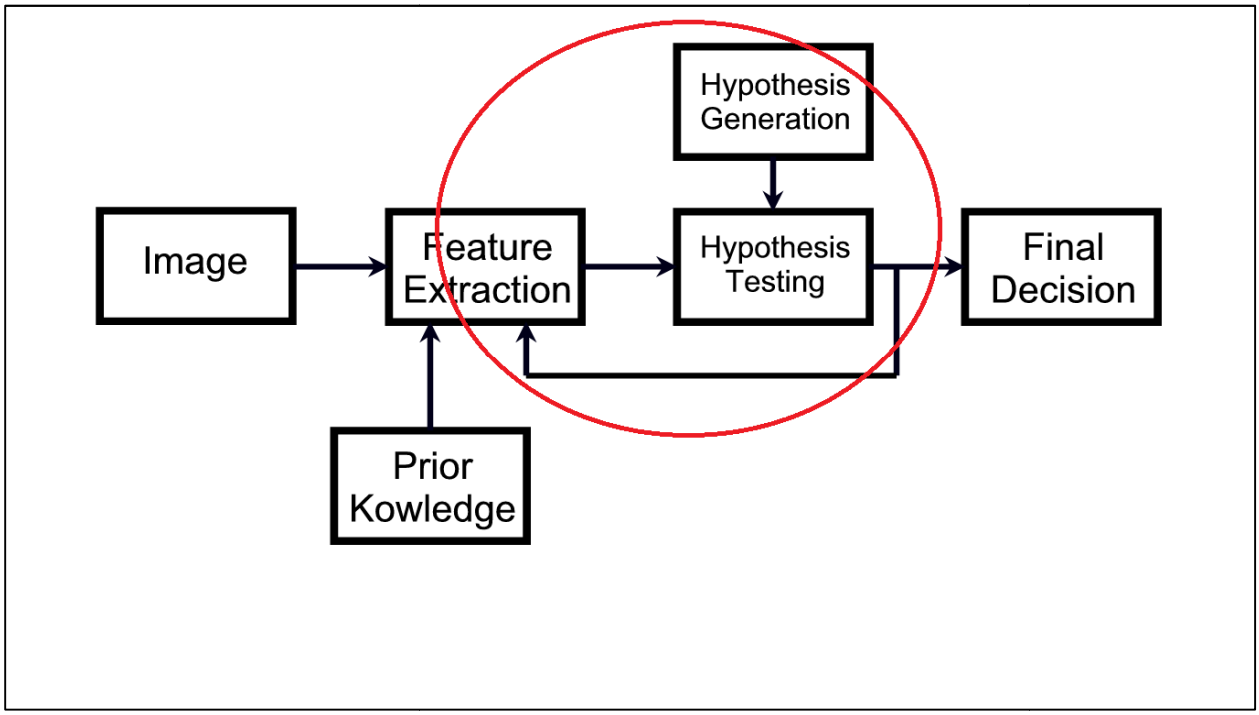

Figure 1: A simplified block diagram of the human visual system. I use the term "feature extraction" loosely to describe the mapping of pictorial data into mathematical structures.

For simplicity, I will use examples with text to illustrate the complexity of human vision. Look at these two sentences

New York State lacks proper facilities for the mentally Ill.

The New York Jets won Superbowl III

Human readers may ignore entirely the shape of individual letters if they can infer the meaning through context.

Figure 2 (adapted from [3]) shows the interaction between bottom up and top down processes. In the left drawing most people will describe the illustration as that of the letters F, H, K, and Z. Yet the gap that separates K from Z is much smaller than the gaps in the strokes of F and H that are ignored. The right drawing shows that tentative binding on the letter shapes (bottom up) is finalized once a word is recognized (top down). Word shape and meaning override early cues. The displayed text does not spell "The behavior of Machines."

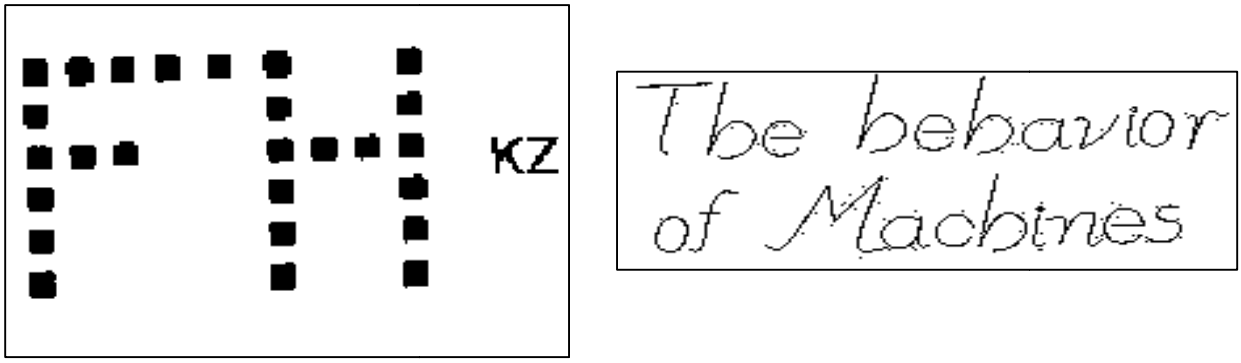

Figure 2: *Left:* Letter recognition overrides stroke detection. *Right:* Word recognition overrides letter recognition.

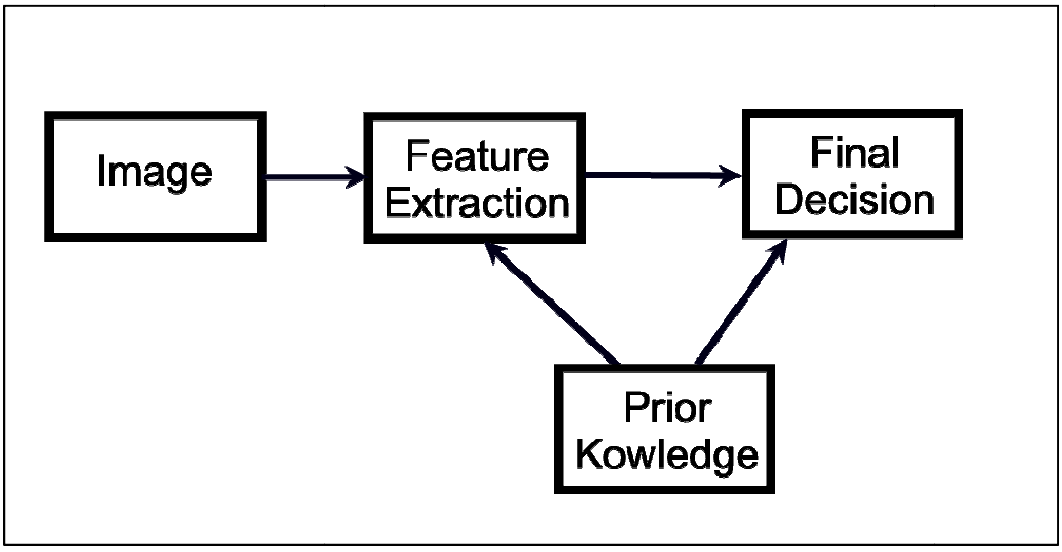

Figure 3: A simplified block diagram characteristic of applications of computer vision in a specific domain.

Successful applications of machine vision are characterized by having sufficient prior knowledge to guide feature extraction as well as the final decision (Figure 3).

One example is Image-Guided Surgery developed at M.I.T. and the Brigham and Women's Hospital at Boston [4]. A key part of the process is segmenting the tumor from healthy tissue in MRI images. However the MRI machines used do not have uniform gain across the image and that complicates the segmentation. Still the problem is physically well defined. We know that the image to be analyzed is an MRI brain scan and we understand the various types of distortion present. The application of machine vision methods has enabled surgeons to operate on tumors that used to be characterized as inoperable because of their proximity to vital areas of the brain.

Another example is the wildfire detection system developed at Bilkent University of Ankara, Turkey [5]. Fire wardens often miss early signs of a wildfire and issue an alarm only after the fire has spread. The imaging system relies on the spectral characteristics of smoke and the complicating factor is the presence of clouds. Again the problem is physically well defined. According to Prof. Enis Cetin, it was politically impossible to eliminate fire wardens so the system was fine tuned to never miss a fire at the expense of some false alarms that would be reset by the wardens. False alarms are caused typically by overcast skies and the system uses a learning algorithm to reduce such incidence. It has been deployed in 77 locations in Turkey and 2 in the U.S. and during 2007-2012 detected 241 wildfires [*Personal communication*].

There are numerous other examples involving relatively simple imaging challenges such as the system identifying and measuring bright spots in micro-arrays in high throughput biology [6, 7].

I conclude this section from a quote by John Tsotsos that summarizes nicely the prospects for general Machine Vision:

"Vision as we know it seems to involve a general purpose processor that can be dynamically tuned to the task and input at hand. This general purpose processor can solve a class of visual perception problems (the class of 'at a glance' problems) very quickly but for more difficult problems time is the major cost factor, that is, those problems take longer to solve using that same general processor but tuned for specific sub-problems that together solve the remaining, 'more than a glance', problems." ([8], p 248)

3. Perceptual similarity is not the same as mathematical similarity

Figure 4 shows an old example. The shape on the right is elliptical, thus pixel by pixel comparison produces a greater distance from the other two than the others

have from each other. On the other hand human observers point to the middle shape as being the one not belonging with the others. One could "fix" this example by using other measures than pixel location differences but then one can construct other examples where the new measures do not provide results in agreement with human perception.

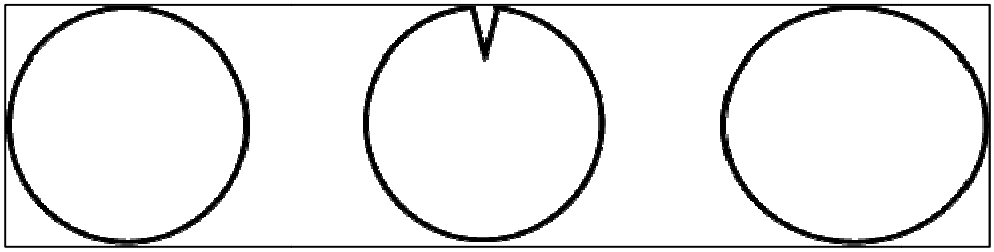

Figure 4: Which one of the three shapes does not belong with the other two?

This example is only the "tip of the iceberg." It turns out that humans do not agree on image similarity unless the images are almost identical or quite different [9, 10]. Figure 5 (Appendix: http://theopavlidis.com/MachineVision/Appendix.pdf) may convince you. There are two groups of six images: in each group try to rank the images A, B, C, D, and E according to their similarity to the target images. While most observers agree on the most similar and least similar images to the target they disagree in the ranking of those in between.

Such findings help explain the difficulties that face Content Based Image Retrieval (CBIR). Figure 6 (Appendix: http://theopavlidis.com/MachineVision/Appendix.pdf) illustrates the poor performance of such a system where images of sport cars are returned in response to a query consisting of a shoe image. Prof. Greg Zelinsky pointed out that such response may also be typical of low level human visual processing: "This is what a normal human might perceive very early in their visual processing. For example, if you are looking for a black shoe in a scene, your eye might be drawn to a black sports car based on low level visual metrics. This is a kind of human image retrieval error (your gaze "retrieved" the wrong pattern) that happens hundreds of times each day." [*Personal communication*]

In contrast to the failures of the general image retrieval problem there have been spectacular successes in biometrics. The reason for such successes is that we look for nearly identical images rather than just similar. In theory the images should be identical but some minor differences may be introduced during the digitization and encoding process.

Since 1999 the FBI has been using an automated fingerprint identification system that has been quite effective in solving crimes [11].

A more recent development is iris recognition [12, 13] for which there have been commercial applications, including the system developed by IrisGuard [14]. According to Imad Malhas, CEO of IrisGuard, one of the customers of the company is the Cairo-Amman Bank that uses iris recognition for customer identification at its ATMs dispensing with the need for cards and PINs. [*Personal communication*]

Fingerprints and iris scans are called hard biometrics because the system deals with images of a very specific kind, therefore facilitating the analysis in terms of the issues discussed in the previous section. A more challenging application is soft biometrics such as scars and tattoos because the system must deal with a broader category of images. However, the matching process still relies on finding nearly identical images and a successful system for soft biometrics has been developed at Michigan State University [15].

Another area where computer vision has made significant progress is industrial inspection because we are trying to closely match an image to a prototype. One example is a wheel alignment system [16] as well as several products of Microscan [17].

4. "Proof" By Example

It has been an old tradition in machine vision to demonstrate the effectiveness of a method by showing its results on a set of examples. Validation by experimentation is accepted in many fields of science provided that the experimental results are repeatable. Unfortunately, the description of methodologies in machine vision is rarely sufficient to allow attempts to repeat the results. Vandewalle *et a*l [18] discuss this problem in the context of general signal processing and offer suggestions to remedy the situation.

There are additional problems with the current practice. Torralba and Efros [19] have shown that the datasets used for validation of machine vision methods have inherent biases that undermine the repeatability of the results. There is an old apocryphal story (dating from the 1970's) that a company developed a system to recognize the presence of a tank in an image. It turned out that all pictures with a tank present had been taken in bright day light and all pictures without a tank had been taken in the evening (it may have been the other way around). So all the system was doing was detecting brighter from dimmer images. This is an extreme example of **fortuitous recognition** that illustrates the need for careful selection of the images in a testing dataset.

But there is an even bigger challenge. Experimental proofs are particularly hard in machine vision because the number of possible images is truly astronomical. I have

shown [20] that 10^{56} is a very conservative lower bound to the number of all possible meaningful and valid images and that number could be as high as 10^{400}. (The ambiguity in the limit is because it is hard to test objectively whether two images can be differentiated by a human observer. See [20] for more on this issue.) Testing a method on a set of only a few thousands images tells us nothing about its general validity.

It is possible to expand a given set of images by generating additional images from it by transformations that do not change the semantics and obtain more reliable statistical results. A few years ago I worked on a method for Image Retrieval (CBIR). The method did quite well on a set of about 5,000 images. I expanded that set by a factor of about 100 by generating new images from the originals by simulating over- and under-exposure, shadows, and other visual artifacts. The method did very poorly on the set of 500,000 images. Figure 7 (Appendix: http://theopavlidis.com/MachineVision/Appendix.pdf) provides an illustration of the results. For details see [21].

An area of Machine Vision that has relied on large volumes of data, including artificially generated images, is Optical Character Recognition (OCR). Today OCR packages come bundled with scanners and perform quite well on printed documents in English using standard fonts. Therefore the OCR problem can be considered as solved for such documents. It still remains open for other languages and even English when printed with uncommon fonts. Recognition of the text in whole books is a topic of active research [22, 23]. It is worth pointing out that the authors of [23] used a set of about two million words (11 million characters) to design their system and an even bigger set (6 million words) to test it. The design of bar code readers that are expected to have a misread rate better than one in a million scans (while maintaining a rejection rate under 1%) relies on artificially generated data based on physical models of the distortions and noise that might be encountered.

A recent publication from Google [24] has used a data set of ten million 200x200 images. The authors claim that they were able to extract features from unlabeled data and obtained 15.8% accuracy in recognizing 22,000 object categories. The recognition rate is clearly much better than a random guess, but it is well below what I would expect from a practical recognition system.

5. A Final Note

I conclude with an example of an industrial system where I had direct personal involvement while I was working at Symbol Technologies. The objective was to measure the dimensions of a rectangular shipping box from one image. Because of the constraints on the shape of the object it is possible to estimate the relative

dimensions from one view. If we know the actual distance of the object from the camera we can find the actual dimensions. (This was measured by the parallax of two laser beams.) A major challenge is that because of labels and other markings on a box the contrast between the box and its surrounding is lower that the contrast within the box. On the other hand we can safely assume that the box occupies most of the image because the scanning device is handheld and aimed at the box. Therefore we developed a method to detect only long linear edges and determine the outline of the box as the convex hull of such lines. If the convex hull is a hexagon the system signals success. The actual method is a bit more complicated and it is described in detail elsewhere [25, 26]. I should add that there was no training data set because of the unpredictability of the challenges that potential customers would provide. Instead all parameters were set adaptively for each image and the design of the algorithm relied on physical modeling. The resulting device performed well in field tests and some shipping companies expressed interest for using it in their shipping hubs. Symbol Technologies had hoped that it would be used by each driver, a much bigger potential market. The hub market was not big enough to justify production of the device.

Acknowledgements

While writing this paper I received significant help from several people who commented on drafts of the paper and/or provided pointers to literature and research activities that I had missed. They are (in alphabetical order): Henry Baird (Lehigh Univ.), Alex Berg (Stony Brook Univ.), Kevin Bowyer (Notre Dame), Enis Cetin (Bilkent Univ., Ankara), Jelena Kovacevic (Carnegie Mellon Univ.), Mike McCann (Carnegie Mellon Univ.), Imad Malhas (IrisGuard), Paul Pavlidis (Univ. of British Columbia), Fabio Roli (Univ. of Cagliari, Italy), Dimitri Samaras (Stony Brook Univ.), Ray Smith (Google), John Tsotsos (York Univ., Canada) and Greg Zelinsky (Stony Brook Univ.).

Obviously, these people need not share all of my views on Machine Vision.

References

[1] B. Julesz “Early vision and focal attention,” *Reviews of Modern Physics*, **63**, (July 1991), pp. 735-772.

[2] V.S. Ramachandran and S. Blakeslee *Phantoms in the Brain*, William Morrow and Company Inc., New York, 1998 (p. 56).

[3] T. Pavlidis "Context Dependent Shape Perception", in *Aspects of Visual Form Processing*, (C. Arcelli, L. P. Cordella, and G. Sanniti di Baja, eds.) World Scientific, 1994, pp. 440-454.

[4] W. Eric L. Grimson, Ron Kikinis, Ferenc A. Jolesz and Peter McL. Black "Image-Guided Surgery," *Scientific American*, June 1999, pp. 62-69.

[5] O. Gunay, B.U. Toreyin, K. Kose, A. E Cetin "Entropy-Functional-Based Online Adaptive Decision Fusion Framework With Application to Wildfire Detection in Video," *IEEE Trans. Image Processing*, **21**, May 2012, pp. 2853-2865.

[6] Figueroa A, Tsai PS, Bent E, Guo R. "Robust spots finding in microarray images with distortions," in http://www.ncbi.nlm.nih.gov/pubmed/19162915

[7] http://www.affymetrix.com/

[8] J. K. Tsotsos *A Computational Perspective on Visual Attention*, MIT Press, 2011.

[9] W. Zhang, D. Samaras, and G. Zelinsky "Classifying objects based on their visual similarity to target categories," *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2008, pp. 1856-1861.

[10] X. Zhao, M. G. Reyes, T. N. Pappas, D. L. Neuhoff "Structural Texture Similarity Metrics for Retrieval Applications," *Proceedings Int. Conference on Image Processing*, 2008, pp. 1196-1199.

[11] A. K. Jain, J. Feng, and K. Nandakumar "Fingerprint Matching," *IEEE Computer*, **43** (February 2010), pp. 36-44.

[12] K. W. Bowyer, K. Hollingsworth, P. J. Flynn "Image understanding for iris biometrics: A survey" *Computer Vision and Image Understanding*, **110** (May 2008), pp. 281–307.

[13] M. J. Burge and K. W. Bowyer (Eds.) *Handbook of Iris Recognition*, Springer, 2013.

[14] http://www.irisguard.com/index.php

[15] J-E. Lee, A. K. Jain, R. Jin "Scars, Marks, and Tattoos (SMT): Soft Biometric for Suspect and Victim Identification," *Biometrics Symposium*, Tampa, Sept. 2008.

[16] R. Burns Leigh Jr., Daniel R. Dorrance, Thomas J. Golab, Mark S. Shylanski, Timothy A. Strege "Common reference target machine vision wheel alignment system," *US Patent No. 7164472 B2, 2007.*

[17] http://www.microscan.com/

[18] P. Vandewalle, J. Kovacevic, and M. Veterlli "Reproducible Research in Signal Processing," *IEEE Signal Processing Magazine*, May 2009, pp. 37-47.

[19] A. Torralba , A. A. Efros "Unbiased Look at Dataset Bias," *CVPR* 2011, pp. 1521-1528.

[20] T. Pavlidis "The Number of All Possible Meaningful or Discernible Pictures," *Pattern Recognition Letters*, **30** (2009) pp. 1413-1415.

[21] http://www.theopavlidis.com/technology/CBIR/index.htm

[22] Pingping Xiu and Henry S. Baird "Whole-Book Recognition," *IEEE Trans. PAMI*, 2012, pp. 2467-2480.

[23] Dar-Shyang Lee and Ray Smith, "Improving Book OCR by Adaptive Language and Image Models," *Proc., 2012 10th IAPR Int'l Workshop on Document Analysis Systems, IEEE*, pp. 115-119.

[24] Q. V. Le et al "Building High-level Features Using Large Scale Unsupervised Learning," *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

[25] T. Pavlidis, E. Joseph, D. He, E. Hatton, and K. Lu "Measurement of dimensions of solid objects from two-dimensional image(s)," *U. S. Patent 6,995,762*, February 7, 2006.

[26] K-F Lu and T. Pavlidis "Detecting Textured Objects using Convex Hull," *Machine Vision and Applications*, **18** (2007), pp. 123-133