WHY MACHINE VISION TURNED OUT TO BE SO HARD?

Theo Pavlidis
Distinguished Professor Emeritus
Stony Brook University
t.pavlidis@ieee.org
http://www.theopavlidis.com/

Why is Machine Vision so Hard?

- ❖ Organisms with complex visual systems have existed for over 300 million years.

- Speech has existed for less than 200 thousand years.
- Writing has existed for less than 5,000 years.

General Challenges to Machine Vision

- We need to replicate complex transformations that the (human/animal) brain has evolved to do over hundreds of millions of years.
- We have to deal with the fact that processing is not unidirectional and it is also affected by other factors besides input (context both inside and outside the image). Visual illusions (far more common than auditory illusions) attest to that.

Three Specific Reasons

1. **Bottom-up and Top-down processes are tightly interwoven and we have no good models for that.**
2. Perceptual similarity is not the same as mathematical similarity.
3. Machine vision has relied a lot on “proofs by example” that are not always valid.

What Neuroscientist Say - 1

- “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”

- Source: B. Julesz "Early vision and focal attention", *Reviews of Modern Physics*, vol. 63, (July 1991), pp. 735-772.

What Neuroscientist Say - 2

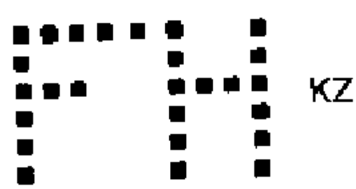
- “Perceptions emerge as a result of reverberations of signals between different levels of the sensory hierarchy, indeed across different senses”. The authors then go on to <u>criticize</u> the view that “sensory processing involves a one-way cascade of information (processing)”

- Source: V.S. Ramachandran and S. Blakeslee *Phantoms in the Brain*, William Morrow and Company Inc., New York, 1998 (p. 56)

In Other Words

- 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 an ability cannot be overstated.

Reading Demo – 1*



It is hard to explain the human ability of reading dot-matrix print and fine laser print by purely bottom up processes.

* Adapted from 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.

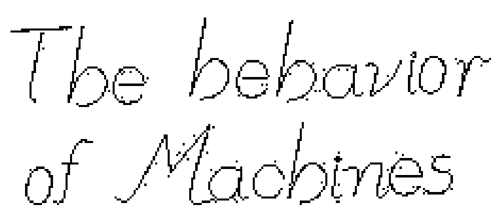
Reading Demo -2

New York State lacks proper facilities for the mentally III.

The New York Jets won Superbowl III.

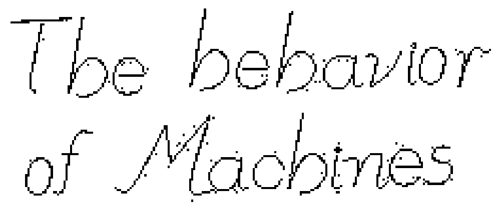
- **Human readers may ignore entirely the shape of individual letters if they can infer the meaning through context.**

Reading Demo – 3*

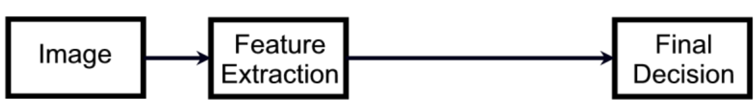


* Adapted from T. Pavlidis "Context Dependent Shape Perception", in *Aspects of Visual Form Processing*, (C. Arcelli, L. P. Cordella, and G. Sanniti di Baja, eds.) World Scientific, **1994**, pp. 440-454.

Reading Demo - 3



Tentative binding on the letter shapes (bottom up) is finalized once a word is recognized (top down). Word shape and meaning override early cues.



Do we really understand human vision?

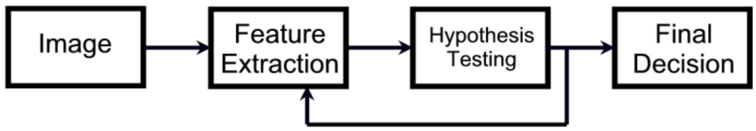
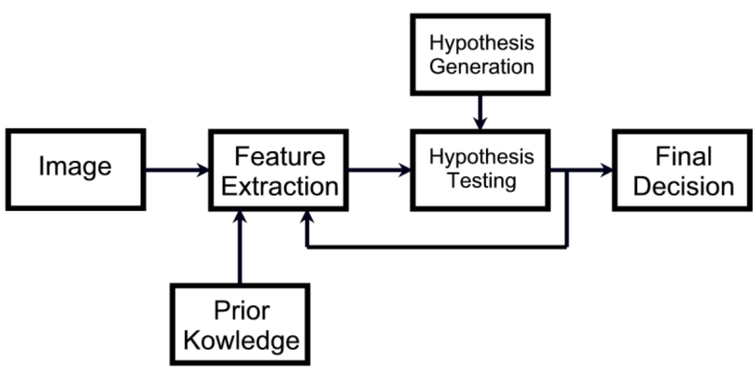
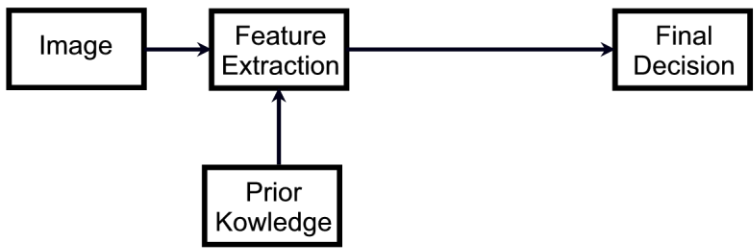


Image
Feature Extraction
Hypothesis Testing
Final Decision


Hypothesis Generation
Image
Feature Extraction
Hypothesis Testing
Final Decision
Prior Kowledge



A Special Case

- In the 1980's Edge Detection was a "hot topic." In retrospect that attention was misplaced.
- A simple edge detector (such as Sobel's) performs quite well, if supplemented by domain knowledge: what kind of shapes we are trying to see.

An Example

Find the outline of the box in the image below
(ultimate goal is to find the dimensions of the box):

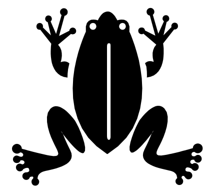
A paradox

- Human viewers have no trouble identifying the box and its outline.
- Application of Edge Detection or Segmentation produces a “mess:”
  - Contrast inside the box may be higher than contrast between the box and the background.

An Inspiration from Nature

- In a classical paper J. Letvin *et al* showed that the frog's visual system responds to only two kinds of stimuli:
  - fast moving, high contrast small shapes (food) or
  - decrease in the ambient illumination (danger).

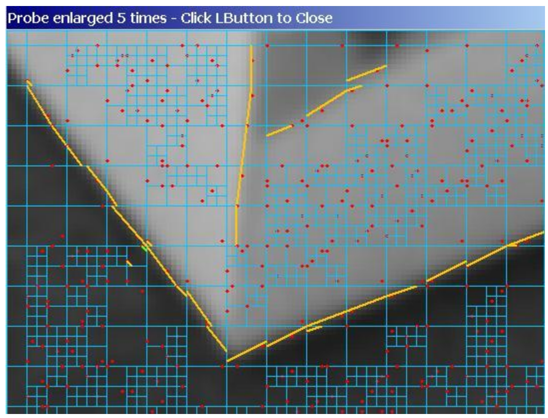
  [*Proceedings of IRE*, 1959]



Implications for the box dimension problem

- The system should look only for hexagonal shapes occupying most of the image.
- This means that the only edges of interest should be lines of length comparable to the dimensions of the field of view.
- Such lines should form a convex set.
- The convex set should be a hexagon.

Use **(Long) Line Detection** as the first step (rather than segmentation or edge detection)



Summary of the Box Outlining Method

- After a collection of long lines is found, their convex hull is determined.
- If the convex hull is a hexagon (outline of a rectangular object) occupying most the image, we are done.
- If not, certain lines, marked as “noise” are removed and we try again until we succeed in finding such a hexagon or determine that such a shape does not exist.

Publications

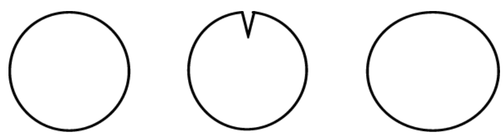
- T. Pavlidis, E. Joseph, D. He, E. Hatton, and K. Lu "Measurement of dimensions of solid objects from two-dimensional image(s)" *U. S. Patent 6,995,762*, February 7, 2006.
- Ke-Fei Lu and T. Pavlidis "Detecting Textured Objects using Convex Hull" *Machine Vision and Applications*, **18** (2007), pp. 123-133.
- On the Web: http://www.theopavlidis.com/technology/BoxDimensions/overview.htm

Three Reasons

1. Bottom-up and Top-down processes are tightly inter-woven and we have no good models for that.
2. **Perceptual similarity is not the same as mathematical similarity.**
3. Machine vision has relied a lot on “proofs by example” that are not always valid.

An Old Example

- Mathematical tools are a poor substitute for knowledge of human perception:



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

Image Retrieval

- The last decade has seen a lot of research on Content Based Image Retrieval (CBIR), trying to replicate for images what Google does for text.
- Results have been, at best, unimpressive.
- Why?

CBIR - 1

- Humans do not agree on image similarity unless the images are almost identical or quite different.
- Sources:
  - 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.
  - 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.

CBIR - 2

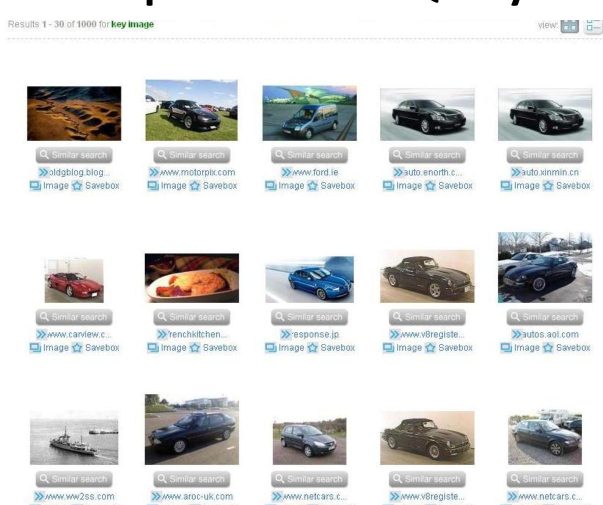
- CBIR that looks for nearly identical images can be quite successful. A prime example is the use of tattoos as soft biometrics in the research led by Anil Jain of MSU.
  - J-E. Lee, A. K. Jain, R. Jin "Scars, Marks, and Tattoos (SMT): Soft Biometric for Suspect and Victim Identification," *Biometrics Symposium*, Tampa, Sept. 2008.
- Human observers bring their own personal and cultural biases in rating the similarity of pairs of different images making automation of the process hopeless.

A Test of CBIR

- The image below was submitted as a query to the GazoPa CBIR site.*

- * No longer in existence.

Response to the Query



A Psychologist’s Explanation

- "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.“

  Greg Zelinsky (Dept. of Psychology, Stony Brook University) – Pers. Communication.

Face Recognition - 1

Face Recognition - 2

Face Recognition - 3

Face Recognition - 4

Find the Terrorist!

Results of the Robotics Institute, CMU program. A green rectangle is overlaid on any face detected. A major miss is evident.

Three Reasons

1. Bottom-up and Top-down processes are tightly inter-woven and we have no good models for that.
2. Perceptual similarity is not the same as mathematical similarity.
3. **Machine vision has relied a lot on "proofs by example" that are not always valid.**

"Proof" by Example

- An algorithm is applied to a set of images and its parameters are chosen to give a satisfactory results on subset of these images (*Learning Subset*). Then the algorithm is tested on the remaining images of the set (*Testing Subset*) and if the results are satisfactory the algorithm is considered a success.
- However, that particular set is unlikely to be a representative example of all images.
- The space of all possible image is huge!!!

What is the Number of All Possible Images?

- 10^{56} is a very conservative lower bound to the number of all possible meaningful/valid images. The number of all meaningful/valid images is at least as high as 10^{400}.
- See:
  - T. Pavlidis "The Number of All Possible Meaningful or Discernible Pictures" *Pattern Recognition Letters,* vol. 30, (2009) pp. 1413-1415.
  - http://www.theopavlidis.com/technology/CBIR/PaperK/draftK1.htm

An Illustration - 1

- 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.

An Illustration - 2

The picture in the middle is a brightened version of the picture on the left but two different sets of feature measures classify it as being closer to the picture on the right. The values were:

First feature set: *dist(Left, Mid)* = 25, *dist(Left, Right)* =25, *dist(Mid, Right)* = 0.

Second feature set: *dist(Left, Mid)* = 42, *dist(Left, Right)* = 50, *dist(Mid, Right)* = 23.

http://www.theopavlidis.com/technology/CBIR/PaperB/vers3.htm

Acknowledgements - 1

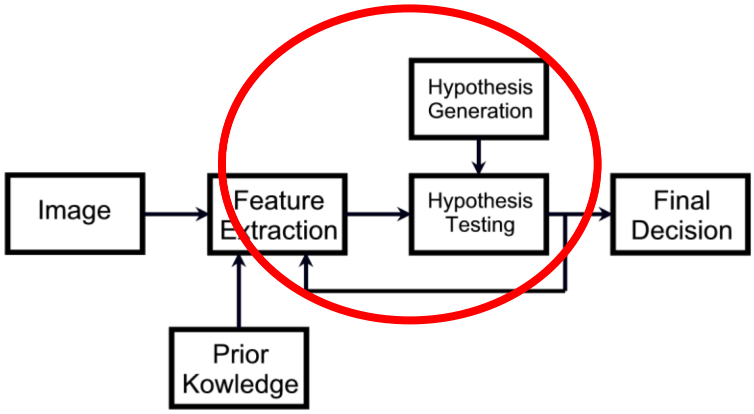
- The ideas expressed in this talk developed over many years and through the interaction with several people.
- The most important interaction was with Eric Grimson (MIT) whom I had invited to be a consultant at Bell Labs when I was there (1980-86).
- We faced the challenge of detecting intermediate structures and concluded that imposing mathematical smoothness constraints was both too severe and too lax a requirement.

Acknowledgements - 2

- Too severe because it did not allow for discontinuities and too lax because it may not capture all the constraints expected in the physical world.
- The conclusion of these discussions was that until we understood such “middle vision” the general machine vision problem could not be solved.
- As a result, the most promising research areas were in specific applications where a fairly complete model of the images under consideration was available.

Acknowledgements - 3

- The first public expression of these ideas on my part was at the 1986 ICPR (Paris) where I gave a talk: ``Why Progress in Machine Vision is so Slow.’’ Reprinted in *Pattern Recognition Letters,* **13** (April 1992), pp. 221-225.
- The block diagrams used to illustrated visual perception have benefitted from comments by Prof. Greg Zelinsky (Psychology, Stony Brook University).


Hypothesis Generation
Image
Feature Extraction
Hypothesis Testing
Final Decision
Prior Kowledge