HOW THE BRAIN DEALS WITH THE COMPUTATIONAL COMPLEXITY OF VISION:
A Different Kind of Dimensionality Curse

John K. Tsotsos
Dept. of Computer Science & Engineering and Centre for Vision Research
York University, Toronto, Canada
Outline

• The General Vision Problem
• The Role of Computational Complexity
  – Formal and Empirical Results
  – Constraints on a Sufficient Architecture
  – Implications
• The Selective Tuning Model of Attention and Vision
• Predictions
• Conclusions
The General Vision Problem?

- the act or power of sensing with the eyes – Dictionary.com 2009

- the special sense by which the qualities of an object (as color, luminosity, shape, and size) constituting its appearance are perceived and which is mediated by the eye - Merriam-Webster's Medical Dictionary 2002

- Vision is the result of some form of unconscious inferences: a matter of making assumptions and conclusions from incomplete data, based on previous experiences - Helmholtz 1896

- Vision proceeds from a two-dimensional visual array (on the retina) to a three-dimensional description of the world as output – Marr 1982

- Vision is the process of deriving purposive space-time descriptions – Aloimonos 1991
The General Vision Problem?

Vision is the main sense that allows humans to navigate and manipulate their world, to purposefully behave in the world.

What object, scene or event that you know best matches a given set of location/measurement pairs?
Visual Search


Feature Search

Conjunction Search

Search mechanisms inferred from slopes

© J.K. Tsotsos 2009
Looks like time complexity?

Sub-area of CS known as Computational Complexity is concerned with the cost of achieving solutions to problems (*not only algorithms*), in terms of time, memory and processing power as a function of problem size.

Since attention is so often accompanied by vague discussions of ‘capacity limits’ and ’bottlenecks’ and the like, it might be that steps towards formally quantifying these notions might be found there.

Maybe complexity theory can help...
Two Kinds of Visual Search


Unbounded Visual Search

Recognition where no task guidance to optimize search is permitted.
Corresponds to recognition with all top-down connections in visual processing hierarchy removed. Pure data-directed vision.

Theorem 1: Unbounded Visual Search is NP-Complete.

Bounded Visual Search

Recognition with knowledge of a target and task in advance, and that knowledge is used to optimize the process.

Theorem 2: Bounded Visual Search has time complexity linear in the number of test image pixel locations.
Definition of Unbounded Visual Match:

Given a test image ($I$ – the image is a set of $P$ pixels) is there a set $I'$ such that it simultaneously satisfies

$$\sum_{a \in I'} \text{diff}(a) \leq \theta \quad \text{and} \quad \sum_{a \in I'} \text{corr}(a) \geq \phi$$

The two separate criteria here require a solution to satisfy an error bound and to also be the maximal set (or rather, a large enough set) that satisfies that bound. In this, way trivial solutions are eliminated.

Not a sufficient characterization for real vision because it does not capture any of the variability in appearance, viewpoint, etc. But it’s hard enough!
Empirical Median Case Analysis


Criticism: The brain did not evolve to solve worst case problems

Developed algorithm for generating random instances of trihedral scenes

Kirousis and Papadimitriou (1988) proved that both the labeling problem and the realizability problem are NP-complete even for the simple case of trihedral, solid scenes.

Examined median case complexity of labelling the scenes with 2 algorithms:

- blind depth-first search with backtracking
  (result - exponential in number of junctions)

- informed best-first search with backtracking
  (result - linear in number of junctions)
Generating Random Polygons

Fig. 3. (A) The support region and a certain number of lines drawn at random. (B) The graph after the intersection points falling outside of the support $S$ have been removed, along with the edges incidents to them. (C) Filling elementary polygons at random may cause the creation of knots. (D) The polygon after the elimination of the knot.
Generating Random Trihedral Scenes

An ideal algorithm A for the generation of random trihedral scenes should have the following properties:

- completeness: every possible polyhedron P (or set of polyhedra) can be generated by A, that is, every P is a possible output of A;
- uniformity: all possible polyhedra appear with the same probability;
- polynomial complexity: the cost of generating a random polyhedron should be polynomial in the number of vertices of the polyhedron.

The extended trihedral world includes all objects which obey these requirements:
(i) they are solid, i.e., there are no hanging edges or faces;
(ii) they are opaque;
(iii) exactly three planes meet at a vertex;
(iv) there are no isolated contact edges; exactly two planar faces meet at every edge;
(v) there are no knots, i.e., isolated contact points between different portions of the scene.

The definition above includes the possibility of having polyhedra with polyhedral holes.

Generation proceeds as for polygons but in 3D (see paper)
Random Scenes

- occluding contour
- visible convex edge
- visible concave edge
Experiment 1 - “Blind” Search

Label scene using simple depth-first search with backtracking:

Time for the search stage is computed as the number of times that the depth-first-search stack containing all nodes which have been visited but not explored (all the nodes adjacent to them have not been visited) is updated.
Results

Each point represents 100 scenes; scenes are pooled into bins of width 5. Plot shows median time complexity.

Theoretical curve is exponential in number of junctions; best fit curve to the data confirms this.
Experiment #2 - Informed Best-first Search

Label scene using best-first search, exploiting the following heuristics:

• choose the node which has the smallest set of legal labelings

• use structure of domain to guide search. For example...

T-junctions have six possible labelings. Four of them are common the other two appear more seldom. Try the common ones first.

For E-junctions, try first the labeling with the middle-segment labeled as a convex segment and the remaining segments labeled as arrows.
Results

Each point represents 100 scenes; scenes are pooled into bins of width 5. Plot shows median time complexity. Time to label is linear in number of junctions.
Extending the Formalization
- the Time-Varying case...


Given a test image sequence in time $I_t$, is there a sequence of sets $\mathcal{I}_t$ for $t=1$ to $\tau$, where $\mathcal{I}_t$ is the union of all sets $I'_t \subseteq I_t$, such that each element of $\mathcal{I}_t$ satisfies,

$$\sum_{a \in I'_t} \text{diff}(a) \leq \theta_t \quad \text{and} \quad \sum_{a \in I'_t} \text{corr}(a) \geq \phi_t$$

such that $\theta_1 \geq \theta_2 \geq \theta_3 \ldots \geq \theta_\tau$ and $\phi_1 \leq \phi_2 \leq \phi_3 \ldots \delta \phi_\tau$?

**Theorem 3:** Active Unbounded Visual Search Is NP-Complete

**Theorem 4:** Active Bounded Visual Search has time complexity linear in the number of test image pixel locations.
Extending the Formalization
– the Stimulus-Action Behavior case...


Simple stimulus-action human behavior may be conceptualized as:

- localize and recognize a stimulus (with or without a pre-specified target)
- link the stimulus to applicable actions
- decide among all possible actions
- generate actuator commands.

**Theorem 5:** Unbounded Stimulus-Behavior Search is NP-hard.

**Theorem 6:** Bounded Stimulus-Behavior Search has time complexity linear in the number of image locations.

Sensory search problems are integral sub-problems of intelligent behavior...
Extending the Formalization
– Sensor Planning for Object Search...


Formally, find $F \subset O_\Omega$ that satisfies $T(F) \leq K$ and maximizes $P[F]$, where
$F$ is the ordered set of actions applied in the search,
$\Omega$ is a 3D region (union of space elements),
$O_\Omega$ all possible actions that can be applied to $\Omega$,
$T(F)$ is the total time required to apply $F$,
$P[F]$ is the probability of finding the target with $F$,
and $K$ is a constant representing the maximum acceptable time for the search.

**Theorem 7:** Object Search is NP-hard
Summary of Formal Results

Seems that any direction we take leads to the same point:

for large input sets, vision is hard!
Dealing with NP-Completeness  (from Garey & Johnson 1976)

NP-Completeness eliminates the possibility of developing a completely optimal and general algorithm. A direct understanding of the size of the problems of interest and the size of the processing machinery may help in determining the appropriate approximations.

(1) Develop an algorithm that is fast enough for small problems, but that would take too long with larger problems (good if anticipated problems are small).

(2) Develop a fast algorithm that solves a special case of the problem, but does not solve the general problem (assumes special cases have practical importance).

(3) Develop an algorithm that quickly solves a large proportion of the cases that come up in practice, but in the worst case may run for a long time.

(4) For an optimization problem, develop an algorithm which always runs quickly but produces an answer that is not necessarily optimal.

(5) Use natural parameters to guide the search for approximate algorithms. Note: Although many approximation algorithms exist that might have relevance here, none seem to have direct biological plausibility (Tsotsos 1991)
Constraints on a Model (from Tsotsos 1987; 1990)

In the unbounded visual search problem, time complexity is $O[N^2P^M]$ (use this because not all vision has a specific task)

Parameters of this computation are
- $N$ - number of prototypes
- $P$ - size of image in pixels
- $M$ - number of features computed at each pixel

$N$ – 30,000+ - any reduction leads to linear improvements
$P$ – 1,070,000 to 125,000,000 - any reduction leads to exponential improvements
$M$ – 100? - any reduction leads to exponential improvements
Consider the natural parameters of a problem and attempt to reduce the exponential effect of the largest valued parameters.

1. Spatio-Temporally Localized Receptive Fields \( O(NP^{1.5}2^M) \)

2. Separable Feature maps \( O(NP^{1.5}2^M) \)

3. Hierarchy \( O(P^{1.5}2^M\log_2N) \)

4. Pyramids \( O(P^{1.5}2^M\log_2N) \)

5. Attention \( O(P^{1.5}2^M\log_2N) \rightleftharpoons O(1) \)

What is Attention?

Attention is the set of mechanisms that tune the search processes in vision to achieve their best performance to a given task (even if the task is free-viewing)
But....is this still the General Vision Problem?

• NO – the formal results shows that the brain cannot solve the general vision problem, even for its known, fixed input

• The problem is re-shaped to fit into the brain
  – no arbitrary location-feature combinations
  – precise location information is sacrificed
  – limited ability to process in parallel (attention)

• This approximation seems to suffice for everyday behavior; but illusions, errors, inadequacies are the price
Not Bellman’s Curse...

- Bellman refers to the exponential growth of a representational space as a function of its dimensionality.

- Not so obvious that it applies to visual information processing in the brain. The differences seem to be at least three.
  - The input representation is fixed by the size and processing nature of the retina and not by the dimensionality of the visual world or our knowledge of it
  - There is no requirement (nor evidence) that processing be completely veridical nor optimal – it must only suffice for the task at hand
  - Although the general form of the problem indeed has an inherent exponential nature, it is not necessarily the problem the brain is solving.

- To be sure, there is an exponential nature to vision, but it is due to the combinatorics of selecting visual entities that drive human action.
But Can’t Stop Here....

What evidence is there that this is the correct problem formulation ?

First explore the nature of these approximations....
The Problems with Pyramids

Blurring

Context

Crosstalk

Multiple Foci

Convergent Recurrence

Search

© J.K. Tsotsos 2009
Attention *controls* the sub-network that feeds neurons at the top of the visual processing hierarchy.

The problems with pyramids constrain the nature of attentional processes.

Attention controls mechanisms that solve the pyramid problems.

*Done from the inside* not from the outside in order to minimize number of connections and total connection length.
Basic Selective Tuning Paradigm

- **a. Initial Network**
- **b. Top-Down Task Priming**
- **c. Bottom-Up Stimulus Activation**
- **d. Selection and Recurrent Localization**
- **e. Completion of Recurrent Localization**
- **f. Bottom-Up Re-Interpretation**

- **Convergence Binding**
- **Partial Recurrence Binding**
- **Full Recurrence Binding**

- Time
  - 250ms and beyond

- **Iterative Recurrence Binding**

- **stimulus onset at 0ms**
- **from 150 to about 215 ms**
- **about 250ms**
- **about 150ms**
- **-300 to -80ms**
Overall Selective Tuning Model
Lattice of Pyramids
A Feed-Forward Motion Hierarchy

Area 7a
4 x 4 columns
input 4 x 4

MST
5 x 5 columns
input 15 x 15

MT
30 x 30 columns
input 4 x 4

V1
64 x 64 columns
input 8 x 8

Input image
256 x 256
Example:
clockwise rotation

local direction of motion

velocity gradient

V1

MT

MST

© J.K. Tsotsos 2009
Correct motion type can be ‘read’ from this representation but NOT localized
Need for Recurrence Binding
Recurrence Binding for Localization

motion-defined object rotating while receding
3 moving objects, one translating, one rotating and one both translating and rotating

No explicit representation for conjunction of rotation and translation
Predictions
**ST Predictions**  (Tsotsos 1990: Tsotsos et al. 1995; Tsotsos et al. 2001)

- Attention appears wherever there is many-to-one neural convergence  
  O’Connor et al. *Nature Neuroscience* 2002
- Attentional effects appear from higher to lower visual areas  
  Mehta et al. *Cerebral Cortex* 2000
- Attentive surround inhibition (spatial, featural, visuomotor)  
  Tombu & Tsotsos, *Perception & Psychophysics* 2008  
  Loach et al., *Psychological Science*, 2008
- Surround inhibition arises from recurrent processing in visual cortex  
  Böhler et al. *Cerebral Cortex* 2009
- Attentional surround suppression not observed for detection tasks  
- Attention has a natural ‘cycle time’  
  VanRullen et al. *PNAS* 2007
- Attentional response enhancement is a side-effect of receptive field modulation  
  Böhler et al. *Cerebral Cortex* 2009
Attentional effects appear after 150ms

“Higher ventral stream areas showed substantial attention effects at the earliest post-stimulus time points, followed by the lower visual areas V2 and V1. In all areas, attentional modulation lagged the onset of the stimulus-evoked response, and attention effects grew over the time course of the neuronal response. The most powerful, consistent, and earliest attention effects were found to occur in area V4, during the 100-300ms post-stimulus interval. Smaller effects occurred in V2 over the same interval, and the bulk of attention effects in V1 were later.”

from Mehta, Ulbert & Schroeder 2000

see also O’Connor, Fukui, Pinsk, Kastner 2002
Attentional modulation suppresses irrelevant competitors within a receptive field at all levels of processing (Tsotsos 1990, 1999)
Conclusions

• The General Vision Problem – Visual Match is ubiquitous

• The Role of Computational Complexity
  – Formal and Empirical Results – vision is hard!
  – Constraints on a Sufficient Architecture – 2 approximations + attention suffice (not necessary nor unique)
  – Implications – set of constraints on search and information flow

• The Selective Tuning Model of Attention and Vision
  – attention controls how the visual cortex is used in any given task

• Predictions
  – many have significant experimental support

• But this is not Bellman’s curse....
Additional References