Perception of structure: Anyone Interested?

(Aaron Sloman, CoSy Project, University of Birmingham)

Most of the work being done on machine vision is concerned with:

- Recognition/classification/tracking of objects (including face recognition).
- Optical character recognition (special case of the previous point)
- Building a representation of a 3-D scene from which views from different viewpoints can be projected.
- Self-localisation and route learning/route following.
- Pushing things, avoiding things, blocking things (as in robot football).
- Various special-purpose applications of the above, e.g. floor cleaning and lawn mowing.

What is missing from all that impressive digital image processing?

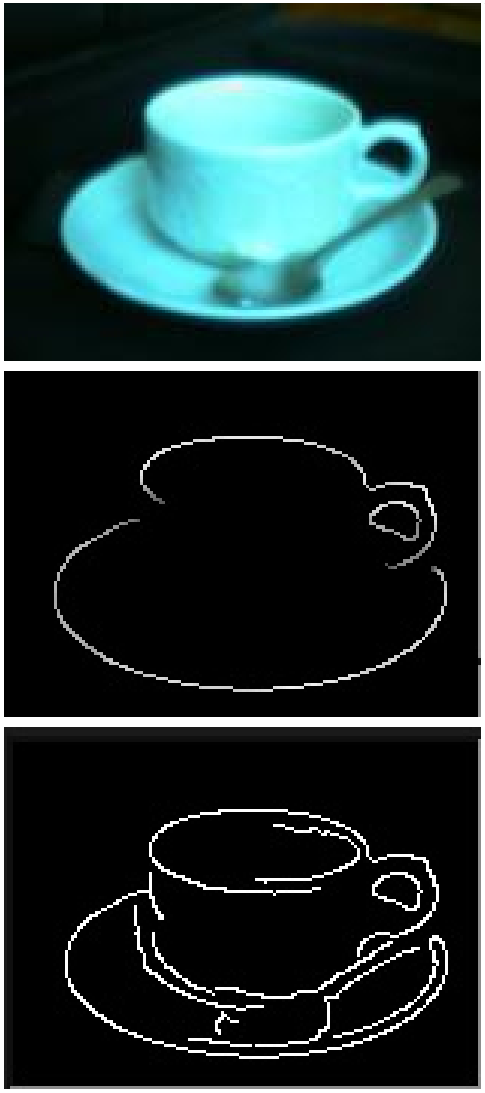
- Perception of structure (at different levels), e.g.
  - perception of 3-D parts and surface fragments and their features and relationships
  - Perception of motion in which relationships between parts of one object and parts of other objects change, including things like sliding along, fitting together, pushing, twisting, bending, straightening, inserting, removing, rearranging.
- Perception of positive and negative affordances and causal relations, e.g.
  - Possibilities for action, for achieving specific effects
  - Obstructions to action, and limitations of actions

  especially as regards parts of complex objects, which can be grasped, pulled, pushed, twisted, rotated, squeezed, stroked, prodded, thrown, caught, chewed, sucked, put on (as clothing or covering), removed, assembled, disassembled, and many more...; as well as many variations of each of the above.

Some challenges

Things for us to think about if we really wish to achieve the major goals of AI and Cognitive Science:

VISUAL INPUT WE SHOULD BE ABLE TO COPE WITH (FOR A MACHINE THAT CAN MANIPULATE OBJECTS):



We need to be able to see things we cannot easily express in language.

- This bit is concave. This bit is shiny. This bit is a reflection of the cup. How is 'this bit' identified?
  (Compare 'virtual finger theories' – Z.Pylyshyn's FINST ?)
- As I move left and right those reflections move on that edge of the cup. How are 'those' and 'that' identified?
- This is how I should move my hand to grasp the spoon. A different kind of 'this' (manner, route, method) – how identified? Perhaps partially instantiated parameters in some operator?
- We need to understand ontologies and representations for active (possibly pre-linguistic) agents
- What details are 'visible' in an image depends on how it is processed (Top picture, processed different ways gives middle or bottom picture). Can the system use intelligent decision making about how to process details in different ways in different places?
  (One of many kinds of focusing of attention.)
- layered interpretations (image, low level image structures, silhouettes (2-d), parsed silhouettes, 3-d structures, affordances, many context-dependent features and relations.
- Attention is different in different parts of the system: different things are selected – image features, objects, object features, locations in the image, locations in the world (relative to: room, table, object, object part...), relations, functions, actions, ways of doing things, routes, other agents, social interactions, kinds of material, properties (e.g. rigid, flexible) ....

Towards the future...

We must not forget concurrently active processing with which vision can be involved (mostly totally unconscious):

- Reactive (including innate and trained reflexes and more complex internal reactive processes)
- Deliberative (capable of detachment from what exists, using structures, relationships, compositional semantics, reference to hypothetical futures (predictions/plans), pasts (explanations), unseen things (what made that noise? what's behind me? etc.)

  `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0604`
- Reflective-meta-management:
  - needs meta-semantic capabilities (reference to things that process information, their states, actions, goals, etc.)
  - both inward and outward directed
  - inward directed meta-semantic processing requires architectural support
- Need for layered action and perception systems.

  Processing intermediate perceptual layers explicitly supports 'hierarchical synthesis' (Neisser and others), and 'low commitment' intermediate inferences (Marr), structure-sharing between hypotheses.

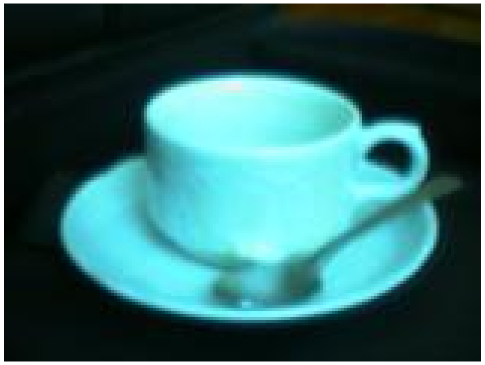
  Intermediate and high level actions can be initiated in different parts of the architecture – e.g. innate and trained reflexes in the reactive subsystems, plan execution in the deliberative system, expressing self assessments in the meta-management system (anger, submissiveness, contriteness, ...)

  As in the CogAff schema and H-Cogaff architecture
  `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk3`

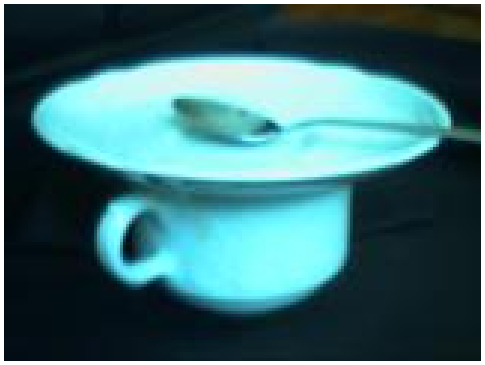
Some tasks for a crow-challenging robot?

THIS IS NOT QUITE THE BLOCKS WORLD

Using a two-finger gripper, what actions can get from this:



to this:



and back again?

Or with saucer upside down?

Consider how, prior to the action, the agent has to

- identify parts of objects, or parts of parts, e.g. the edge of the handle, or the far edge of the handle or a certain portion of the edge of the saucer
- identify possible actions: grasping this thing here from this direction

  Could such deliberative premeditation use the action schema (operator) with approximate, qualitative parameters instead of the more definite actual parameters that would be used if the action were performed?
- think about various effects of actions, including changing effects of continuous processes

NOTE: there are problems here partly analogous to problems of reference and identification in language, except that the mode of reference is not linguistic and what is referred to typically cannot be expressed in language because it is anchored in non-shared structures and processes.
(Internal ‘attention’ processes are partly like external pointing processes: virtual fingers.)

TO BE CONTINUED

Conjecture: (added 6 Feb 2005)

Solving these problems will require revival and extension of the ideas put forward in this generally forgotten paper:

H.G. Barrow and J.M. Tenenbaum,
'Recovering intrinsic scene characteristics from images',
in *Computer Vision Systems*,
Eds. A. Hanson and E. Riseman,
Academic Press, New York, 1978

Mention of a crow-challenging robot is a reference to Betty,
the hook-making New Caledonian crow
`http://users.ox.ac.uk/~kgroup/tools/tools_main.shtml`
`http://news.nationalgeographic.com/news/2002/08/0808_020808_crow.html`

More Challenges for Vision Researchers

Seeing a toy (meccano) crane

`http://www.cs.bham.ac.uk/research/projects/cosy/photos/crane/`

Seeing impossible objects

`http://www.cs.bham.ac.uk/research/projects/cogaff/challenge-penrose.pdf`