An Analysis of the Required Capabilities of Cognitive Vision

Version 1.0

1. Introduction

The 4th ECVision Six-Monthly Meeting on the 26th March 2004 was devoted to the further development of the research roadmap. One objective of the meeting was to characterize exactly what is meant by cognitive vision. After some discussion, it was agreed that one way to do this is by reference to existing biological models, such as human perception. To make this concrete, four break-out groups were formed to identify the principle capabilities that a cognitive vision system should exhibit, taking a reference model as an exemplar. The four reference models that were chosen were:

1. A surveillance system
2. A home assistant
3. Young infants and children
4. An autonomous vehicle

As well as the requisite capabilities, each group was asked to identify the scientific and technological challenges to be overcome in achieving these capabilities.

The goal of this exercise was to produce a variety of characterizations. Ideally, these would overlap considerably so that it would be possible to subsequently produce a list of generic capabilities that are not specific to the scenario associated with the reference model (*e.g.* threat detection, cleaning, gesturing, or driving).

The following four sections document the findings of the break-out groups. This is followed by a draft of the generic capabilities abstracted from these four sections.

2. The Cognitive Vision Capabilities of a Surveillance System

This break-out group took a single scenario – tracking a person – as their point of reference and then identified a range of capabilities, sorted in order of increasing levels of complexity and sophistication. These are:

a) 2-D fronto-parallel tracking with spatio-temporal continuity (*i.e.* inter-frame object persistence);
b) The ability to deal with changes in scale;
c) Theability to cope with occlusions (*i.e.* object persistence through occlusion);
d) Multi-camera configurations with inter-camera integration and 3-D tracking with object persistence through occlusions.
e) Representation of small number of classes of objects;

f) Representation of categories of behaviours;
g) The ability to learn objects classes and behaviours;
h) The ability to generalize (*i.e.* to form new categories);
i) Portability to new contexts (within the same scenario);
j) Portability to new applications scenarios (*e.g.* the ability to reconfigure a train-station surveillance system for an airport).

3. The Cognitive Vision Capabilities of a Home Assistant

The second group nominated a home-assistant as their reference model and identified two proto-typical scenarios to illustrate the required capabilities. The first is goal invocation, in which one addresses the home assistant and instructs it, for example, to 'go to the meeting room, pick up the empty bottles, and put them in this box'. The second scenario, guided enumeration, is a version of 'show-and-tell' wherein the assistant takes you on a tour of the house and indicates the name and purpose of household items.

The home assistant, therefore, would have the capability to do the following:

a) Form mental maps of the world and associate labels with objects;
b) Explore the environment;
c) 'Show and tell': identify objects and understand their purpose and function;
d) Detect novelty and change in the environment (including introduction of new objects, removal of objects, and alteration of object location);
e) Interpret the actions of humans.

The challenges these capabilities pose are many and include:

- Semantic labeling and localization;
- Categorization of objects and behaviours;
- Formation of motor maps and spatial maps;
- The ability to deal with inconsistent representations without explicitly resolving the inconsistencies;
- Functional categorization (related to the concept of affordance: what in the room is 'throwable' as a weapon, for example);
- Understand a subset of human activities;
- The ability to verbalize knowledge and to describe the content of representations.

4. The Cognitive Vision Capabilities of Young Infants and Children

The third break-out group took young infants and children as their reference model for cognition and cognitive vision. They then identified several abilities that are typical of their cognitive and perceptual faculties. These include the ability to:

a) Recognize expressions;
b) Comprehend object persistence in space and time, and in the presence of visual occlusion;
c) Point to and gesture at specific objects;
d) Develop hand-eye coordination and grasp for objects;
e) Develop an understanding of the structure of one's local space;
f) Imitate the actions of others;
g) Use vision to enable self-locomotion;
h) Understand naïve physics;
i) Formulate hypotheses and reason visually;
j) Engage in counter-factual reasoning (whereby one can reason as if one is someone else); this implies the ability to (a) model someone else's model of the environment, and (b) deliberate on that instantiated model, rather than on one's one innate model).

These capabilities then give rise to a series of challenges. In order of increasing difficulty, these are:

- The development of a software and hardware platform to investigate cognitive vision;
- The need to model object invariance and persistence in a spatio-temporal context;
- The ability to reason from different points of view;
- The ability to develop awareness of one's body (typically through a correlation of proprioceptive information and visual information).

5. The Cognitive Vision Capabilities of an Autonomous Automobile

The fourth group decided to catalogue a list of visual behaviours with autonomous automobiles as their reference model. They identified three environments that exhibit increasing levels of difficulty for visual interpretation. These are:

- *Highways*: these are highly structured and can be controlled in the sense that there are a strictly limited number of visual scenarios and object behaviours;
- *Cities streets*: these are poorly structured and are not easy to control in the sense that the visual environment will vary a great deal and there will be many objects, each of which may exhibit entirely unexpected behaviours
- *Off-road*: in theory, this should be the easiest scenario because the behaviour of the autonomous vehicle is not greatly constrained. On the other hand, the recent DARPA Grand Challenge competition shows that this scenario is far from trivial.

The group then identified the following capabilities that would be expected in a cognitive vision system:

a) Detection of driveable free-space, dealing with:
   1. Static obstacles;
   2. Moving obstacles (*e.g.* pedestrians, bicycles, animals);
   3. Other vehicles.

b) Detection of legal driveable free-space; this implies the need to understands rules and laws;

c) Situation assessment, dealing with
   1. The position and velocity of the vehicle with respect to free-space;
   2. The relative position and velocity of obstacles and other vehicles;
   3. Models of behaviour (*e.g.* actions based on right of way) and prediction of behaviour;
   4. Models of intent (*e.g.* acceleration on approach to an amber traffic lights suggests an intent not to stop and give way);
   5. Legal situation assessment.

In turn, these capabilities give rise to the following challenges:

- Achievement of robust operation in highly variable environmental conditions (*e.g.* illumination and precipitation);
- Reliable and safe operation: *e.g.* estimating the probability of a crash given a dangerous situation;
- Ability to exploit strong temporal constraints.

The group also identified several metrics that should be used to assess a cognitive vision system in the context of the autonomous vehicle reference model. These are:

- Mean Time Between Failure;
- Speed of operation;
- Variability of environmental conditions;
- Ability to pass a driving test;
- Maximum/minimum travel time, speed, comfort, risk, fuel consumption.

6. Common Cognitive Vision Capabilities

This section sets out a first draft of the generic capabilities of a cognitive vision system (*i.e.* capabilities that are not specific to a given scenario).

A complete cognitive vision system will have the ability to:

1. Track objects / visual entities / points of interest in 3-D along both short-term and long-term persistent (non-discontinuous spatio-temporal) paths through partial and complete occlusion, integrating multiple viewpoints acquired from one or more possibly-moving cameras.

2. Classify or label objects / visual entities / points of interest into one (or more) of a small number of classes, optionally with the use of *a priori* constraints.

3. Categorize behaviours (*i.e.* temporally-extended spatio-temporal configurations of visual entities); characterize these behaviours, *e.g.* in terms of position, orientation, velocity.

4. Learn classes of objects / visual entities / points of interest and categories of behaviours.

5. Generalize classes and categories (*i.e.* form new classes and categories from old ones).

6. Recognize and adapt to novel variations in the current visual environmental context.

7. Generalize to new contexts (*i.e.* to new application scenarios); this ability is probably implied by the learning and generalization abilities 4. and 5. above.

8. Explore or investigate the visual environment (as a forcing mechanism for learning and generalization capabilities).

9. Form maps of the visual environment using single or multiple eco- or ego-centric frames of reference, denoting the location of labeled objects / visual entities / points of interest.

10. Interpret the intent underlying behaviour, specifically to predict future spatio-temporal configurations of the visual environment, across a variety of time-scales, optionally with the use of *a priori* constraints.

11. Categorize functionality on the basis of appearance of an object / visual entity (*i.e.* map from classes of objects to categories of behaviour).

12. Communicate, either by gesture or verbally, an understanding of the environment to other systems, including humans.

13. Develop hand-eye coordination and an ability to grasp objects.

14. Imitate the actions of other agents.

15. Develop an understanding of naïve physics.

16. Formulate hypotheses about spatio-temporal configurations and deliberate about them visually.

17. Engage in counter-factual deliberation (whereby one can deliberate as if one is someone else); this implies the ability to (a) model someone else's model of the environment, and (b) deliberate on that instantiated model, rather than on one's one innate model).

These capabilities give rise directly to some difficult challenges, including the need to:

1. Develop a software and hardware platform to investigate cognitive vision;

2. Create a theoretical model for each capability; validate & evaluate that model empirically (the list of capabilities can be prioritized if necessary).

3. Create a theoretical meta-model whereby individual models can be integrated into a coherent framework.

4. Facilitate multiple inconsistent, time-varying, incomplete instantiations of each model in that framework.

5. Achieve robust operation in highly variable environmental conditions.

6. Develop an awareness of one's body and its context in the visual environment.