A Research Roadmap of Cognitive Vision

DRAFT Version 3.2

David Vernon (Ed.)
ECVision: The European Research Network for
Cognitive Computer Vision Systems

website: www.ecvision.org
email: coordinator@ecvision.org

September 6, 2004

Contents

Preface

On the 1st March 2002, a European research network for cognitive computer vision systems — *ECVision* — was inaugurated. This network is funded for three years by the European Commission under the Information Society Technologies (IST) programme as project IST-2001-35454.

The goal of *ECVision* is to promote research, education, and application systems engineering in cognitive computer vision in Europe. It pursues this goal by a variety of means, most of which are based on facilitating peer-to-peer interaction amongst the foremost researchers in the area. The network targets four main activities: research planning, education & training, information dissemination, and industrial liaison. In the research planning area, one of the main goals is to maximize the effectiveness of future research by creating a detailed research agenda: a research roadmap.

Research roadmaps are often conceived as a statement of the state-of-the-art in a given discipline accompanied by a prioritized list of problematic issues and a strategy for their investigation. The state-of-the-art survey constitutes a point of departure on the roadmap, the problematic issues constitute the destination, and the strategy the path by which one should go proceed to the destination. However, there isn't unequivocal consensus on the right approach to take in addressing the problems posed by cognitive vision. Consequently, the approach adopted here is one of inclusiveness: we don't assume that there is a single point of departure, a single destination, and a clearly-mapped path leading from one to the other. Instead, we allow that there are several different destinations (*i.e.* types of cognitive vision system) and that there are multiple paths by which these destinations can be achieved. Indeed, depending on how you choose to view or define cognitive vision, there are many points of departure, some based squarely in artificial intelligence and image processing, others in developmental psychology and cognitive neuroscience, and others yet in cognitive robotics and autonomous systems theory. Our goal in this document is to avoid disenfranchising any particular community (because we don't know which one will ultimately be successful) and, instead, simply to state what the options might be, to suggest ways in which each paradigm might best proceed, and in the process to identify common pathways, the investigation of which can benefit all concerned. The goal, then, is to formulate a solid scientific research agenda, defined within a framework that is relevant to all, which doesn't compromise on the need to address the very real diversity of possible approaches to the creation of cognitive vision systems.

The roadmap is not, however, an amalgam of disparate views. Notwithstanding that there are indeed different perspectives, a number of unifying themes emerged when formulating the roadmap so that the final picture is actually far more cohesive than we had hoped it might at the outset.

D.V. 5-9-2004

Acknowledgements

This research roadmap is the product of several generations of evolution and many people contributed directly and indirectly to its creation over the past couple of years.

The initial versions were collated and edited by James Crowley, based on an extensive process of consultation with the members of *ECVision*[1].

Colette Maloney organized a European Commission workshop in June 2003 to formulate a support document for the forthcoming call for proposals on cognitive systems [2]. This workshop was attended by Henrik Christensen, James Crowley, Mark Keane, Kevin O'Regan, Aaron Sloman, Hans Uszoreit, and David Vernon. The discussions at that workshop first flagged the need to look more broadly at the issues of cognition and the implications for vision.

Hans-Hellmut Nagel and Henrik Christensen organized a Dagstughl seminar [3] on cognitive vision systems in October 2003. This seminar was pivotal in re-shaping our thinking, alerting us to the necessity of keeping open a broad research agenda without losing the focus we originally had on well-engineered computer vision system.

An *ECVision* workshop in March 2004 and a follow-up meeting in May 2004 provided the input and impetus to create the first draft of this particular roadmap. Ales Leonardis and Horst Bischof deserve special mention for their suggestion on how to structure the discussion on the nature and scope of cognitive vision (see Section 4).

On-going discussions with Giulio Sandini were very helpful in sifting out the key issues and showing a way to actually achieve an inclusive and realistic research agenda that is, hopefully, relevant to all.

Apart from those mentioned already, many other people contributed to this research roadmap. They include: Peter Auer, Aude Billard, Isabelle Bloch, Pia Boettcher, Heinrich Bueltoff, Hilary Buxton, Tony Cohn, Patrick Courtney, Andrew Crookell, Sven Dickinson, Christof Eberst, Jan-Olof Eklundh, Bob Fisher, Wolfgang Förstner, John Gilby, Goesta Granlund, Vaclav Hlavac, Josef Kittler, Walter Kropatsch, Jim Little, Giorgio Metta, Hans-Hellmut Nagel, Bernhard Nebel, Bernd Neumann, Heinrich Niemann, Lucas Paletta, Fiora Pirri, Gerhard Saegerer, Bernt Schiele, Rebecca Simpson, Gerald Sommer, John Tsotsos, Monique Thonnat, Markus Vincze. Sincere apologies to anyone that has been inadvertently omitted.

This work was made possible by the European Commission through funding under the Information Society Technologies (IST) programme, Project IST-2001-35454: *ECVision*—European Research Network for Cognitive Vision Systems. Special thanks go to Colette Maloney for her unstinting and far-sighted support and encouragement over the past three years.

Executive Summary

Computer vision is an important and maturing engineering science. However, despite significant success in some application areas, contemporary computer vision is still a relatively brittle technology.

The term *cognitive vision* has been introduced to encapsulate an attempt to achieve more robust, resilient, and adaptable computer vision systems by endowing them with a cognitive faculty: the ability to learn, adapt, weigh alternative solutions, develop new strategies for analysis and interpretation, generalize to new contexts and application domains, and communicate with other systems, including humans.

Cognitive vision is an emerging discipline in a pre-paradigmatic state. There are several competing perspectives on the nature of cognition and the scientific approach that will be ultimately successful in achieving its ambitious aims is not yet known.

This document presents a research roadmap of cognitive vision: it defines the discipline in a model-neutral manner and provides a inclusive strategy whereby we can move beyond the pre-paradigmatic position we are now in, to develop cognitive vision — the science of visually-enabled cognitive systems — and create a discipline with well-understood aims, sound constitutive theories, and exploitable technologies.

Cognitive vision is cast here in a 3-D space comprising three orthogonal considerations: scientific techniques, functional capabilities, and instantiated competences. The scientific techniques include *visual sensing, architecture, representation, memory, learning, recognition, deliberation, planning, communication, and action.*

The functional capabilities include *detection and localization, tracking, classification and categorization, prediction, concept formation and visualization, inter-agent communication and expression, embodied exploration, and hand-eye coordination,* though not all need necessarily feature in the same system.

The instantiated competences represent the application-specific functionality based on capabilities but developed over time through experience, through learning, interaction, and practice in task-specific circumstances.

Because of the inter-dependencies between the scientific techniques and the functional capabilities, research in cognitive vision should be carried out in the context of specific major challenges. These are concerned with *the advancement of methods for continuous learning, the establishment of minimal architecture(s), goal achievement, generalization, utilization of systems engineering methodologies, development of complete systems with well-defined competences, and the creation of research tools.*

The involvement of industrial interests is crucial to the development of the area, to provide focus on potential applications, to identify essential functionality, and to seek opportunities for timely commercial exploitation as the discipline develops.

This research roadmap represents a 20 year plan, not a 5 year plan. However, successful results can be achieved in the short-term, probably in the priority challenges of Learning, Architecture, Tools, and Systems Engineering. Since cognitive vision is a systems science, and every technique, capability, and challenge is inter-related, it is essential that the entire roadmap be followed, not just isolated parts of it.

1 Introduction

1.1 The Emerging Discipline of Cognitive Vision

Computer vision is an important and maturing engineering science. It underpins an increasing variety of applications that require the acquisition, analysis, and interpretation of visual information. However, despite recent success in such areas as computational projective geometry [4, 5] and appearance-based recognition [6], contemporary computer vision is still a relatively brittle technology. Consequently, its successful exploitation has been limited to relatively narrow application domains such as machine vision for industrial inspection, the analysis of video data for remote monitoring, and the creation of special effects in the film industry. The focus of much recent research has been on finding ways to reduce this brittleness.

The term *cognitive vision* has been introduced in the past few of years to encapsulate an attempt to achieve more robust, resilient, and adaptable computer vision systems by endowing them with a cognitive faculty: the ability to learn, adapt, weigh alternative solutions, and even the ability to develop new strategies for analysis and interpretation.

The key characteristic of a cognitive vision system is its capacity to exhibit robust performance even in circumstances that were not forseen when it was designed [7]. Furthermore, a cognitive vision system should be able to anticipate events and adapt its operation accordingly. Ideally, a cognitive vision system should be able to recognize and adapt to novel variations in the current visual environment, generalize to new contexts and application domains, interpret the intent of underlying behaviour to predict future configurations of the visual environment, and communicate an understanding of the environment to other systems, including humans.

The following is a working definition of the area:

> A cognitive vision system can achieve the four levels of generic computer vision functionality of detection, localization, recognition, and understanding.[1] It can engage in purposive goal-directed behaviour, adapting to unforeseen changes of the visual environment, and it can anticipate the occurrence of objects or events.
>
> It achieves these capabilities through learning semantic knowledge (*i.e.* contextualized understanding of form, function, and behaviour); through the retention of knowledge about the environment, about itself, and about its relationship with the environment; and through deliberation about objects and events in the environment (including the cognitive system itself) [8, 9].

Cognitive vision is in essence a combination of computer vision and cognition.[2] Consequently, the study of cognitive vision requires both the study of computer vision and cognition. Unfortunately, this is where the trouble begins. The term cognition has several interpretations, each of which is dependent on very disparate underlying models of fundamentally different positions on the nature of cognition. Neither is the situation fixed and static, with the discipline of cognitive science is itself going through something of a metamorphosis [11].

[1] By understanding, we mean the ability to comprehend the role, context, and purpose of a recognized entity and its categorization into meta-level class on some basis other than visual appearance alone.

[2] See [10] for an alternative viewpoint on the possible nature of cognitive vision.

Quite apart from the disparate perspectives on cognition, there is also the question of the nature of the combination of vision and cognition. Some view cognitive vision as a cognitively-enabled vision, others view it as a visually-enabled cognition. For reasons that will become clear later, we adopt the latter view in this roadmap.

The foregoing definition of cognitive vision above does not necessarily imply that cognitive vision systems are human-like in structure, organization, or operation. It implies only a requirement for robust, adaptive, anticipatory behaviour. It is not necessary that the cognitive vision system be based on a model of human cognition. On the other hand, there are two good reasons why we would want a cognitive vision system to be based in some respects at least on human cognition. First, humans offer an existence proof that cognitive vision is possible. Second, the applications of some cognitive vision systems will have to interact directly with humans and therefore it seems reasonable that they share some form of common understanding of their respective worlds. We will return to this issue again in a later section when we consider the need for embodiment in cognitive vision systems.

1.2 The Pre-Paradigmatic Status of Cognitive Vision

It should be clear from the previous section that cognitive vision is an emerging discipline. This is a point that needs some emphasis because it isn't a case of cognitive vision being well-established and simply needing more time to grow and develop. On the contrary, the discipline is in a pre-paradigmatic state where the scientific approach that will ultimately be successful in achieving the ambitious aims is not yet known. There are two complementary issues at stake here:

1. the definition of the discipline itself: *what cognitive vision is*;
2. the scientific foundations upon which a discipline will be built: *how cognitive vision can be effected.*

Once these issues have been fully addressed, we will have moved beyond the pre-paradigmatic state to a position where the scientific community at large can work together to develop the discipline.

However, we are not yet at that point and this roadmap has a pivotal role to play in getting there. We have already made a good start on the first issue — the definition of the discipline — in the previous section and we will return to consider it in much greater detail in Section 4. We address the second issue — the scientific foundations — in Section 5 with a survey of the possible approaches that can be adopted in treating cognitive systems in general and visually-enabled cognitive systems in particular. Each of the approaches has different strengths and weaknesses and they are all at different states of scientific maturity. Our goal in this document is, having carefully defined the discipline of cognitive vision in a model-neutral manner, and having surveyed the scientific options and competing positions, to present a clear roadmap by which we can move forward effectively beyond the pre-paradigmatic position we are now in, to reach a point where there is a solid body of scientific knowledge which can be used to address the complex applications that motivate the development of cognitive vision in the first place.

1.3 The Roadmap Objectives

With the essential goal of the *ECVision* research roadmap now established, we can now summarize the constituent objectives. They are:

1. To provide a definition of cognitive vision that is neutral with respect to possible approaches, to set out what capabilities should be exhibited by such a system, and to explain what task-specific abilities these will enable;

2. To identify competing approaches and to characterize their similarities and differences, flagging research issues that are common to each approach;

3. To create an inclusive research agenda that is relevant to all of the cognitive vision research community, one that is maximally effective in making scientific progress but without unnecessarily biasing the direction of advancement;

4. To identify critical gaps in our scientific understanding and/or technical know-how;

5. To propose a strategy for research that will allow these gaps to be filled;

6. To highlight contentious and significant issues (*e.g.* the necessity for embodiment, the nature and need for representations, the nature and role of knowledge, the role of language, the inter-dependence of perception and action) that require special attention.

In the following, Section 4 addresses the further definition of cognitive vision, Section 5 surveys the possible approaches, while Section 6 addresses the core scientific problems to be solved. Finally, Section 7 sets out the *ECVision* research roadmap proper. Before embarking on these issues, we will first take a small detour to consider what obstacles may lie in the way of progress and suggest reasons why they can be surmounted.

2 Obstacles to Progress

2.1 Unchartered Territory: The Relationship between Computer Vision and Cognitive Vision

The first obstacle in our path is that we are treading unchartered territory. Although we know where we want to go, we don't know for certain how to get there and it's not even clear from what position we should depart. The apparently obvious candidate departure point is state-of-the-art traditional computer vision, and the equally obvious direction in which to travel is towards the domain of cognitive science and artificial intelligence. This leads us naturally to view cognitive vision as a point on a spectrum of theories, models, and techniques with computer vision at one end and cognitive systems at the other (see figure 1). The question then is where are the boundaries between computer vision and cognitive vision, and between cognitive vision and cognitive systems? Unfortunately, it is extremely difficult to place any such boundaries. As we saw in our definition of cognitive vision above, the standard computer vision functionality is an important component of cognitive vision but it is not obvious how the cognitive processing can be simply grafted on. It would seem then that such a conceptual view of cognitive vision is not helpful.

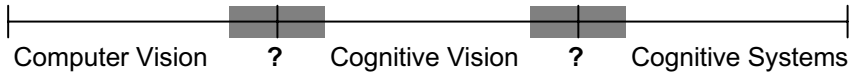

Figure 1: Cognitive Vision as a part of the Computer Vision / Cognitive Systems spectrum.

An alternative is to consider the more general space of cognitive systems embracing all perceptual modalities and to view cognitive vision as the projection of that general space onto the sub-space of cognitive vision systems (see figure 2). This projection leaves in those capabilities that are necessary to cognitive vision but excludes those that are not necessary or are less relevant (such as pure reasoning, for example). This viewpoint suggests then that a cognitive vision system can best be viewed as a *visually-enabled cognitive system*, rather than a cognitively-enabled vision system.

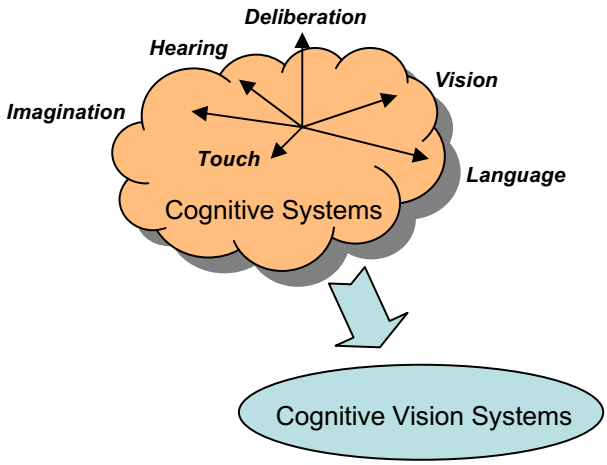

Figure 2: Cognitive Vision as projection of the space of Cognitive Systems.

Such a viewpoint extends naturally to other perceptual faculties such as hearing, touch, smell, and modes of interaction such as speech, gesture, manipulation, and exploration. It also re-directs our attention to the core attributes of cognitive systems: robust performance even in circumstances that were not foreseen when it was designed, the ability to learn, adapt, weigh alternative solutions, and develop new strategies for analysis and interpretation. The challenge then becomes one of trying to see how we can build a theory that produces systems with these attributes in the context of visual perception. This does not diminish the critical significance of advanced traditional computer vision nor does it relegate it to a position of lesser relative importance. It simply places it in the context

of systems-oriented research in cognition. The difficult task that faces us in this roadmap is to see how to project from the general space of cognitive systems to visually-enabled cognitive systems: to decide what issues in cognitive science we can leave out and what issues in computer vision it is essential to include.

2.2 The Multidisciplinary Nature of Cognitive Vision

The switch in perspective on cognitive vision from cognitively-enabled vision to visually-enabled cognition serves to highlight the second obstacle to progress: the multidisciplinary nature of cognitive vision. To study cognitive vision in depth and in all its guises, one must tackle a huge array of disciplines: computer vision, pattern recognition, artificial intelligence, cognitive science, perceptual psychology, developmental psychology, cognitive neuroscience, neurophysiology, cognitive robotics, semiotics, epistemology, systems sciences, cybernetics, autonomous systems theory, and probably others too. Add to this list the several branches of specialized mathematics that underpin many of these areas and you begin to see the breadth of the area. Of course, the focal point is still computer vision, but computer vision in a very multidisciplinary context. It is a rare polymath that would be fully conversant with all of these topics. And yet, because of the pre-paradigmatic nature of cognitive vision, and the consequent need to set the research agenda as broadly as possible, there is really no alternative but to attempt to address all these issues. It would be fundamentally wrong-headed to confine one's attention to, say, a postulated theory that was based exclusively on a neuroscientific embodiment of state-of-the-art computer vision. At the moment, we simply don't know enough yet about cognition and cognitive vision to confine research solely to this, or any other, potentially restrictive approach.

2.3 Achieving Definitive Knowledge: Exploiting The Scientific Method

One of the features that characterizes young disciplines is the difficulty in creating a body of knowledge that can be shared and built upon by others in the field. This was the case in early days of computer vision and is evident today in cognitive vision. Whilst there is now a solid base of both theoretical knowledge, empirical know-how, and supporting technology in some fields of computer vision, much work remains to achieve this in cognitive vision. This again is a natural consequence of the pre-paradigmatic status and the multidisciplinary nature of the discipline. However, until we have such a framework, long-term stable progress will be unlikely, and quick progress will be impossible. The aim must be to set out the theoretical foundations of each of the contributing paradigms in a way that leaves them open to formal scrutiny, replication, and evaluation by others. This almost inevitably implies a mathematical framework. Without the exactitude of mathematical exposition, it will be extremely difficult to create a framework of results that can be built upon by all in the certainty that the foundations are solid. It is of course a great challenge to identify and deploy the right type of mathematics to facilitate this. This theoretical framework must be accompanied by the creation of transferrable technology (*i.e.* algorithms, hardware, software) and support tools (*i.e.* development environments, languages, benchmarking test suites).

3 The Changing Scientific Context: Overcoming Obstacles

3.1 External Influences on Cognitive Vision

The previous section sketched a somewhat gloomy picture, focussing on all the reasons why making progress in cognitive vision is going to be hard: the difficulty in formulating the problem in the first place, the daunting multi-disciplinary nature of the area, the challenge of ensuring that the results are developed in a formal — repeatable and reusable — fashion. But it is not all bad news. This research roadmap goes a some way towards tackling the first difficulty and, whilst it won't solve any specific scientific problems, it will at least tell us what the problems are and how we might approach them.

The difficulty posed by the many facets of the discipline is actually less of a problem and more of an opportunity. Science has succeeded well in the past because problems were constrained to limited boundaries to ensure that the issues then allowed clear formulation and tractable solution. As science expands, it is increasingly having to relax these constraints and admit many more factors that before. Multi-disciplinary research and cross-disciplinary collaboration is becoming the new *modus operandi* in many branches of science today. The multi-disciplinary nature of cognitive vision mirrors exactly this general trend. Truly multi-disciplinary research institutes such as COGS – The Centre for Research in Cognitive Sciences at the University of Sussex[3], CALD – The Center for Automated Learning and Discovery in Carnegie Mellon University[4], and the Santa Fe Institute[5], which used to be viewed as exceptional in their outlook, are increasingly being looked at as models of how future research should be organized. Multi-disciplinary research is not simple, however. Apart from the problems associated with the disparate pre-dispositions of constituent disciplines and the need to achieve a shared understanding and language to facilitate a common research agenda, there is also the difficulty presented by having to integrate the ideas from both hard and soft sciences without losing the quantitative focus and the formal framework of mathematical modelling that are required for any computational emulation of a visually-enabled cognitive system.

3.2 Enabling Technologies

It is perhaps a little premature to cite enabling technologies for an emerging science that has not yet developed a universal and widely-accepted model. However, three things are sure.

First, the emulation of cognitive systems will consume as much computational power and processor memory as we can supply. Moore's Law is still in effect and today's microprocessors deliver sufficient power in small footprint machines to facilitate real-time or near real-time processing of video data in embedded applications. Of course, it isn't clear yet what will be the computational requirements of cognitive processing but, unless the algorithmic complexity of successful techniques is high-order polynomial or exponential, there is reason to be hopeful that at least some level of realistic cognition can be emulated.

Second, cognitive vision systems will require small, high-quality, robust imaging devices and high-bandwidth image acquisition systems. The past five years or so has seen

[3] http://www.informatics.sussex.ac.uk/cogs/
[4] http://www.cald.cs.cmu.edu/
[5] http://www.santafe.edu/

considerable advances in camera technology, in particular with the current generation of high-resolution colour CCD cameras with high-speed digital interfaces such as USB-2 and IEEE 1394 (i-Link or Firewire). These serial-bus cameras, with bandwidths of more than 400 M bits per second, can easily be connected in a multi-camera configuration, with synchronous triggering if necessary, and they interface directly with contemporary computers without the need for expensive image acquisition equipment (for example see `http://www.ptgrey.com/`).

Third, cognitive vision systems will be large software-intensive systems. Whilst software engineering methodologies have not advanced as much as the industry would like, we are better equipped today to build re-usable component-based multi-functional software systems than at any time in the past forty years. That said, there is still a strong need to create cognitive vision development environment that can be used to enable the effective construction of complete visually-enabled cognitive systems, to facilitate the sharing of algorithms and software between researchers, and to help accelerate subsequent commercial exploitation. We will return to this issue again in Section 7.

3.3 Advances in Related Disciplines

As we have seen, cognitive vision involves many related disciplines and it would be unrealistic to attempt a comprehensive survey of all advances in these areas. That said, it is also important to understand that the emerging discipline of cognitive vision is developing in the context of other disciplines which are themselves evolving and changing. New insights will come not only from the cognitive vision community but also from the broader multi-disciplinary community. We note here just two examples of how advances in related disciplines will have an impact on our own efforts to create the solid foundations of visually-enabled cognitive systems. The disciplines involved are cognitive science, neuroscience, and epigenetic robotics.

In the last 10 years or so, an ever growing number of cognitive scientists [12, 13] have begun to appreciate the possibility of instantiating cognitive models in robotic systems. The space of research spanned is quite wide, starting from the locomotion and organizational behaviors of insects and early vertebrates [14, 15] through models of high order cognitive skills in humans such as social behaviors [16], imitation [17, 18, 19], communication, and language [18, 20, 21, 22, 23]. More recently a new strain of research explicitly included developmental aspects and the modeling of development [24, 25], and epigenetic robotics [24]. Examples are the work of Metta and Sandini [26, 27, 28, 29], of the group of Pfeifer [30, 31, 32], of Dautenhahn *et al.* [33, 34].

Imitation is one of the key stages in the development of more advanced cognitive capabilities. While the study of the ability of infants and adults to imitate has remained for the most part a field of the psychological literature, recently, it has found a ground in the neurological literature with the discovery of the mirror neuron system in monkeys.[35] The mirror neuron system is formed by premotor neurons discharging both when the animal acts and when it sees similar actions performed by other individuals. A system, similar to that found in monkeys, has been indirectly shown to exist also in humans by transcranial magnetic stimulation studies of the motor cortex during action observation [36]. Further investigations have shown that the mirror system can be activated not only by visually perceived actions but also by listening to action-related sounds [37] and, in

humans, by speech listening [38]. In addition to these electrophysiological data, a number of brain imaging studies in humans all point to a network of brain areas responsible for the visuo-motor transformation mechanism underlying action recognition [39, 40]. It is plausible that the motor resonant system formed by mirror neurons is involved in the understanding of someone else's action and in imitation.

4 The Nature and Scope of Cognitive Vision

Cognitive vision is a complex and multi-faceted discipline. In addressing it, we must distinguish clearly between, on the one hand, identifying what it is and, on the other, how it can be modelled and effected. This section deals with the former issue, identifying the problem issues that together define the area, while the next section surveys the different scientific approaches that have been proposed to model cognitive systems, including cognitive vision. This decoupling is not absolute, however, since the different approaches themselves are predicated upon some quite disparate perspectives on what cognition proper actually is. For the moment, we won't let this stand in our way and we will proceed in this section to circumscribe the area of cognitive vision, drawing out its nature and scope. In doing this, we will take up where we left off in the definition of cognitive vision in Section 1, and proceed in the same way to create as far as possible a paradigm-neutral treatment of the discipline. The treatment reverts to specific paradigms once we consider *how* the cognitive vision functionality can be effected.

Mindful of the need to deal effectively with the complexity of the subject, we propose that cognitive vision be cast in a 3-D space comprising three orthogonal considerations (see Figure 3).

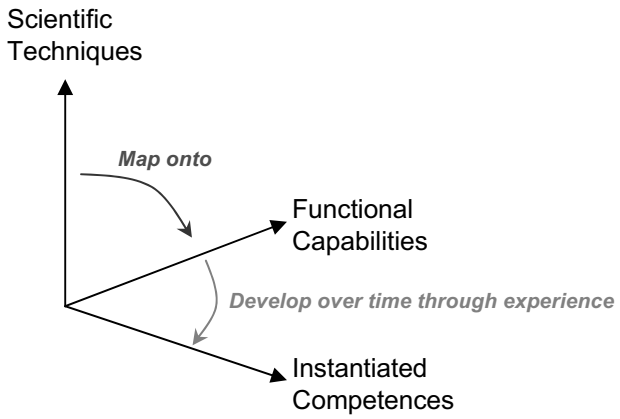

Figure 3: Three dimensions of cognitive vision: one or more scientific techniques map onto each functional capability and one or more functional capability contributes to the development of realized task-oriented system competences.

The first dimension embraces the *scientific techniques* that form the foundations of the subject: the constituent theoretical models, stating in a generic way the scientific issues without going so far as to say how they are to be tackled. For example, learning is a central scientific issue which can be modelled in several ways, depending on the paradigm one adopts.

The second dimension comprises the generic *functional capabilities*. By capability, we mean the potential faculty to achieve some function, purpose, goal, or desired state. This contrasts with the *competence* of a cognitive vision system by which we mean the actual power or faculty to achieve something. Competences are based on capabilities but are developed over time through experience, through learning, interaction, and practice in task-specific circumstances.

The system's *instantiated competences* then form the third dimension of our space of cognitive vision and constitute the verifiable task-dependent operation, behaviour, skills of the system, the performance of which can be subject to either quantitative or qualitative assessment in specific application scenarios.

Before proceeding, we need to remark on the relationship between these three dimensions. Each capability is brought about through the inter-operation of some sub-set of the scientific techniques. Thus, certain scientific techniques map onto certain capabilities (see Figure 3). Equally, each of the instantiated compentences is developed through experience from a sub-set of capabilities over some finite time period. This time period might be quite short (*e.g.* for hard-wired capabilities), or long (*e.g.* for capabilities that themselves have to be developed).

The key aim, therefore, in the specification of a cognitive vision system is to say what particular techniques are required for each capability and to say how each capability is transformed into an instantiated verifiable task-specific competence, which can be subjected to appropriate performance evaluation. Note that this schema is not static or fixed: the development of both capabilities and compentences proceed in time, and with experience, so that the manner in which some capabilities can develop will be affected by what the system's instantiated competences are. In short, experience influences not just competence but also the potential to develop new capabilities.

4.1 Enabling Scientific Techniques

Since Sections 5 and 6 address the approaches and scientific techniques that underpin the discipline, we will say very little here about specific techniques. Instead, we will simply identify the broad scientific issues that must be taken into consideration in some form or other by any candidate theory of cognitive vision. It should be noted that the terms we use to describe each issue are meant in their most general sense and are not to be interpreted in a paradigm-specific sense. For example, when we refer to reasoning we don't necessarily imply symbolic reasoning (as would be required in the cognitivist paradigm; see Section 5.1.1) but we include any mechanism whereby a space of possible options can be explored by the cognitive agent without necessarily committing to them or acting them out. Note also that there are clear inter-relationships between several of these issues. Indeed, each issue should be tightly linked since each forms (or should form) a necessary component in a complete visually-enabled cognitive system. In separating out these issues, we do not mean to imply that they are decoupled; the separation is merely a device to gain some

grip on this complex area. These scientific issues in question then are:

Visual Sensing. For the purposes of this document, we are concerned solely with sensing using vision. This is the mechanism or process whereby the cognitive system can be influenced by the environment around it. In a way, it is the mechanism by which the environment perturbs the cognitive system. We couch it these terms as it then allows for several interpretations of the purpose of sensing. For example, it allows for system autonomy and homeostasis, or self-regulation, as the primary focus of cognitive activity with the sensory apparatus providing the means for adapting to environmental changes and activities. On the other hand, this formulation also allows for sensing to be construed as an input mechanism whereby the system acquires information about the environment. As we will see later, both of these essentially antagonistic positions are held by proponents of different paradigms in cognition. This antagonism — or opposition in viewpoint — recurs repeatedly and is, as we have noted, an unavoidable consequence of the pre-paradigmatic status of the discipline.

Architecture. Here, the term architecture is intended in the sense of the minimal configuration of a system – *i.e.* its operational architecture – that is necessary for certain capabilities and cognitive behaviours. For approaches to cognition that have a focus on development from a primitive state to a fully cognitive state over the life-time of the system, the architecture of the system is equivalent to its phylogenic configuration: the innate capabilities with which it is endowed at the beginning of its life-time and which don't have to be learned (but may be developed further). For cognitivist systems, the architecture of the system is equivalent to the minimal set of information processing modules and their network of inter-relationships. In either case, one of the biggest open questions is 'what are the necessary and sufficient architectural requirements for cognitive behaviour?' That is, accepting that cognition is achieved in large part through development and learning, what operational architecture is needed to get the cognitive development started?

Representation. The term *representation* is very contentious in the realm of cognitive systems. The different approaches which we will survey in the next section take emphatically different positions on the issue of representations. Here, we again choose to use the term in its least pejorative and weakest sense, taking it to mean any stable state of a cognitive system. The meaning ascribed to the system state by either ourselves, as designers and observers of cognitive systems, or by the system itself is left quite properly to the discussion of a given cognitive paradigm. Thus, whether or how a given system state represents something is dependent on the cognition paradigm one adopts.

Memory. It seems that memory must play a pivotal part in cognitive systems as otherwise it would be impossible for the system to learn from experience, develop, adapt, recognize, plan, deliberate, reason; indeed, almost all of the other issues listed here depend to a lesser or greater extent on there being some form of persistent state. Clearly, the issue of memory is strongly linked to that of representation and, consequently, what you understand by the nature and form of memory is intimately linked with the approach you take to cognition. That said, we should note that

whatever theory is being developed in whatever paradigm, it seems that there will have to be room for different forms (or functions) of memory, such as propositional, episodic, modal (visual/acoustic), short-term, and long-term memory. There is also the key issues of graceful degradation of memory and the need to avoid catastrophic forgetting (*i.e.* the loss of important memories when something new is learned).

Learning. The issue of memory leads naturally to that of learning or the development of memory states. Here, learning is taken both in the sense of the acquisition of knowledge (model learning / parameter estimation) and in the sense of the acquisition of new techniques and capabilities. That is, learning is taken to be equivalent to the more general term *development*. Learning is crucial to cognitive systems and there are many forms of learning, which will be more or less appropriate depending on the cognition paradigm in which one is working.

Recognition. A cognitive vision system must be able to discriminate between visual entities, be their regions in the visual field, simple features, complicated behaviours, and even the potential of visual entities to fulfil certain functions. We place this ability to accomplish such discrimination under the heading of recognition but it should be noted that it is a much richer idea than simple pattern classification. In essence, a cognitive vision system must include a scientific theory that facilitates several forms of grouping using a variety of criteria, some based purely on visual appearance, some based on temporal change or behaviour over time, and some based on perceived functionality (*i.e.* the apparent potential to fulfil given functions). This last criterion is linked closely to the concept of affordance. We use the term *categorization* in its most general sense as a meta-level classification to denote this ability to identify affordance.

Deliberation & Reasoning. The ability to reason explictly about a problem is probably the functional characteristic that most people would associate with a cognitive system. The term *reasoning* is somewhat pejorative however as it is often equated with the symbolic reasoning methods of cognitivist approaches (see Section 5.1.1) even though the reasoning process can be achieved in other ways in different paradigms. Consequently, we will use the term *deliberation* instead to convey the intended meaning of explicit weighing of alternative options in selecting a given strategy or system behaviour. Unless otherwise stated, the term reasoning is also meant in this paradigm-neutral sense.

Planning. The concept of planning is closely allied to deliberation & reasoning, memory, and representation. By planning, we mean the ability to deal with events in the future: events that haven't happened yet and events that may never happen. It is thus concerned with anticipation, expectation, and managing contingencies (*i.e.* adapability is not purely an adaptive or reflex behaviour). Together with deliberation and memory, planning allows a cognitive system to extend its working into the future: to escape from being a reactive reflex system. Note again that we are not saying what form a scientific theory of planning should take — and we are certainly not saying that it should necessarily be symbolic — just that it is an issue that must be treated in any complete scientific theory of cognition and cognitive vision.

Communication. A key characteristic of advanced cognitive systems is the capability to communicate with other systems or agents. However, there are different forms of communication and we need to be clear that by communication in this context, we mean any form of communication that is effected *as a consequence of the system's cognitive activities.* We use the term *semiotic communication* to denote this form of communication. It does not include, for example, data-communication involving direct access to internal states of the system.

Action. We come finally to the last scientific issue and perhaps the most contentious one in cognitive vision systems (even more contentious perhaps than representation). It is the issue of action. The question that arises immediately is: What constitutes an action? Is it a physical act requiring forcible interaction or can the initiation of a simple change of state constitute an action? These questions are tightly related to the issue of embodiment which we will consider in some depth in Section 6.1.2 later in this document. For the present, we will satisfy ourselves simply by flagging the fact that a scientific theory of cognitive vision does have to address and settle the issue of action and embodiment.

4.2 Functional Capabilities of Cognitive Vision

Having dealth briefly with the various scientific techniques, we come now to the second dimension of cognitive vision: the functional capabilities. These are grouped under eight broad headings, each dealing with somewhat distinct issues. Note well however that we are not suggesting that the capabilities form a strict hierarchy. As we will see, many capabilities are co-dependent. On the other hand, within each group, we will suggest a form of ordering, not dissimilar from subsumption, so that one capability is exploited by (and subsumed by) another, typically in the sense that the first capability is required for the development of the second. The eight groups are:

1. Detection and Localization;
2. Tracking;
3. Classification and categorization;
4. Prediction;
5. Concept formation and visualization;
6. Inter-agent communication and expression;
7. Embodied exploration;
8. Hand-eye coordination.

We will catalogue the capabilities under these headings in the following. Note that any complete theory of cognitive vision should address the following five issues: the purpose for which the capability can be used, its inter-dependency with any other capability, its origin (*i.e.* how is the capability obtained: is it hard-wired or developed?), the underlying scientific technique, and the evaluation of the performance in carrying out that capability.

4.2.1 Detection and Localization Capabilities

A cognitive vision system must be equipped with a minimal set of computer vision capabilities. These base-line computer vision capabilities do not have to be independent: it is reasonable for some to exploit other vision capabilities to achieve their functionality. In the following, we will go through the various capabilities in the order of decreasing independence from one another.

First and foremost, a system must have the capability to discrimination between regions in the visual field of view on the basis of colour, texture, or depth. Note that no meaning is attributed to the discrimination. It is simply the capability to make a distinction between regions. Next, a system should be capable of detecting hard-wired features; this capability will typically exploit the discrimination capability. Together, we refer to this joint capability as *detection*.

Following the detection of features and regions, the next capability is that of localization, in the sense that the position and extent of hard-wired features can be determined. We assume here an ego-centric frame of reference.

Next comes the capability to establish the spatial relationships between regions or features, including mutual occlusion and relative ordering. Again, we assume an ego-centric frame of reference.

Similarly, there is the capability to establish temporal relationships between regions or features, *i.e.* between events. These relationships include simultaneity and relative ordering.

Finally, if the cognitive vision system we are considering is embodied, then the system should have the capability to acheive all the above under ego-motion.

4.2.2 Tracking Capabilities

A cognitive system should have a number of capabilities associated with tracking regions or features. Typically, these will build on the base-line computer vision capabilities. However, whereas the base-line vision capabilities are all based on hard-wired functionality, the tracking capabilities now include learned or developed functionalities. As before, some tracking capabilities depend on others and we will again work through the capabilities in order of decreasing independence.

First, a cognitive vision system will have a hard-wired tracking capability that can track pre-determined regions or features in continuous 3-D spatio-temporal paths.

It should be capable of integrated tracking based on multiple viewpoints or multiple cameras.

It should be capable of tracking learned regions or features.

It should also be capable of dealing with partial and complete occlusion of a tracked region or feature; typically, this will be accomplished with either a hard-wired or developed short-term predictive capability.

It should be capable of detecting unexpected regions, features or event; that is, it should be capable of detection of novelty in the visual scene.

Finally, a cognitive vision system should have a long-term predictive capability to allow it to exploit expectation and effect planning.

4.2.3 Classification and Categorization

Whilst the base-line computer vision capabilities incorporate some form of region or feature discrimination, a cognitive vision system must also be capable of several forms of grouping, *i.e.* classification and categorization. We note again here that by categorization, we mean the process of grouping based on more than visual form alone. In this sense, categorization is a meta-level classification based on function (temporally-extended behaviour) as well as on pure form or appearance.

The classification and categorization capabilities of a cognitive vision system, in order of decreasing independence, are as follows.

First, there is the straight-forward classification of regions or features, perhaps with the use of hard-wired constraints.

There is then the capability to categorize behaviours (*i.e.* temporally-extended configurations of regions or features) in terms of, for example, position, orientation, and velocity.

Just as with tracking, the capabilities must also include the facility to adapt and change. In the first instance, this means a capability to detect an instance of a new class, followed by a capability to learn classes of regions, features, and categories of behaviours. These capabilities should be able to work from simple basic physical manipulation behaviours all the way up the hierarchy to complex social behaviours.

A cognitive vision system should have the capability to form new classes and categories from old ones, *i.e.* to generalize classes and categories.

Continuing in the same vein, cognitive vision involves the recognition of and adaptation to novel variations in the current visual environmental context and the capability to generalize to new contexts, *i.e.* to new application scenarios.

Finally, a cognitive vision system should be capable of categorizing functionality on the basis of appearance of regions or features; that is, it should be capable of mapping from classes of objects to categories of behaviour.

4.2.4 Prediction

The capabilities of a cognitive vision system include being able to interpret the intent underlying behaviour, specifically to predict future spatio-temporal configurations of the visual environment, across a variety of time-scales, perhaps with the use of *a priori* or hard-wired constraints. This may involve the formulation of hypotheses about spatio-temporal configurations and subsequent deliberation about them either through active exploration or through some form of passive assessment.

4.2.5 Concept Formation and Visualization

A cognitive vision system should be able to create some form of conceptualization of its perceived environment and it should also be able to visualize those concepts. Naturally, this implies some capability to associate behaviours or visual entities derived using other capabilities with these concepts or abstractions.

This leads to the capability to establish conceptual object decomposition, with the decomposition being either relational or hierarchical.

Finally, a cognitive vision system should have the capability to engage in counterfactual deliberation, that is, the capability to view the world from the perspective of

another agent or system. This implies the capability (a) to instantiate someone else's model of the environment, and (b) to deliberate on that instantiated model, rather than on one's own innate model.

4.2.6 Inter-agent Communication and Expression

The communication capabilities of a cognitive vision system include the expression, either by gesture or or by verbal interaction, of an understanding of the environment to other systems, including humans. It also includes the capability to recognize and respond to facial expressions. Recall again the comment we made in the section on communication under the heading of scientific techniques that by communication we mean only communication that is effected as a consequence of the system's cognitive activities. It does not include, for example, data-communication involving direct access to internal states of the system. Typically, it will exploit capabilities for concept formation and visualization. Some authors refer to this as symbolic expression but, as before, we should be aware that this use of the term symbolic does not necessarily imply the physical symbol system manipulation of the cognitivist information processing paradigm of cognition.

4.2.7 Embodied Exploration

Exploration is a force that is needed to drive the development of a cognitive system. It is an in-built goal related to understanding the world, in the sense of extracting sense or order from the world, by driving the learning and generalization capabilities.

The exploration capabilities of an embodied cognitive vision system includes the formation of maps of the visual environment using single or multiple eco- or ego-centric frames of reference, denoting the location of labeled regions or features. These maps are the passive results of exploration. On the other hand, the capability to point to or gesture at objects or people is an active consequence of embodied exploration.

Finally, the capability to learn affordances is included under this heading. This capability, whereby the usability of an object or part of the visual environment in a specific context to achieve a particular purpose is perceived or inferred by the system, is an advanced and very important cognitive capability.

4.2.8 Hand-Eye Coordination

Embodied cognitive visions systems, in particular those with the ability to manipulate objects in the environment must be able to achieve a considerable degree of hand-eye coordination. There are several constituent capabilities. First, there is the capability of visual accommodation and focussing at various distances; this should be achievable both continuously and after saccades. Next, there is capability to control the gaze direction for scene scanning. The vision system should have the capability to develop of basic hand-eye coordination to allow it to follow, reach for, grasp at, and possibly grip presented objects. More advanced capabilities include learning specific coordination, *e.g.* to hit a ball with a bat, and to treat gripped objects as an extension of its own body. Associated with this is the capability to predict object, hand, and body motions. This leads then to a capability for self-awareness of systems's own body, and thence to a capability to imitate the actions of other agents.

4.3 Realized Task-Specific Competences of Cognitive Vision

Having dealt with the enabling scientific techniques and the functional capabilities, we come finally to the third dimension of our space of visually-enabled cognitive systems: the realized task-specific competences of cognitive vision. As we noted above, these are based on capabilities but are developed over time through experience, through learning, interaction, and practice in task-specific circumstances. Since these competences of a cognitive vision system are inevitably set in some practical context, it is more meaningful to describe them with respect to that context than to attempt the creation of a catalogue of a set of generic competences. In any case, such a catalogue would be too close to the underlying enabling capabilities to be very instructive. The approach we will take in this section therefore is to describe four representative application scenarios and list the competences that need to be realized to make that application possible. The four scenarios are: a surveillance system, a home assistant, an autonomous automobile, and finally the cognitive vision competences of infants and young children.

4.3.1 Competences of a Cognitive Surveillance System

The surveillance scenario is limited here to the single task of tracking an individual person. The competences that this task involve draw from several of functional capabilities, most notably classification and categorization, and tracking. The surveillance system would have to be able to recognize an individual to the tracked either on the basis of physical appearance or on the basis of behaviour. The behaviour could be either one that matches a given pattern or one that is anomalous in the context of the behaviours of other people in the scene. The cognitive vision competences required to achieve this include the ability to represent and classify a small number of objects, the ability to represent and classify categories of behaviours, the ability to learn object classes and behaviours, and the ability to generalize (*i.e.* to form new categories).

The system would then have to be able to track the individual through the scene. The system would need to have the competence to deal with changes in scale and occlusion, and changes in appearance (for example, when the individual turns his back or puts on or removes clothing). The system should also be able to exploit multiple camera configurations with inter-camera integration.

A well-developed cognitive vision surveillance system should also be portable to new contexts within the same application scenario as well as to new applications scenarios (*e.g.* the ability to easily reconfigure a train-station surveillance system for an airport).

4.3.2 Competences of a Cognitive Home Assistant

The home assistant that is envisaged here is a physical robotic mobile agent. There are two proto-typical scenarios to illustrate the required competences. 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 a house, for example, and indicates the name and purpose of household items.

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

1. Form mental maps of the world and associate labels with objects;

2. Explore the environment;

3. Identify objects and understand their purpose and function;

4. Detect novelty in the environment, including the introduction of new objects, removal of objects, and change in object location;

5. Interpret the actions of humans.

4.3.3 The Cognitive Vision Competences Required for an Autonomous Automobile

The following is a list of the the visual competences that would be required to build an autonomous automobile. To give some structure to the list, three environments that exhibit increasing levels of difficulty for visual interpretation are identified: highways, city streets, and off-road environments. Highways are highly-structured and controlled environment in the sense that there are a strictly limited number of visual scenarios and object behaviours. On the other hand, city streets 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 unexpected behaviours. In theory, the off-road environment should be the easiest scenario because the behaviour of the autonomous vehicle is not greatly constrained but, on the other hand, the variability in the visual environment is considerably increased.

The following competences would have to be realized in a cognitive vision system to effect autonomous navigation in these environments:

1. Detection of driveable free-space, dealing with static and moving obstacles (*e.g.* pedestrians, bicycles, animals), and other vehicles.

2. Detection of legal driveable free-space, implying the need to understand rules and laws.

3. Situation assessment, dealing with:

   (a) The position and velocity of the vehicle with respect to free-space.

   (b) The relative position and velocity of obstacles and other vehicles.

   (c) Models of behaviour (*e.g.* actions based on right of way) and prediction of behaviour.

   (d) Models of intent (*e.g.* acceleration on approach to an amber traffic light suggest an intent not to stop and give way).

   (e) Legal stipulations.

4.3.4 The Cognitive Vision Competences of Young Infants and Children

The final scenario is less focussed on applications and more on the flexible behaviour of cognitive agents. It is still task-specific because it is set in the particular context of human infants. A cognitive vision system that can emulate the competences of young infants would have to be able to recognize facial expressions and infer some sense of meaning or

intent from them. It would comprehend object persistence in space and time, and do so in the presence of occlusion. For example, if a ball rolled behind a door, it would understand that the ball was still there but was just hidden. In this sense, it would understand of the structure of its local space, though not necessarily in the same frame of reference as another cognitive agent. Similarly, it would be able to formulate hypotheses and reason visually. Ultimately, it would also be able to extend these competences to other frames of reference and be able to engage in counter-factual reasoning, *i.e.* reasoning from the point of view of another cognitive agent. A cognitive vision system would have the ability to point to and gesture at specific objects, and it would have the hand-eye coordination to grasp for objects. It would also be able to imitate of the actions of others and be able to use vision to enable self-locomotion.

5 Research Paradigms: Meeting the Capability Challenge

At this point, we have established a clear concept of the task-specific competences typical cognitive vision systems should exhibit and capabilities that a system should have to develop these competences. In this section, we will look more deeply at the scientific approaches that might underpin these capabilities. This is not quite as straightforward as it sounds because, as noted at the outset, there are several competing views on what cognition is and on how it should be effected. Therefore, we will begin with a survey of the different paradigms and follow this by a discussion of how well each paradigm is currently equipped to achieve the required capabilities. In the next section, we will consider the key issues in the development of the relevant scientific theories and then conclude in Section 7 with the final roadmap of priority research topics.

5.1 Survey of Existing and Emerging Paradigms in Cognition and Cognitive Vision

There are several quite distinct approaches to the understanding and synthesis of cognitive systems. These include physical symbol systems, connectionism, artificial life, dynamical systems, and enactive systems[41, 11]. Each of these makes significantly different assumptions about the nature of cognition, its purpose, and the manner in which cognition is achieved. Among these, however, we can discern two broad classes: the *cognitivist* approach based on symbolic information processing representational systems; and the *emergent systems* approach, embracing connectionist systems, dynamical systems, and enactive systems, based to a lesser or greater extent on principles of self-organization.

5.1.1 Cognitivist Models

Cognitivism asserts that cognition involves computations defined over symbolic representations, in a process whereby information about the world is abstracted by perception, represented using some appropriate symbol set, reasoned about, and then used to plan and act in the world. This approach has also been labelled by many as the *information processing* approach to cognition[42, 43, 44, 45, 46, 47, 41]. Traditionally, this has been the dominant theme in cognitive science[43] but there are indications that the discipline is migrating away from its stronger interpretations[11].

For cognitivist systems, cognition is representational in a strong and particular sense: it entails the manipulation of explicit symbolic representations of the state and behaviour of an objective external world[48]. Reasoning itself is symbolic: a procedural process whereby explicit representations of an objective world are manipulated and possibly translated into language.

In most cognitivist approaches concerned with the creation of artificial cognitive systems, the symbolic representations are the product of a human designer. This is significant because it means that they can be directly accessed and understood or interpreted by humans and that semantic knowledge can be embedded directly into and extracted directly from the system. However, it has been argued that this is also the key limiting factor of cognitivist vision systems: these designer-dependent representations are the idealized descriptions of a human cognitive entity and, as such, they effectively bias the system (or 'blind' it [48]) and constrain it to an domain of discourse that is dependent on and, a consequence of, the cognitive effects of human activity. This approach works well as long as the system doesn't have to stray too far from the conditions under which these descriptions were formulated. The further one does stray, the larger the 'semantic gap' [49] between perception and possible interpretation, a gap that is normally plugged by embedding programmer knowledge or enforcing expectation-driven constraints [50] to render a system practicable in a given space of problems.

This approach usually then goes hand-in-hand with the fundamental assumption that 'the world we perceive is isomorphic with our perceptions of it as a geometric environment'[51]. The goal of cognition, for a cognitivist, is to reason symbolically about these representations in order to effect intelligent, adaptive, anticipatory, goal-directed, behaviour.

The vast majority of computer vision systems, both cognitive and classical, adopt an essentially cognitivist position, especially with regard to their approach to representational issues (*e.g.* see [52, 53, 54]), although there are some notable exceptions (*e.g.* [55]). Since real perceptual systems work with inherently uncertain, time-varying, and incomplete information, computer vision systems are increasingly turning to the use of machine learning to improve the resilience of these systems (*e.g.* [56]). However, this doesn't alter the fact that the representational structure is still predicated on the descriptions of the designers. The significance of this will become apparent in later sections.

5.1.2 Emergent Systems

Emergent systems, embracing connectionist, dynamical, and enactive systems, take a very different view of cognition. Here, cognition is a process of self-organization whereby the system is continually re-constituting itself in real-time to maintain its operational identity through moderation of mutual system-environment interactions and co-determination[57]. Co-determination implies that the cognitive agent is specified by its environment and at the same time that the cognitive process determines what is real or meaningful for the agent. In a sense, co-determination means that the agent constructs its reality (its world) as a result of its operation in that world.

Co-determination is one of the key differences between the emergent paradigm and the cognitivist paradigm. For emergent systems, perception provides appropriate sensory data to enable effective action[57] but it does so as a consequence of the system's actions. In the emergent paradigm, cognition and perception is functionally-dependent on the richness of

the action interface[58].

One of the key features of emergent systems is that 'the system's connectivity becomes inseparable from its history of transformations, and related to the kind of task defined for the system'[41]. Whereas in the cognitivist approach the symbols are distinct from what they stand for, in the emergent approach, meaning relates to the global state of the system. Indeed, the meaning is something attributed by an external third-party observer to the correspondence of a system state with that of the world in which the emergent system is embedded.

Dynamical systems theory is one of the most promising approaches to the realization of emergent cognitive systems. Advocates of the dynamical systems approach to cognition (*e.g.* [43, 47, 59]) argue that motoric and perceptual systems, as well as perception-action coordination, are dynamical systems, that self-organize into meta-stable patterns of behaviour.

Proponents of dynamical systems point to the fact that they directly provide many of the characteristics inherent in natural cognitive systems such as multi-stability, adaptability, pattern formation and recognition, intentionality, and learning. These are achieved purely as a function of dynamical laws and consequent self-organization. They require no recourse to symbolic representations, especially those that are the result of human design.

Clark[11] has pointed out that the antipathy which proponents of dynamical systems approaches display toward cognitivist approaches rests on rather weak ground insofar as the scenarios they use to support their own case are not ones that require higher level reasoning: they are not 'representation hungry' and, therefore, are not well suited to be used in a general anti-representationalist (or anti-cognitivist) argument. At the same time, Clark also notes that this antipathy is actually less focussed on representations *per se* (dynamical systems readily admit internal states that can be construed as representations) but more on objectivist representations which form a isomorphic symbolic surrogate of an absolute external reality.

It has been argued that dynamical systems allow for the development of higher order cognitive functions such as intentionality and learning in a straightforward manner, at least in principle[43]. Although dynamical models can account for several non-trivial behaviours that require the integration of visual stimuli and motoric control, including the perception of affordances, perception of time to contact, and figure-ground bi-stability[60, 61, 43, 62, 63], the principled feasibility of higher-order cognitive faculties has yet to be validated.

Enactive systems take the emergent paradigm even further. In contradistinction to cognitivism, which involves a view of cognition that requires the representation of a given objective pre-determined world[59, 41], enaction[64, 65, 66, 57, 67, 41, 48] asserts that cognition is a process whereby the issues that are important for the continued existence of the cognitive entity are brought out or enacted: co-determined by the entity as it interacts with the environment in which it is embedded. Thus, nothing is 'pre-given', and hence there is no need for symbolic representations. Instead there is an enactive interpretation: a real-time context-based choosing of relevance. The advantage is that it focusses on the dynamics by which robust interpretation and adaptability arise.

A key postulate of enactive systems is that reasoning, as we commonly conceive it, is the consequence of reflexive use of the linguistic descriptive abilities to the cognitive agent itself. Linguistic capability is in turn developed as a consequence of the consensual co-development of an epistemology in a society of phylogenically-identical cognitive agents.

This is significant: reasoning in this sense is a descriptive phenomenon and is quite distinct from the self-organizing mechanism (*i.e.* structural coupling and operational closure[57]) by which the system/agent develops its cognitive and linguistic behaviours. Since language (and all inter-agent communication) is a manifestation of high-order cognition, specifically co-determination of consensual understanding amongst phylogenically-identical and ontogenically-compatible agents, symbolic or linguistic reasoning is actually the product of higher-order social cognitive systems rather than a generative process of the cognition of an individual agent.

Theoretical support for the emergent position can be found in recent studies which have shown that an organism can learn the dimensionality and geometry of the space in which it is embedded from an analysis of the dependencies between motoric commands and consequent sensory data, without any knowledge or reference to an external model of the world or the physical structure of the organism[68, 69]. The conceptions of space, geometry, and the world that the body distinguishes itself from arises from the sensorimotor interaction of the system, exactly the position advocated in developmental psychology[47].

5.1.3 Hybrid Models

Recently, effort has gone into developing approaches which combine aspects of the emergent systems and cognitivist systems [58, 7, 70]. These hybrid approaches have their roots in strong criticism of the use of explicit programmer-based knowledge in the creation of artificially-intelligent systems [71] and in the development of active 'animate' perceptual systems [72] in which perception-action behaviours become the focus, rather than the perceptual abstraction of representations. Such systems still use representations and representational invariances but it has been argued that these representations should only be constructed by the system itself as it interacts with and explores the world rather than through *a priori* specification or programming [58]. Thus, a system's ability to interpret objects and the external world is dependent on its ability to flexibly interact with it and interaction is an organizing mechanism that drives a coherence of association between perception and action. Action precedes perception and 'cognitive systems need to acquire information about the external world through learning or association' [7]. Hybrid systems are in many ways consistent with emergent systems while still exploiting programmer-centred (but not programmer-populated) representations (for example, see [56]).

5.2 Paradigms and Capabilities: How Do They Measure Up?

It is important to realize that the foregoing paradigms are not equally mature and it isn't clear which paradigm will ultimately be successful. The arguments in favour of dynamical systems and enactive systems are compelling but the current capabilities of cognitivist systems are actually more advanced. However, they are also quite brittle and have achieved little in all of the cognitive capabilities associated with generalization. Enactive and dynamical systems should in theory be much less brittle because they emerge through mutual specification and co-development with the environment, but their cognitive capabilities are actually very limited at present. The extent to which this will change, and the speed with which it will do so, is uncertain. Hybrid approaches seem to offer the best of both worlds but it is unclear how well one can combine what are ultimately highly antagonistic underlying philosophies. What is clear is that all paradigms will continue to evolve and should

be supported in that evolution. What is important for us in this document is to accord relative priorities in the research agenda that enables this evolution.

Ideally, we would present here a survey of the scientific techniques that can be deployed in each paradigm to achieve each capability, including where possible examples of their use in achieving task-specific competences. Unfortunately, time and space restrictions preclude this. However, the creation of such a survey is a valid exercise in its own right and will be attempted by the members of *ECVision* at some point in the future.

6 Scientific Development

The basis for our research roadmap is now becoming apparent: we have a clear view of the different approaches that one can adopt in developing visually-enabled cognitive systems. We know what enabling capabilities need to be realized and we can see how these capabilities combine to form task-specific competences. We also have to a lesser extent some idea of the scientific techniques that can be deployed in each paradigm to achieve each capability. A comprehensive research programme must ultimately address each capability for which we don't have a strong scientific theory in whatever paradigm seems the most appropriate and then set about creating one. However, this straightforward approach is not appropriate for formulating a research roadmap because many of the functional capabilities and their underpinning scientific theories are inter-dependent and should not be treated in isolation from one another. This is a natural consequence of the system-orientation of the area. Furthermore, we need to be conscious of the relative maturity and associated risk of each possible approach. To formulate the research roadmap well, we need to know and understand both the key issues that strongly influence the shape any scientific theory will take as well as the scientific problems that underpin the various capabilities. In this section, we will look at the first point and we will discuss two core scientific concerns. The approach one takes to these has a strong bearing on the way one will subsequently address the creation of a scientific theory of visually-enabled cognitive systems. In the next section, we will then treat each of the specific scientific and technological challenges that underpin the cognitive vision capabilities.

6.1 Core Scientific Concerns

6.1.1 The Balance between Phylogeny and Ontogeny: Hard-Wired Functionality *vs.* Learned Capabilities

In Section 4.1 we already noted that one of the biggest open questions in cognitive vision is the minimal architecture required to configure a cognitive vision system and enable it to boot-strap cognitive development. There are two perspectives on this, depending on whether one takes a cognitivist stance or an emergent stance.

In the cognitivist stance, the issue comes down to the balance between required 'pre-knowledge' and acquirable knowledge (and the processes that acquire that knowledge). The question then is: How much does one need to know and to be able to do in order to be capable of learning new things, such as concepts or actions? In other words, we need a clear cut set of conditions under which certain learning can take place.

In the emergent stance, there is a trade-off between phylogenic configuration and ontogenic development. Phylogeny — the evolution of the system configuration from generation to generation — determines the visuo-motor capability that a system is configured with at the outset and which facilitates the system's innate behaviours. Ontogenic development — the adaptation and learning of the system during its lifetime — gives rise to the cognitive capabilities that we seek. Since we don't have the luxury of having evolutionary timescales to allow phylogenic emergence of a cognitive system (we can't wait around to evolve a cognitive system from nothing) we must somehow identify a minimal phylogenic state of the system. In practice, this means that we must identify and effect visuo-motor capabilities for the minimal reflex behaviours that ontogenic development will subsequently build on to achieve cognitive behaviour.

In both stances, we need to decide what visual processing capabilities are needed for a minimal cognitive vision system. This is essentially the same thing as saying we need to decide how we will effect the first few generic vision functionalities outlined in Section 1.1 and expanded upon in Section 4.2.1.

6.1.2 The Necessity of Embodied Cognition

The question as to whether cognitive vision systems necessarily have to be embodied is one of the most divisive issues in the field. The divisiveness arises from the different stances taken by the different paradigms. We begin by looking at the issue from both the cognitivist and the emergent perspectives, and then consider the implications.

From the perspective of the cognitivist paradigm, there is no case for embodiment, at least none for it as a mandatory requirement of cognition. Cognitivist systems don't necessarily have to be embodied. The very essence of the cognitivist approach is that cognition comprises computational operations defined over symbolic representations and these computational operations are not tied to any given instantiation. They are abstract in principle. It is for this reason that it has been noted that cognitivism exhibits a form of mind-body dualism [47, 73]. Symbolic knowledge, framed in the concepts of the designer, can be programmed in directly and doesn't have to be developed by the system itself through exploration of the environment. Some cognitivist systems do exploit learning to augment or even supplant the *a priori* designed-in knowledge and thereby achieve a greater degree of adaptiveness, reconfigurability, and robustness. Embodiment may therefore offer an additional degree of freedom to facilitate this learning, but it is by no means necessary.

The clear advantage of this position is that a successful cognitivist model of cognition could be instantiated in any physical context and, theoretically at least, be ported to any application domain.

The perspective from emergent systems is diametrically opposed to the cognitivist position. Emergent systems, by definition, must be embodied and embedded in their environment in a situated historical developmental context [47].

To see why embodiment is a necessary condition of emergent cognition, consider what cognition means in the emergent paradigm. It is the process whereby an autonomous system becomes viable and effective in its environment. In this, there are two complementary things going on: one is the self-organization[6] of the system as distinct entity, and the

[6] The self-organization is achieved through an operationally-closed network of activities characterized by circular causality [43] and possibly modelled by a dynamical system defined over space of order parameters

second is the coupling of that entity with its environment. 'Perception, action, and cognition form a single process' [73] of self-organization *in the specific context of environmental perturbations of the system.* This gives rise to the co-development of the cognitive system and its environment and thereby to the ontogenic development of the system itself over its lifetime. This development is identically the cognitive process of establishing the space of mutually-consistent couplings. Put simply, the system's actions define its perceptions but subject to the strong constraints of continued dynamic self-organization. The space of perceptual possibilities is predicated not on an objective environment, but on the space of possible actions that the system can engage in whilst still maintaining the consistency of the coupling with the environment. These environmental perturbations don't control the system since they are not components of the system (and, by definition, don't play a part in the self-organization) but they do play a part in the ontogenic development of the system. Through this ontogenic development, the cognitive system develops its own epistemology, *i.e.* its own system-specific knowledge of its world, knowledge that has meaning exactly because it captures the consistency and invariance that emerges from the dynamic self-organization in the face of environmental coupling. Thus, we can see that, from this perspective, cognition is inseparable from 'bodily action' [73]: *without physical embodied exploration, a cognitive system has no basis for development.* Although this argument is compelling, it has one weakness: it requires you to accept the legitimacy of the emergent thesis. Many don't. If you do accept it, then the necessity of embodiment follows directly.

If this argument for embodied perception is valid then it's necessary to consider what exactly it is to be embodied. One form of embodiment, and clearly the type envisaged by proponents of the dynamical and enactive systems approach to cognition, is a physically-active body capable of moving in space, manipulating its environment, altering the state of the environment, and experiencing the physical forces associated with that manipulation [73]. This 'strong' form of embodiment clearly satisfies the conditions of the emergent argument for embodiment and it seems to be a good place to begin because, having satisfied the boundary conditions, one can then focus on the core problem: the development of rigorous models of cognitive and perceptual processing. However, many computer vision and cognitive systems researchers have concerns about accepting this scenario as it seems to suggest that the only possible cognitive vision systems are ones that are part of robotic systems. This goes against much of the motivation for the creation of cognitive vision systems: resilience, robustness, re-configurability, open-ended improvement of performance, and especially automatic adaptability to unforseen operating conditions. Robotic applications are certainly not the only ones that can benefit for these capabilities. But yet we seem to be concluding that this is the only domain in which a cognitive system can be developed. There are two issues at stake here: first, is there a 'weaker' form of embodiment that still satisfies the needs of emergent systems, and second, even if there isn't does this necessarily imply that the only domain of application of cognitive systems is robotics?

The first issue comes down to the question of what it means to act in the environment. Is a speech act an action? Does action requires mobility? Does action required any physical contact with the environment? Or, is it sufficient for a system to be able to effect some change in the environment? And, if this is the case, what exactly constitutes a change in the environment: a change in physical configuration or just a modification in its state,

and control parameters.

such as switching on and off some electrical device?

If one looks closely at the emergent paradigm, one finds two cornerstones: the operational closure (or circular causality) of system, and the structural coupling of the system with its environment. Operational closure by itself does not imply a need for embodiment: it is an organizational principle and applies to systems of many temporal and spatial scales. Coupling with the environment is a little trickier. The key requirement is that the mutual perturbations implied by the coupling, *i.e.* the mutual system-environment interactions, should be rich enough to drive the ontogenic development but not destructive of the self-organization. *It would seem then, that there is nothing in principle that requires the 'action' to be physical in any strong sense and, therefore, that it should be possible to develop an embodied cognitive vision system in any application that offers a suitably rich set of interactions.* There is, however, an important caveat. In such a system, there is no guarantee that the resultant cognitive behaviour will be in any way consistent with human models or preconceptions of cognitive behaviour (but that may be quite acceptable, as long as the system performs its task adquately). If we want to ensure compatibility with human cognition, then it would seem that we do indeed have to admit the stronger version of embodiment and adopt a domain of discourse that is the same as the one in which we live: one that involves physical movement, forcible manipulation, and exploration.

This brings us to the second issue: is a cognitive vision system that has been developed in a robotics setting only of use in that setting? The answer is probably not: once the cognitive capacity has been developed, removal of the robotic interaction doesn't diminish the capacity, though it may inhibit further development. Thus, in principle, a cognitive vision system might be developed in a robotic setting and then transplanted to an embedded passive setting.

7 The Research Roadmap

We come finally to the research roadmap proper: a list of priority topics that require immediate investigation if the emerging area of cognitive vision — the science of visually-enabled cognitive systems — is to develop to the point where it can be considered a discipline with well-understood aims, sound constitutive theories, and exploitable technologies. This section sets out this schedule, first by enumerating the topics and then by considering the timing and resource implications. It concludes with a brief analysis of ways to maximize the impact of the roadmap.

It should be clearly understood at the outset that there is a need to engage in research of each and every one of the ten scientific techniques identified in Section 4.1:

1. Visual sensing;
2. Architecture;
3. Representation;
4. Memory;
5. Learning;
6. Recognition;
7. Deliberation & reasoning;
8. Planning;

9. Communication;
10. Action.

Similarly, there is the complementary need to develop all eight of the functional capabilities identified in Section 4.2 (but not necessarily all in the same system):

1. Detection and Localization;
2. Tracking;
3. Classification and categorization;
4. Prediction;
5. Concept formation and visualization;
6. Inter-agent communication and expression;
7. Embodied exploration;
8. Hand-eye coordination.

However, because of the inter-dependencies both between the scientific techniques and between the functional capabilities that we have already spoken of, this research should be carried out in the context of major 'challenges'. We propose here seven such challenges that constitute the priority topics for research in cognitive vision. These are:

1. Advancement of methods for continuous learning;
2. Identification of minimal system architecture(s);
3. Goal identification and achievement;
4. Generalization of operation;
5. Utilization of systems engineering methodologies;
6. Development of complete systems with well-defined competences;
7. Creation of research tools;

We discuss each of these in detail below. In addition, Table 1 shows these priority challenges set against the individual scientific techniques. A '$\times$' symbol indicates a clear and straightforwardly identifiable relationship between the challenges and the techniques. We could have presented similar tables relating the functional capabilities to the scientific techniques and relating the functional capabilities to the challenges. However, the first exercise forms part of the architecture challenge and the second then follows automatically.

7.1 Priority Challenges

7.1.1 Methods for Continuous Learning

In the foregoing, we have seen that the creation of cognitive systems, be they based on the premises of cognitivist approaches or emergent approaches, depend to a great degree on inherent systemic development, through learning, development, and exploration. Cognitive systems — visual or otherwise — are shaped by their experiences. Furthermore, cognitive development is an open-ended process. Consequently, there is a great need to build on our present understanding of methods for continuous learning and develop them

	Visual Sensing	Architecture	Representation	Memory	Learning	Recognition	Deliberation	Planning	Communication	Action
Continuous Learning	×		×	×	×	×				
Minimal Architecture	×	×	×	×	×	×	×	×	×	×
Goal Achievement					×	×	×	×	×	×
Generalization			×		×	×	×			
Systems Engineering		×								
Complete Systems	×	×	×	×	×	×	×	×	×	×
Research Tools	×	×								×

Table 1: Relationships between the priority challenges (down) and the individual scientific techniques (across). A '×' symbol indicates a clear and straightforwardly identifiable relationship. Other relationships may also exist.

to work on an extended range of issues such as invariances, affordances, actions/activities, concepts, representations, and categories.

More specifically, research on three aspects of continuous learning is required. These are:

1. Learning mechanisms;
2. Representations and feature sets;
3. Learning domains.

The learning mechanisms should facilitate fast, incremental, and continuous learning, with large capacity and graceful degradation.

The representations and feature sets need to be sparse, efficient, and extendable. Visual sensing techniques need to be developed to create these feature sets.

Finally, the use of learning in several domains should be investigated. These domains include the development of representations, both spatial (or perceptual) and symbolic (or conceptual), and the mapping between these representations.[7] The mapping from perceptual to conceptual is intended to facilitate expression, communication, categorization, and deliberation, whereas the mapping from conceptual to perceptual is intended to facilitate contextualization and/or embodied action. It must be emphasized that we are concerned in both domains with learning in two distinct senses: first, parameter estimation (the population of designed representations and mappings); and, second, system identification (the development of new representations and mappings). Learning in the second sense is the more difficult research challenge.

[7]Note well that we use both of the terms *representation* and *symbolic* in the neutral senses we have adopted throughout this document and without any prejudice as to the cognition paradigm (cognitivist, connectionist, dynamical, or enactive) in which learning is effected.

There is also a need to investigate training strategies for all learning mechanisms to overcome possible bottlenecks in learning rates such as those caused by the requirement of real-time coupling and interaction between cognitive agent and its environment and those caused by embodied cognition. Possible stategies include mixed virtual and real-world training.

7.1.2 Minimal Architectures

The balance between phylogenic configuration and ontogentic development and learning should be investigated to find out what is the minimal system configuration or architecture that will permit effective cognitive development. For example, in the case of a conventional physical symbol representational system (*i.e.* a cognitivist system) we need to identify the minimal set of information processing modules and their network of inter-relationships. In the case of a dynamical system, we need to identify at least the dynamical equations, the collective variables (order parameters), and the control parameters.

A significant part of this exercise is to specify how each scientific techniques should be mapped to each of the functional capabilities, acknowledging that not every system will require all techniques or all capabilities.

7.1.3 Goals Identification and Achievement

Goals are crucial for cognitive systems. With cognitivist approaches, goals are specified or stipulated explicitly by external observers in terms that are based on the required abilities as assemblages of contributing capabilities. That is, they are specified in terms of the outcome of cognitive behaviour. With emergent approaches, goals are much more difficult to specify since cognitive behaviour is a non-specific emergent consequence of a set of system dynamics. Consequently, they have to be stipulated in terms of constraints or boundary conditions on the system configuration, either through phyogenic make-up or ontologenic development, or both. It is a significant research challenge to understand how to do this effectively.

Recognizing that different approaches or paradigms have fundamentally different viewpoints on how goals can be introduced into the system and what part they play in the specification of the system, it is necessary nontheless for all approaches to investigate the issue. In particular, it is necessary to consider:

1. how implicit goals, which can be a function of the phylogenic make-up of the system, *i.e.* of its initial configuration, can be specified and incorporated.

2. how externally-communicated goals can be introduced to the sytem from its environment or from those interacting with it, *e.g.* through some form of communication or through training.

7.1.4 Generalization

The transferrability of competences or skills from one context to another is one of the ultimate goals of cognitive vision systems. This applies both in the sense of adaptability to new application scenarios and in the sense of being able to hypothesize about different

'virtual' situations in a given scenario. This transferrability requires some mechanism for generalization based on experience, achieved perhaps through analogical reasoning, metaphorical deliberation, or manipulation of visual memory.

As always, the particular approach will be dependent on the chosen paradigm. Of particular concern is the need to investigate the extent to which this generalization capability is dependent on having established some form of cognitive communication in the system (such as language, speech, or gesture).

7.1.5 Systems Engineering

Two things are relatively certain: (a) cognitive vision systems will exhibit a very high degree of system complexity and (b) they will depend heavily on software-based processing. Quite apart from the challenges of developing cognitive characteristics, the system and software complexity brings with it their own problems such as brittleness and ungraceful degradation of performance. There is a need to address the deployment of advanced systems science to achieve autonomic — self-regulation, self-description, self-healing, self-critical — system behaviour as a necessary complement to the processes of cognitive development. Furthermore, the relationship between these systems engineering considerations and cognitive processes should be addressed. For example, homeostasis or self-regulation is a plausible (and possibly necessary) precursor to the emergence of cognition.

7.1.6 Development of Complete Systems with Well-defined Competences

In cognitive vision, the systems aspect is crucial. A cognitive system, visual or otherwise, is first and foremost a system: a collection of necessary and mutually-dependent parts which together achieve a desired function or behaviour. A priority then in the development of cognitive vision is to focus on the construction of complete visually-enabled cognitive systems. The functionality should be specified along the 'competence' dimension of the 3-D space and the performance can be assessed accordingly. It is not enough to focus on just one or two distinct capabilities: any strong research programme must target the whole system and at least a significant majority of the capabilities detailed in the sections above. Furthermore, these complete systems should be developed in the context of well-specified application scenarios that require the development of explicitly-identified task-specific competences. Thus, research in visually-enabled cognitive systems should be carried out in the context of complete systems, with a critical mass of functional capabilities, driven by the need to exhibit or develop specific task-oriented competences. Finally, the manner in which each scientific technique is integrated into the whole system must be set out in detail (see also Section 7.1.6).

In creating complete systems, one needs to be cautious in to ensure that the complexity of the system is properly bounded by keeping in mind the need to say why each component is a necessary part of the system, what role it plays, and how it interacts with every other component. The key goal in this endeavour is the achievement of a critical mass of capabilities in a given application context rather than the incorporation of the full complement of capabilities catalogued above.

7.1.7 Research Tools

There is a need therefore to create the research tools that provide at least the minimal infrastructural pre-requisites to enable the construction of complete systems. Different approaches to cognitive vision will have different needs and ideally each should be catered for. Inevitably, several tools can be shared by different approaches.

The tools that need to be targetted include:

- Physical robotics systems with a sufficiently rich set of interfaces and a sufficiently large number of degrees of freedom to (a) allow the development of cognitive behaviour, and (b) exercise the full spectrum of capabilities identified in the 3-D space of cognitive vision detailed above.
- Software development environments with common device interfaces.
- Benchmarking scenarios, such as are outlined in the competence scenarios in Section 4.3.

It needs to be emphasized that the aim in developing these research tools is to create a common research framework to facilitate collaborative and even competitive work, to enable sharing and effective comparison of results, and to promote the establishment of a common body of theoretical and empirical knowledge in cognitive vision systems.

7.2 Time Horizons and Resource Implications

We have noted the pre-paradigmatic status of cognitive vision several times already in this document. This status means that the time horizon for the the achievement of the ultimate goals of cognitive vision systems are quite distant: as a whole, this research roadmap represents a 20 year plan, not a 5 year plan. With that said, there is reason to be optimistic that much will be accomplished on the road to the ultimate goals and that significant results can be attained in the short-term as well as the long-term. In particular, successful results in the short-term will probably be achieved in the priority challenges of Architecture and Learning, insofar as scientific breakthroughs are concerned, and in the priority challenges of Tools and Systems Engineering, insofar as enabling technologies are concerned. The complete attainment of the aims of the Generalization and Goals Achievement priority challenges may take somewhat longer. Work on the Complete Systems challenge will impact on all others in both the short-term and the long-term.

The pre-paradigmatic status also has an effect on the resource implications and on the level of required investment because several competing strands of research need to be accommodated. This is further accentuated by the system nature of cognitive vision since it is essential that all of the priority challenges be worked on at the same time, beginning now. One can't simply cherry-pick a topic and focus on that to the exclusion of the others; it is essential that the entire roadmap be followed, not just isolated parts of it.

7.3 Leveraging Impact

Even though we focus primarily on the scientific development of cognitive vision systems in this roadmap, it is important not to forget that the ultimate purpose of cognitive vision

is to facilitate more robust commercial vision-based applications. Consequently, the involvement of industrial interests is crucial to the development of the area, both to provide focus on potential applications and identify essential functionality. Typically, cognitive vision systems will provide adaptable and adaptive interfaces between humans and machines. These interfaces will be part of applications that either monitor human behaviour or interact with humans, or both, especially in surroundings and situations that cannot be modelled completely when the system is being designed. Section 4.3 detailed four application scenarios (surveillance and monitoring, home assistance, autonomous navigation, and emulation of infant behaviour). Other applications include the assessment of the behaviour of shoppers in retail outlets, interactive toys, semantic annotation of image databases, and monitoring of adaptive advertisements. To appreciate the impace of cognitive vision, it is worth considering one application scenario in a little more detail. For example, home assistance for the elderly and infirm is one class of applications that has much potential for the exploitation of various degrees of cognitive vision functionality. With the increase in the age profile of the global population, and the consequent rise in the cost of institutional healthcare, home-based care for the elderly has become not only socially-desirable but economically-necessary. This brings with it many problems, typically derived from the need to monitor the condition and activities of the client in an non-intrusive manner and to guide or prompt their actions in a natural way. This is in essence an exercise in the creation of advanced interfaces between humans and assistive technology and cognitive vision has a significant role to play.

In addition to providing application drivers for visually-enabled cognitive systems, industrial involvement is also important to ensure effective utilization of the techniques being developed: different aspects of the discipline will mature at different times and by having industry directly involved it is easier to spot opportunities for timely commercial exploitation.

This document has set out what needs to be done to advance the discipline of cognitive vision beyond it pre-paradigmatic status. If the document is to have long-term impact, this work must now be carried out by the scientific community, with or without EU funding. Either way, it is hoped that the ideas put forward here will also be a positive influence on those who are charged with creating the workplans for Framework 7 and beyond.

This research roadmap, three years in the making, is being finalized as the end date for the *ECVision* contract approaches. This timing is not quite as unfortunate as it seems for *ECVision*, as a thematic network, is precluded by its contract from funding actual research. The network may identify what needs to be the subject of research, as we have done here, but it can't then take responsibilty for carrying out the subsequent work. On the other hand, instruments do exist under Framework 6 that facilitate the conduct of research. These include Networks of Excellence. In this document, we have seen that the emerging discipline of cognitive vision can best be viewed as a form of visually-enabled cognitive system and, therefore, it is a special case of a more general area. Since the research projects in cognitive systems have only recently commenced, and since the research agenda in the general area of cognitive systems is only now at the stage *ECVision* was three years ago, it might be appropriate to delay the launch of a Network of Excellence in cognitive systems for some time. However, it would be very opportune and timely to create a Network of Excellence in visually-enabled cognitive systems as a pilot precursor to a full cognitive systems Network of Excellence which would subsequently subsume an *ECVision* Network

of Excellence under its broad umbrella of topics, together with other special interest groups in the other modalities. The proposal of such a Network of Excellence in visually-enabled cognitive systems is of course a matter for the community, the members of *ECVision*, and the European Commission, but it would be one very effective way of ensuring the future of cognitive vision.

8 Epilogue

The ultimate aim and guiding principle in our research should be to seek for simple cognitive principles and mechanisms from which complex visually-enabled cognitive behaviour can emerge, in whatever paradigm one is operating. The hope is that this roadmap will aid the search.

References

[1] J. L. Crowley. *Cognitive Vision Research Roadmap, Version 2.5.* 2003. `http://www.ecvision.info/research_planning/ECVisionRoadMapv2.5.pdf`.

[2] European Commission. *Call for Proposals on Cognitive Systems.* 2003. `http://www.ecvision.org/news/CS-Support_Document-v2.pdf`.

[3] H. I. Christensen and H.-H. Nagel. *Report on Dagstuhl Seminar 03441: Cognitive Vision Systems.* 2003. `http://www.dagstuhl.de/03441/Report/`.

[4] O. D. Faugeras. *3-D Computer Vision.* MIT Press, 1993.

[5] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2000.

[6] T. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In H. Burkhardt and B. Neumann, editors, *Proceeding of the 5th European Conference on Computer Vision*, volume 2, pages 484–498. Springer-Verlag, 1998.

[7] G. H. Granlund. Does vision inevitably have to be active? In *Proceedings of SCIA99, Scandanavian Conference on Image Analysis*, 1999.

[8] D. Vernon. The space of cognitive vision. In H.-H. Nagel and H. I. Christensen, editors, *Cognitive Vision Systems*, Lecture Notes in Computer Science. Springer-Verlag, 2004. In Press.

[9] *ECVision: European Research Network for Cognitive Computer Vision.* Network Website, 2003. `http://www.ecvision.org`.

[10] H.-H. Nagel. Reflections on cognitive vision systems. In J. Crowley and *et al*, editors, *Proceedings of the Third International Conference on Computer Vision Systems, ICVS 2003*, volume LNCS 2626, pages 34–43, Berlin Heidelberg, 2003. Springer-Verlag.

[11] A. Clark. *Mindware – An Introduction to the Philosophy of Cognitive Science*. Oxford University Press, New York, 2001.

[12] G. M. Edelman. *Neural Darwinism: The Theory of Neuronal Group Selection*. Oxford University Press, Oxford, 1988.

[13] M. I. (Ed.) Posner. *Foundations of Cognitive Science*. The MIT Press, Cambridge, MA, fifth edition, 1996.

[14] A. J. Ijspeert. Synthetic approaches to neurobiology: review and case study in the control of anguiliform locomotion. In *Fifth European Conference on Artificial Life (ECAL99)*, 1999.

[15] R. Moller, M. Marinus, and D. Lambrinos. A neural model of landmark navigation in insects. In *Computational Neuroscience Meeting CNS'98*, 1998.

[16] C. Breazeal. *Sociable Machines: Expressive Social Exchange Between Humans and Robots*. Unpublished Doctoral Dissertation. MIT, Cambridge, MA., 2000.

[17] M. Arbib, A. Billard, M. Iacoboni, and E. Oztop. Mirror neurons, imitation and (synthetic) brain imaging. *Neural Networks*, 13(8/9):953–973, 2000.

[18] A. Billard. Imitation. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA, 2002.

[19] S. Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3(6):233–242, 1999.

[20] A. Billard and K. Dautenhahn. Grounding communication in autonomous robots: an experimental study. *Robotics and Autonomous Systems*, 24(1/2 "Scientific methods in mobile robotics"):71–79, 1998.

[21] A. Billard and K. Dautenhahn. Experiments in social robotics: grounding and use of communication in autonomous agents. *Adaptive Behavior*, 7(3/4):415–438, 2000.

[22] L. Steels and P. Vogt. Grounding adaptive language games in robotic agents. In P. Husbands and I. Harvey, editors, *Fourth European Conference on Artificial Life, ECAL97*, pages 473–484. MIT Press, 1997.

[23] P. Varshavskaya. Behavior-based early language development on a humanoid robot. In *Epigenetic Robotics*, pages 149–158, Edingburg, UK, 2002. Lund University.

[24] J. Zlatev and C. Balkenius. Introduction: Why "epigenetic robotics"? In *First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, volume 85, pages 1–4. Lund University Cognitive Studies, 2001.

[25] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini. Developmental robotics: A survey. *Connection Science*, Forthcoming, 2003.

[26] G. Metta. *Babybot: a Study on Sensori-motor Development*. Ph.d. thesis, University of Genova, 2000.

[27] G. Metta and P. Fitzpatrick. Early integration of vision and manipulation. *Adaptive Behavior*, 11(2):109–128, 2003.

[28] G. Metta, G. Sandini, and J. Konczak. A developmental approach to visually-guided reaching in artificial systems. *Neural Networks*, 12(10):1413–1427, 1999.

[29] G. Metta, G. Sandini, L. Natale, and F. Panerai. Development and robotics. In *IEEE-RAS International Conference on Humanoid Robots*, Tokyo, Japan, 2001. Humanoids2001 organizers.

[30] M. Lungarella and L. Berthouze. Adaptivity through physical immaturity. In *2nd International Workshop on Epigenetic Robotics (EPIROB'02)*, pages 79–86, Edinburgh, Scotland, 2002.

[31] R. Pfeifer and C. Scheier. Sensory-motor coordination: The metaphor and beyond. *Robotics and Autonomous Systems*, 20:157–178, 1997.

[32] R. Pfeifer and C. Scheier. Representation in natural and artificial agents: an embodied cognitive science perspective. In *Natural Organisms, Artificial Organisms, and Their Brains*, pages 480–503, Bielefeld, Germany, 1998. Verlag.

[33] K. Dautenhahn and S. Coles. Narrative intelligence from the bottom up: A computational framework for the study of story-telling in autonomous agents. *JASSS, The Journal of Artificial Societies and Social Simulation*, 2000.

[34] K. Dautenhahn, B. Ogden, and T. Quick. From embodied to socially embedded agents - implications for interaction-aware robots. *Cognitive Systems Research*, 3(3):397–428, 2002.

[35] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141, 1996.

[36] L. Fadiga, L. Fogassi, G. Pavesi, and G. Rizzolatti. Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology*, 73(6):2608–2611, 1995.

[37] E. Kohler, C. Keysers, M. A. Umilta, L. Fogassi, V. Gallese, and G. Rizzolatti. Hearing sounds, understanding actions: action representation in mirror neurons. *Science*, 297:846–848, 2002.

[38] L. Fadiga, L. Craighero, G. Buccino, and G. Rizzolatti. Speech listening specifically modulates the excitability of tongue muscles: a tms study. *European Journal of Neuroscience*, 15:399, 2002.

[39] J. Decety, T. Chaminade, J. Grezes, and A.N. Meltzoff. A pet exploration of the neural mechanisms involved in reciprocal imitation. *NeuroImage*, 15:265–272, 2002.

[40] M. Iacoboni, R.P. Woods, M. Brass, H. Bekkering, J.C. Mazziotta, and G. Rizzolatti. Cortical mechanisms of human imitation. *Science*, 286:2526–2528, 1999.

[41] F. J. Varela. Whence perceptual meaning? a cartorgraphy of current ideas. In F. J. Varela and J.-P. Dupuy, editors, *Understanding Origins – Contemporary Views on the Origin of Life, Mind and Society*, volume 130 of *Boston Studies in the Philosophy of Science*, pages 235–263. Kluwer Academic Publishers, 1992.

[42] J. Haugland. Semantic engines: An introduction to mind design. In J. Haugland, editor, *Mind Design: Philosophy, Psychology, Artificial Intelligence*, Cambridge, Massachusetts, 1982. Bradford Books, MIT Press.

[43] J. A. S. Kelso. *Dynamic Patterns – The Self-Organization of Brain and Behaviour.* M.I.T. Press, 3rd edition, 1995.

[44] J. F. Kihlstrom. The cognitive unconscious. *Science*, 237:1445–1452, September 1987.

[45] D. Marr. Artificial intelligence – a personal view. *Artificial Intelligence*, 9:37–48, 1977.

[46] S. Pinker. Visual cognition: An introduction. *Cognition*, 18:1–63, 1984.

[47] E. Thelen and L. B. Smith. *A Dynamic Systems Approach to the Development of Cognition and Action.* MIT Press / Bradford Books Series in Cognitive Psychology. MIT Press, Cambridge, Massachusetts, 1994.

[48] T. Winograd and F. Flores. *Understanding Computers and Cognition – A New Foundation for Design.* Addison-Wesley Publishing Company, Inc.,, Reading, Massachusetts, 1986.

[49] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[50] J. Pauli and G. Sommer. Perceptual organization with image formation compatibilities. *Pattern Recognition Letters*, 23:803–817, 2002.

[51] R. N. Shepard and S. Hurwitz. Upward direction, mental rotation, and discrimination of left and right turns in maps. *Cognition*, 18:161–193, 1984.

[52] N. Maillot, M. Thonnat, and A. Boucher. Towards ontology based cognitive vision. In J. Crowley and *et al*, editors, *Proceedings of the Third International Conference on Computer Vision Systems, ICVS 2003*, volume LNCS 2626, pages 44–43, Berlin Heidelberg, 2003. Springer-Verlag.

[53] Peters G. Efficient pose estimation using view-based object representations. In J. Crowley and *et al*, editors, *Proceedings of the Third International Conference on Computer Vision Systems, ICVS 2003*, volume LNCS 2626, pages 12–21, Berlin Heidelberg, 2003. Springer-Verlag.

[54] G. Heidemann, R. Rae, H. Bekel, I. Bax, and H. Ritter. Integrating context-free and context-dependent attentional mechanisms for gestural object reference. In J. Crowley and *et al*, editors, *Proceedings of the Third International Conference on Computer Vision Systems, ICVS 2003*, volume LNCS 2626, pages 22–33, Berlin Heidelberg, 2003. Springer-Verlag.

[55] M. Jogan, M. Artac, D. Skocaj, and A. Leonardis. A framework for robust and incremental self-localization of a mobile robot. In J. Crowley and *et al*, editors, *Proceedings of the Third International Conference on Computer Vision Systems, ICVS 2003*, volume LNCS 2626, pages 460–469, Berlin Heidelberg, 2003. Springer-Verlag.

[56] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In T. Pajdla and J. Matas, editors, *Proceeding of the 8th European Conference on Computer Vision, ECCV 2004*, volume I of *LNCS*, pages 28–39. Springer, 2004.

[57] H. Maturana and F. Varela. *The Tree of Knowledge – The Biological Roots of Human Understanding*. New Science Library, Boston & London, 1987.

[58] G. H. Granlund. The complexity of vision. *Signal Processing*, 74:101–126, 1999.

[59] T. van Gelder and R. F. Port. It's about time: An overview of the dynamical approach to cognition. In R. F. Port and T. van Gelder, editors, *Mind as Motion – Explorations in the Dynamics of Cognition*, Cambridge, Massachusetts, 1995. Bradford Books, MIT Press.

[60] J. J. Gibson. *The Perception of the Visual World*. Houghton Mifflin, Boston, 1950.

[61] J. J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979.

[62] W. Köhler. *Dynamics in Psychology*. Liveright, New York, 1940.

[63] W. H. Warren. Perceiving affordances: Visual guidance of stairclimbing. *Journal of Experimental Psychology: Human Perception and Performance*, 10:683–703, 1984.

[64] H. Maturana. *Biology of Cognition*. 1970.

[65] H. Maturana. *The Organization of the Living: a Theory of the Living Organization*. 1975.

[66] H. R. Maturana and F. J. Varela. *Autopoiesis and Cognition*, volume 42 of *Boston Studies on the Philosophy of Science*. D. Reidel Publishing Company, 1980.

[67] F. Varela. *Principles of Biological Autonomy*. 1979.

[68] D. Phillipona, J. K. O'Regan, and J.-P. Nadal. Is there something out there? inferring space from sensorimotor dependencies. *Neural Computation*, 15(9), 2003.

[69] D. Phillipona, J. K. O'Regan, J.-P. Nadal, and O. J.-M. D. Coenen. Perception of the structure of the physical world using unknown multimodal sensors and effectors. NIPS 2003, page 8 pages, 2003.

[70] G. H. Granlund. Cognitive vision – background and research issues. Linköping University, 2002.

[71] H. L. Dreyfus. From micro-worlds to knowledge representation. In J. Haugland, editor, *Mind Design: Philosophy, Psychology, Artificial Intelligence*, Cambridge, Massachusetts, 1982. Bradford Books, MIT Press. Excerpted from the Introduction to the second edition of the author's *What Computers Can't Do*, Harper and Row, 1979.

[72] D. H. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

[73] E. Thelen. Time-scale dynamics and the development of embodied cognition. In R. F. Port and T. van Gelder, editors, *Mind as Motion – Explorations in the Dynamics of Cognition*, pages 69–100, Cambridge, Massachusetts, 1995. Bradford Books, MIT Press.