INFORMATION SOCIETY TECHNOLOGIES

(IST) PROGRAMME

Cognitive Vision Research Roadmap

Draft - Version 2.4

1 March 2003

ECVision

European Research Network for

Cognitive AI-enabled Computer Vision Systems

IST-2001-35454

Table of Contents

Overview and summary

(Jim - Last thing to write)

Co-authors include
David Vernon (Captec) Definition of Cognitive Vision
Patrick Courtney - Industrial Applications - many contributions.
Pia Böttcher - Industrial Applications and many contributions
Bernd Neumann (Univ. of Hamburg) - AI and Cognitive Vision.
Rebecca Simpson (Sera) Section on Enabling Technologies
Markus Vincze (Vienna University of Technology) - Robotics and Automation
Jan-Mark Geusebroek (University of Amsterdam)
Arnold W. M. Smeulders (University of Amsterdam)
Wolfgang Forstner (University of Bonn) - Self Diagnosis.

The key things to address here are:

The objective of the document;
The reasons it is needed;
The people it is directed at;
How it is organized (and why it is organized this way);
The basis on which it was founded.
These will be written one the document is stable. However, here is a start.

Computer vision is the science of machines that see. The goal of computer vision is to endow a machine with the ability to understand the structure, composition, and behaviour of its physical surroundings through the use of visual data. Over the last few years, exponential growth in computing power has provided inexpensive computing devices capable of processing images in real time. In parallel, rapid progress has been made in the development of theories and methods for using visual information to reconstruct and model the spatio-temporal world. The convergence of these two innovations has triggered significantly increased growth in the use of visual sensing. However, current growth rates are a small fraction of the potential growth

rates for a technology of machines that see. Further more, the diversity and economic scale of potential applications offers the promise of a substantial impact.

Cognitive computer vision is a scientific domain that emerges from the convergence of Cognitive Systems and Computer Vision. The results of this discipline will enable the creation of new classes of technologies in fields that require machine to perceive and interact with its environment. The goal of this Research road map is to provide the potential for impact of cognitive vision and to identify the best way in which this potential can be realized. It establishes an ontology of fundamental concepts so that we can unambiguously discuss its constituent parts. It then examines a variety of application that illustrate the potential technologies that can be enabled by appropriate breakthroughs in cognitive vision.

This research roadmap documents innovations required to advance from a science of visual reconstruction to a science of visual perception where the machine can develop an understanding of its environment and thus behave adaptively, where it can reason about what it sees, anticipate events, and learn from its behaviour.

In Section 1, we define the scientific discipline of cognitive vision within the structure of established scientific disciplines and delineate boundaries and relations with related domains. We begin with a definition of cognitive systems. We then define the sub-dmain of cognitive systems as a convergence of computer vision and cognitive systems We review relations to the parent domains of computer vision and artificial intelligence. We conclude with a summary of enabling technologies that are driving progress in cognitive vision.

In section 2 we draw from the parent domains of artificial intelligence, cognitive systems, and computer vision to synthesize an ontology of fundamental concepts. These concepts enable the definition of a glossary defining the technical vocabulary for the doimain.

Section 3 motivates the reseach roadmap by describing the innovation process. In particular we examine the nature of innovation and the potential for innovation in cognitive vision.

Section 4 surveys applications and potential markets. For each survey, we identify fundamental scientific or technological problem whose solution would enable advancement. The production of these surveys has been used as a tool to verify and complete the ontology as well as the list of enabling technologies. These surveys illustrate the ontology and need for the enabling technologies.

Section 5 summarizes the open fundamental problems identified in section 4 and describes the potential impact of progress.

The key question this document seeks to answer is: what should we be doing to provoke the scientific rupture that will lead to strong take-up of cognitive vision and exponential growth in the dependent application domains? The answer is a set of fundamental research issues that complement existing computation vision in order for it to achieve the required innovative functionality. Ultimately, it is about creating new theories and methods, what Marr called computational theories, algorithmic formulations, and implementation realizations, that will allow us to do existing things in new more effective and more efficient and, more importantly, that will allow us to do new things altogether.

1. The Domain of Cognitive Vision

1.1 Cognitive Systems

Cognitive Systems is the study theories and methods for the construction of artificial systems that exhibit intelligent behaviour. While cognitive systems may derive inspiration from biological intelligence, Cognitive Systems is a separte scientific discipline, concerning artificial systems [Simon 87] that combine perception, action and reasoning. As a scientific discipline, Cognitive Systems seeks to provide an enabling technology for robotics and automation, natural language understanding, man-machine interaction and complex systems. However cognitive systems is not about applications in any of these domains.

The goal of the EC research programme in Cognitive Systems is to create and develop a scientific foundation that will apply across these and other domains of engineering science. The starting point for this program is a view of intelligence as rational behaviour. Rationality is the ability to choose actions to accomplish goals. To be intelligent, a system must be able to act, to have goals, and must be able to choose actions that attain its goals. Cognition, or knowledge, is the means by which intelligent systems choose actions. Research in cognitive systems seeks to develop theories and methods for constructing intelligent systems that can learn, reason, know and explain.

Past attempts to provide a formal foundation for cognitive systems have relied on symbolic logics. A purely symbolic foundation for cognitive systems has been found to have limited utility because of the problem of grounding (providing meaning) for symbols. Purely symbolic systems provide only syntactic processing. Semantics (or meaning) requires perception and action. Without semantics, symbolic systems have no basis for learning or for common sense reasoning. Such systems can be made to imitate intelligence. However, they do not exhibit the generality that is characteristic of intelligence.

Cognitive systems will require convergence of action, perception and reasoning. Action, taken in a broad sense, provides the foundation for semantics. Actions may involve applying and controlling the energy to a mechanical device. They may have

the form of communicating or recording a symbolic description of the world. Indeed, generating natural language is a highly desirable form of action. Actions may also have the form of changes to the internal state of the system, such as a change in focus of attention, with no immediate external manifestation.

Perception is the interpretation of sensory input. Cognitive systems must be perceptually enabled in order to generate appropriate actions and behaviours. In its most sophisticated form, perception provides a model of a situation that enables reasoning. However, perception may also directly result in the selection of behaviours, the execution of actions, a change in focus of attention.

Reasoning coordinates perception and action. Such coordination can occur over multiple time scales and at multiple levels of abstraction. At the lowest level of abstraction, a system exhibits reflexive behaviours in which sensory signals are mapped directly to actuator controls. Reasoning at this level selects and regulates the transformations. At intermediate levels, compositions of actions or behaviours bring the system and the world to desired state. Reasoning may be used to select and apply a predetermined plan of action. Reasoning may also be used to adapt an existing plan or to generate new plans to attain a goal from a new situation, or to attain a new goal.

To be general, a reasoning system must be able to form and exploit new symbolic concepts. The system must learn. Such learning is not restricted to recognition and categorization. It extends to automatic acquisition of perception-action cycles, to parameter control and to formation of abstract concepts and behaviours. The ability to learn perception-action cycles, to learn control and coordination of perception-action, to learn procedures to accomplish goals, to learn new concepts, and to learn and improve new plans of actions are all important problems for cognitive systems.

1.2 Cognitive Computer Vision

Cognitive Computer Vision is the study of the acquisition and use of knowledge and reasoning in vision. Cognitive vision represents a convergence of computer vision and cognitive systems. As a sub-field of cognitive systems, cognitive vision seeks to provide the perceptual component necessary for rational intelligence. As a sub-field

of computer vision, cognitive vision seeks to evolve computer vision from the science of visual reconstruction to a science of machines that see.

Seeing requires abilities to use visual perception to know, understand and learn about an environment and, possibly, to facilitate adaptive and/or anticipatory interaction with that environment by the perceptual system. Thus, cognitive vision implies capabilities or functionalities for:

- **Recognition and categorization**, *i.e.*, mapping to previously-learned or *a priori* embedded knowledge bases. The majority of vision systems today rely on recognition or classification to attain their goals. This makes them inherently application-specific and quite difficult to re-use. Systems are able to recognize instances of particular objects or events (a car or type of car, a table or type of table) rather than being able to identify objects or events as instances of a meta-level concept of car (or road-vehicle) or table (or object-supporting platform). Cognitive vision systems would ideally have this categorisaton capability. It is, however, a difficult issue because objects of the same category can have completely different visual appearances.
- **Knowledge representation** of events and structures. Ideally, the representation should exhibit some form of object or category invariance with respect to events and/or vision system behaviour. Many of the representations in the past have either been too application-specific to be adaptable to general (*i.e.* unanticipated) scenarios or environments or they have been too abstract to be applied in any meaningful way.
- **Learning**. There are at least two aspects to learning. First, there is *learning to see* (*i.e.* learning about the perceived visual environment.) Second, there is *learning to do* (*i.e.* learning to act and/or learning to achieve goals). Other forms of learning may also be necessary in a fully-functional vision system.

- **Reasoning** about events and about structures (or some unified combination of both). One might distinguish three types of reasoning: reasoning to facilitate learning, reasoning to facilitate recognition/categorization, reasoning to facilitate hypothesis formation (*e.g.* ‘what-if’ scenario evaluation to aid system planning).

- **Goal specification**, *i.e.*, identification of required system behaviour (this is the very essence of application development). Goal specification does not mean simply identifying the required information transformation from sense data to symbolic description – it may well include this but this in itself is inherently insufficient for a cognitive vision system which will typically have a number of often conflicting goals.

- **Context Awareness (Focus of Attention):** Perception requires context. Context determines the entities to observe and the properties and relations to measure. Context provides a means to focus attention in order to observe the what is important to the current goals.

Cognitive vision may also require some form of embodiment, autonomy, articulation, or agency. The need for embodiment and autonomy is an open question of exactly the kind that Cognitive Vision will address.

Cognitive vision is a multi-disciplinary area, drawing for example on the disciplines of computer vision, artificial intelligence, cognitive science, and control and systems engineering. It is ultimately intended as a complement to conventional computational vision to facilitate an overall system functionality that is adaptive, anticipatory, and, possibly, interactive.

1.3 Cognitive Vision and Computer Vision

1.4 Cognitive Vision and Artificial Intelligence

As a scientific discipline, Artificial Intelligence (AI), encompasses numerous research areas which are relevant for or even overlap with Cognitive Vision. This is reflected by the topics of large conferences such as IJCAI, AAAI and ECAI. In its early years, Artificial Intelligence was understood to include vision: As early as 1955 Selfridge proposed vision as a task integrated in a cognitive context [Selfridge 55] and interacting with other cognitive processes. But vision research was in its infancy then,

and a much narrower view of the vision task on the one hand and of artificial intelligence on the other had to be pursued for several decades.

An obvious theme of common interest and indeed one of the fundamental themes of AI is the acquisition, representation and utilization of common-sense knowledge. In vision, we know that humans interpret what they see in the light of diverse knowledge and of experiences about the world. The scope of this knowledge can best be seen when we consider silent-movie watching as a vision task, for example, watching and understanding a movie with Buster Keaton. If a vision system were to interpret the visual information of such a movie in a depth comparable to humans, the system would have to resort to knowledge about typical (and atypical) behaviour of people, intentions and desires, events which may happen, everyday physics, the necessities of daily life etc.

The same holds true for several other subareas of AI, e.g. natural language understanding, robotics, planning, diagnosis, information management and others. Human-like performance seems to depend decisively on the width and depth of common-sense knowledge which is available for a task. Of course, there are domains and tasks where limited, specialized knowledge suffices, but dealing with common-sense knowledge in its widest sense remains a fundamental challenge of AI.

We will review now major research areas of AI pertaining to the acquisition, representation and utilization of common-sense knowledge and comment on their relevance for Cognitive Vision.

One of the most interesting endeavors of AI in the area of common-sense reasoning is naive physics, a term coined by Pat Hayes in his seminal paper "The Naive Physics Manifesto" [Hayes 79]. The goal is to develop a theory for common-sense physical inferences, e.g. about the behaviour of fluids, objects moving in gravity etc. It is evident that such a theory could be very useful for a cognitive agent interacting with the everyday world. For vision this would mean, for example, that educated predictions about the development of a dynamic scene or inferences about unvisible parts of a scene may be generated.

The interest in naive physics has led to extensive research in diverse subspaces of the physical world [Weld and de Kleer 90]. The common thread is to model physical phenomena with qualitative as opposed to quantitative laws. A qualitative theory is based on symbols, each typically corresponding to a class of physical situations. Qualitative physical descriptions abstract from numbers and provide a level of abstraction useful for higher-level understanding of scenes in Cognitive Vision.

The need to relate quantitative visual data to symbolic descriptions has often been addressed as a particular challenge, even as an unsurmountable obstacle preventing the application of qualitative AI theories to vision. This is not the case as work on interfaces between vision and language show [Badler 75, Neumann 89] where symbols are grounded by well-defined mappings into quantitative subspaces.

For Cognitive Vision, qualitative spatial and temporal theories are of special interest. Such theories may provide a suitable language for the description of complex, temporally extended scenes, both at the conceptual and factual level, as well as reasoning facilities based on this language. Allen´s interval algebra is a popular scheme [Allen 83] for reasoning with temporal interval relations. A convex time point algebra which also includes quantitative expressions has been employed for event recognition in [Neumann 89]. Vila [Vila 94] provides a survey of qualitative temporal theories.

The goals of Cognitive Vision overlap to a considerable degree with Cognitive Robotics. The term "Cognitive Robotics" has been created by Ray Reiter for research which is concerned with "the theory and the implementation of robots that reason, act and perceive in changing, incompletely known, unpredictable environments. Such robots must have higher level cognitive functions that involve reasoning, for example, about goals, actions, when to perceive and what to look for, the cognitive states of other agents, time, collaborative task execution, etc. In short, Cognitive Robotics is concerned with integrating reasoning, perception and action within a uniform theoretical and implementation framework" [Reiter 96].

Reiter adheres to a strictly logic-based approach (the situation calculus) for Cognitive Robotics, and it is justified to question its applicability to Cognitive Vision. Indeed, there appears to be overwhelming evidence that a probabilistic approach may be better suited to deal with the inherent uncertainties in vision. However, Reiter´s work on reasoning with uncertainty (among others he introduced default reasoning) and more recent work on the combination of logics and probabilistic reasoning show that the two worlds of "scruffy" and "clean" reasoning may be brought together.

As Cognitive Vision strives to exploit common sense, knowledge representation comes into play. This is a well-understood core area of AI where valuable insights are available about the expressiveness and inferential power of representation languages. Deriving implicit knowledge by inference engines will gain importance in vision as the task domain becomes richer and involves a larger body of knowledge. Recently, Description Logics have been shown to offer interesting compromises between expressiveness and inferential complexity [Lutz 02]. A Description Logic has also been proposed as the inference engine for the Semantic Web, an effort to turn web information into a machine-readable knowledge base.

A word of caution is in order, however, concerning the use of logic-based inference systems for the vision task per se. It is well-known [Reiter and Mackworth 87] that image interpretation can be formally described as partial model construction ("model" in the logical sense). Hence standard inference services such as classification or subsumption are not of immediate value. Rather, evidential reasoning processes are required which are not (yet) part of formal knowledge representation.

Learning is another area of AI research that is eminently relevant for Cognitive Vision. Visual behaviour may be predicted from a spatio-temporal context, models may be determined from the statistics of a large number of observations. Various forms of learning have been used for a long time in low-level vision, more recently also for high-level tasks such as learning of traffic patterns [Fernyhough et al. 98], see Buxton for a recent survey [Buxton 03]. In Cognitive Vision, learning at a high level of abstraction will become important both in a supervised and unsupervised form in order to establish the experience required for high-level scene understanding. AI

offers a rich repertoire of learning methods for such tasks. As a prominent example, concept learning using the version space method in the framework of logic-based knowledge representation is one well-established methodology [Mitchell 97].

In the last ten years, however, learning in a probabilistic framework has gained increasing importance due to the introduction of Bayesian Nets (in AI also known as belief nets). Bayesian Nets provide a sound basis for evidential reasoning which may be applied to diverse subtasks in vision. Moreover, recent work on combining logic-based and probabilistic approaches, for example the work of Koller on Bayes Nets in connection with relational structures [Getoor et al. 01], indicate that the gap between the quantitative world of classical Computer Vision and the symbolic world of classical AI is rapidly closing.

One of the research goals of Cognitive Vision is to be able to interpret scenes in terms of intended actions or plans, to ascribe intentions to agents, and to infer goals. This is closely related, of course, to planning, one of the oldest and best established subareas of AI, for a representative collection see [Allen et al. 90]. Research on planning in AI also includes plan recognition, for a probabilistic approach see [Charniak and Goldman 91], for multi-agent aspects see [Huber and Durfee 95]. Hence there is a rich body of methods to build on. In addition, however, the vision task of interpreting observations as intended actions also requires a deeper understanding of causality at the physical level. When can one conclude that two occurrences are causal? For recent AI research into this topic see [Pearl 00].

In summary, it is fair to say that AI, in particular common-sense reasoning, and vision, although early recognized to belong together, have not yet been integrated to a significant extent. The main reason seems to be that Computer Vision and AI have gone separate ways in order to obtain sufficient maturity in their own respective disciplines before attempting an integration. But given the substantial relevant work outlined above, time has definitely come to establish Cognitive Vision as a research program which brings the two together.

1.5 Enabling Technologies

The roots of cognitive vision lie in the explosion of interest in human perception which resulted from developments of artificial intelligence in the 1970's. However, cognitive vision is only now beginning to realise its potential, based partly on developments in our scientific understanding but also, crucially, on improvements in imaging, information and communication technologies.

Technology is driving the development of cognitive vision systems in two ways: firstly as an enabler for cognitive vision, such as the enormous improvements in cameras allowing for cost-effective acquisition of high resolution digital images; and secondly in providing new applications and markets for cognitive vision exemplified through the demand for content re-use over multiple media, which is itself a consequence of the development in broadcast and internet technologies.

A number of key technology drivers can be identified to illustrate why the renewed focus on cognitive vision is timely. These technologies are enabling in the sense that they make technically possible aspects of cognitive vision systems and that they have reached a level of performance which is also acceptable in economic terms. Although it can be argued that many of these technological advances have had beneficial impact on all forms of computer vision, it is the combination of these drivers, combined with new approaches to artificial intelligence which are influencing the development of *cognitive* vision systems.

1.5.1 Image acquisition devices

Core to all vision systems is the ability to acquire, store and where necessary display images to the user. Over the past twenty years, solid-state devices such as CCD (charge coupled device) and CMOS (complimentary metal oxide semiconductor) have ousted tube devices, although for many years, consumer-grade chips were tailored to provide signals in the analogue television formats (NTSC, PAL, SECAM) usually digitised to provide a spatial resolution of ~0.3 M pixels and 3 x 8 colour bit depth (65k colours). Cameras which acquired images over different aspect ratios or at higher resolutions were available only at high cost, and were targeted towards specialist applications such as astronomy. In 2003, the cost of a several mega-pixel

CCD chip is less than 0.5 USD, and there have been corresponding decreases in the price of readout electronics and optics (lenses, colour masks etc). The consequence for cognitive vision is that cost of acquiring images is relatively low and hence there are a greater number of images available, but also that the *information content* of these images is much higher.

1.5.2 Computing devices

In parallel with the general trends in semiconductor manufacture which have enabled the production of high quality, low cost imaging devices, computing devices (processors and memory) have also improved hugely over recent times. This has applied equally to desktop, specialised research machines and embedded systems. As an example, a desktop user in 1993 could expect to be using a 486 PC with 32 Mb RAM and low 100s Mb disk space. The processing power, and volatile and non-volatile storage available to today's user is two orders of magnitude greater. This power is also more *accessible* to the cognitive vision system designer – as well as the improvements in hardware, software tools have also simplified the implementation of complex algorithms.

1.5.3 Networks

In addition to the obvious improvements in bandwidth provided by advances in digital systems and photonics, and which enable more rapid and frequent transmission of high information content images, improvements in networking technologies have had some specific consequences for cognitive vision. In particular, the ubiquity of computer networks, their increased spatial coverage (fixed cable and wireless) and standardisation (primarily in internet protocols) are making new applications of cognitive vision possible. For example, the seamless use of mobile (wireless) and fixed (cabled) cameras in implementing a monitoring system which needs to track an individual over a wide spatial area. Access through networks to the enormous processing power organised in the GRID will also be a factor in the viability of some potential applications of cognitive vision.

2 Fundamental concepts for Cognitive Vision

Ontology is the branch of philosophy that deals with being [Morris 69]. In philosophy of science, ontology refers to the irreducible conceptual elements in a logical construction of reality [Cao 97]. In other word, an ontology provides the set of logical elements from which theories and models are constructed. For example, in physics, rival ontologies propose particles, forces, or space-time as the fundamental elements of theories and models.

An ontology provides a formal definition of the domain of discourse for a scientific discipline. The theories and analytical methods of a domain are expressed in terms of an ontology. Without an ontology, there are no theories or analytical methods. However, scientific ontologies evolve dynamically. Unless a domain is mature, no ontology is perfect. Admitting rival ontologies is pivotal for a developing science. Such competition allows competing approachs view points to develop and to cross fertilize.

A research roadmap must allow for this multiplicity of ontologies and, hence, the roadmap must be presented at the level of a meta-ontology: not a single ontology of a cognitive vision system per se, but in the context of a class of ontological entities or concepts. Thus we develop our ontology as a hierarchical structured glossary of concepts, theories, methods and representations.

The following provides an ontological foundation for cognitive computer vision. This ontology is a logical construction of concepts that allows us to make well defined statements about the problems and methods that define cognitive vision. Because cognitive vision is an emerging discipline, any proposition of any ontology will elicit controversy and debate. Such controversy is healthy.

We start with what we believe are terms that are unambiguous for any experienced engineer or scientist in informatics. These are concepts that are fundamental to any measurement system. From these we progressively move to concepts that are specific to vision and artificial intelligence.

We begin our ontology with an assertion that there exists an unknowable and unique reality. This reality can only be partially observed. The task of any cognitive agent or machine that attempts to see is to estimate a partial description of this reality from partial and incomplete perceptions. In modern terms we say that reality is "sensed" by a sensor (or transducer). A sensor produces a measurement based on reality. A measurement from the sensor is the basic input into the sensing system.

A measurement is the simplist and most primitive form of "observation". Observations are the outcomes of an "observational process" performed by an "observer". The observation is the "answer to a question to nature" that is posed by the observer. The observation process can be based on measurements or on observations derived from measurements. Although the result is typically numeric it may alternatively be symbolic or binary.

Time is a unique measurement whose values are derived from an ordered directed sequence. Time is a measurement provided by special sensor called a clock. Any other observation may be associated with a time. A point is an entity described by a vector of N observations (N is an integer). One of the variables may be time. A space is a set of points. A space can be ordered along each of the N component variables.

Predicates are conclusions from observations. A predicate is a function that produces a truth value. The truth value may be binary(t/f) or probabilistic [0,1] or any other form of belief function.

A rational definition of cognition and intelligence requires is based on the concept of state as used in Artificial Intelligence [Korf 89]. We note that the use of the word "state" poses problems because of different usages in the neighboring domains of estimation theory and control theory. In those domains, "state" is a characteristic vector of estimated observations. In cognitive systems we adopt the usage from Artificial Intelligence that a state is a logical proposition defined over a set of observations. A state is defined by a predicate.

An entity is an associations (or correlation) of observations. Entities may represent objects, or concepts, or classes, or instances of classes. Entities may represent physical objects (This-pen) or phenomena (night). Entities allow the system to make propositions about the state of reality. Entities are defined by predicates (Boolean or probabilistic). The predicate may be defined by extension (by explicitly listing members of the class) or by intention (by an acceptance test).

A property is a measurement associated with an entity. Relations are predicates (Boolean or probabilistic) defined over entities and their properties. A structure is a set of entities and their relations. A situation is a kind of structure.

The identity of an entity is value that is unique for each entity. An identity variable is produced by an identification function. For example, an identification function may return the same identity variable for an entity that has the same position in successive images. Identification functions may also be constructed based on complex predicate expressions. (The blue box with the red spot).

A category is an observable variable whose value is unique for a set of entities. The set may be closed (entities A, B, C) or open (green entities).

Recognition is the ability to determine the identity of an entity. Classification (or categorization) is the ability to determine if an entity belongs to a category.

Knowledge is the ability to predict or specify situations. Learning is the ability to acquire knowledge. Reasoning is the ability to generate knowledge from descriptions of other knowledge.

3. The potential for innovation in Cognitive Vision

Predictions about the emergence of new technologies are notoriously inaccurate. For one thing, innovation is a cumulative process, resulting in exponential changes. Humans are surprisingly poor at predicting exponential processes [Gladstone 00]. Human intuition tends to rely on linear extrapolations, resulting in predictions that are often over optimistic on the time scale of a year or two while pessimistic for scales beyond 5 years. More-over, the events that trigger a cumulative exponential process are unpredictable, difficult to discern and difficult to estimate.

A number of theories about the innovation process have emerged over the last few decades. These theories draw from the roots in econometrics and philosophy. Authors such as Mowery [Mowery and Rosenberg 98] and Utterback [Utterback96] have documented the emergence of a diverse array of technologies and synthesized theories that lead to general phenomena and predictable phases. This section reviews some of these phenomena and their possible sequences in order to establish an objective foundation for rational policies for triggering or accelerating innovation in cognitive vision.

3.1 The nature of innovation

In order to make statements about innovation we must have some objective properties with which to characterize innovation. The primary objective properties used to characterize innovation are performance metrics. These are properties that can be measured in an objective and reproducible manner independent of a specific technological implementation. An essential step in defining a technology is elucidating the performance metrics that can be used to characterize the technology. Without performance metrics it is impossible to make statements about a technology.

Innovations are ideas that permit improvements in performance. Such ideas may have many forms. In the following, we will describe several forms of innovations. We illustrate these forms using examples from the domain of electricity in the 19th century.

Innovations may have the form of an ontology of concepts that enable analysis. For example, statements about electricity required an ontology composed of such

concepts as voltage, current, resistance, capacitance, and inductance. Each of these concepts is based (that is, grounded) on observable measurements (performance metrics). The emergence of this ontology, and the accompanying metrics, made it possible to reason and communicate about electric circuits.

Innovations may also have the form of theories that predict and explain phenomena. For example, Faraday discovered a relation between magnetism and electric charge. This relation made it possible to design a plethora of electrical devices. However, the theory could only be stated formally once the ontology and its measures were established.

Innovation may have the form of methods of analysis. For example, analysis of electrical circuits were made possible by Kirchoff's laws for current and voltage. Analysis of communication systems was made possible by the introduction of Fourier domain analysis by Nyquist, Shannon, and others.

Innovations may have the form of the design for a device. Faraday's law made is possible to design devices that converted between kinetic energy and electric potential (motors and generators). Over the years, many new forms of electric motors and electric generators have been invented, refined and exploited. Each motor can be compared to others based on a common set of measurable properties such as efficiency or torque.

Innovations may have the form of systems architectures composed of multiple independent devices. Thomas Edison's creation of an electrical power distribution system was a revolutionary innovation that can be measured in terms of number of electrical devices constructed or total electric power generated .

Innovations often have the form of improvements to existing devices or architectures. George Westinghouse surplanted Thomas Edison by transmitting power using alternating current. The improvement in transmission efficiency led to important economic advantages.

In summary innovations may have the form of a proposal for or an improvement to an ontology, a theory, a method, a device, or a system. In order to map the domain of Cognitive Computer Vision, we must identify the ontology for the domain, and catalog the existing theories, methods, devices and systems.

3.2 The virtuous cycle of innovation.

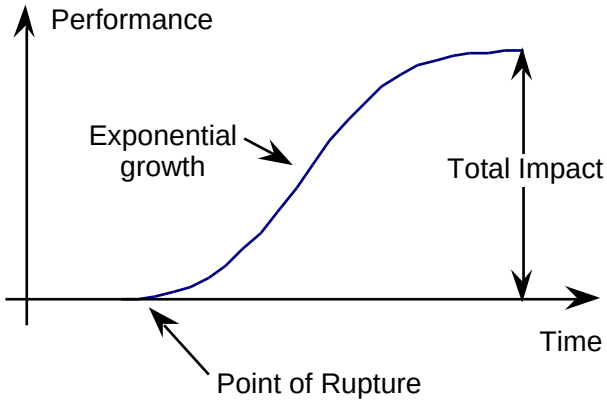

Figure 1. The S Curve can be described by three parameters: rupture , exponential growth and total impact.

The interaction of innovations tends to multiply performance. As a result, the accumulation of innovations leads to exponential growth in performance. The exponential nature of such growth in peformance has been widely documented throughout the history of human technology, and is widely known as the "S" curve, shown in figure 1 [Rogers 95], [Dent 93].

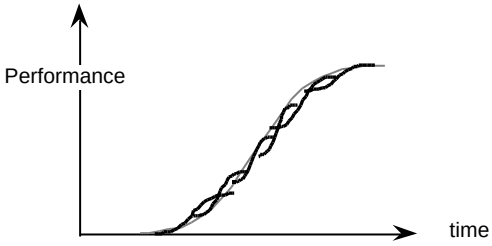

Figure 2. An S curve is the result of the composition of smaller S-curves.

The "S" curve is, in fact, a cumulative Gaussian distribution resulting for the accumulation of a multitude of interacting innovations. Each of these interacting innovations provides its own smaller S - curve, as shown in figure 2. The exponential growth in performance is the result of a virtuous cycle composed of the interaction of research, innovation and economic impact. In this spiral, the promise of economic

gain leads to allocation of resources for research. Research produces innovation. Innovation triggers generation of wealth. The resulting wealth provides resources for research. All three components are fundamental. As long as the three components (or phases) of the cycle continue, innovations will accumulate and performance grows exponentially.

The virtuous spiral breaks when any of its three components attains a limit. For example, saturating the potential for generating economic benefit will result in a decline in the economic impact of innovations, resulting in a decline in resources for research, resulting in a decline in innovation. In some cases, the decline is due to a physical limit. For example, commercial air travel is currently restrained by the cost surpassing the sound barrier. While technologies exist for flying faster than sound, there is little economic incentive. As a result EADS, Boeing and their competitors concentrate on decreasing the cost of air travel.

The impact of a line of research can be described by the parameters of its accompanying S curve. For measuring past innovation the three most convenient parameters might be the growth rate, impact, and date.

The growth rate is the exponential coefficient that characterises the growth. For example, Gordon Moore has observed that transistor density in Integrated Circuits doubles every 12 months, leading to a double in the computing power on a chip every 18 months. An important measure of the impact of innovation is the discontinuous change in the exponential growth rate. Such a change is called "rupture". The increment in growth rate is an important measure of the impact of innovation or of a public policy decision.

The incremental growth rate is a direct result of the economic return on investment in research. Such growth rate is conditioned by a number of factors, including the propensity of the economic and administrative climate towards innovation. One property of innovation that interests us here might be called the "power" of the innovation. By this we means the effect that the innovation has on the efficiency of research. The term "power" corresponds to the coefficient of exponential growth. A

"powerful" innovation enhances the efficiency of research on the design of devices and systems, thus leading to more rapid return on investment and a greater exponential growth.

The impact is the total growth in performance that can be achieved. For example, the introduction of jet engines led to improvements in the cruise speed of commercial aircraft from 300 to 600 knots. The duration of an S-curve is determined by the exponential coefficient and the impact. For a Cumulative Guassian, this is measured by the second moment (or standard deviation) of the derivative (the Gaussian density function). The impact is an important measure for choosing between competing public policy decisions.

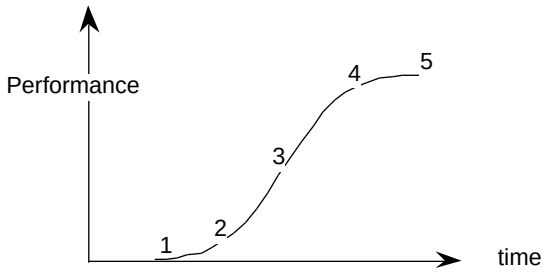

Figure 3. Phases of technology innovation: (1) rupture, (2) early development, (3) expansion, (4) maturation and (5) saturation

3.3 The phases of innovation

The S curve permits the process of technology innovation to be described in terms of 5 phases, as shown in figure 3. The phases tend to determine the nature of the innovation. For example, rupture tends to result from a new theory or method, often made possible by a new ontology or the emergence of an enabling technologies. Early development tends to be characterised by refinement to ontologies, theories and methods, and by innovation in the form of new devices or systems. Expansion is generally characterised by the emergence of a dominant design for devices or systems, and refinements to component devices or technologies. Maturation and saturation mark a period of diminishing returns on investment in refinements to devices or systems.

The above concepts allow us to make some simple recommendations about public policy on research funding. These recommendations are derived directly from the role that this can have in generating innovation.

a) Sustained exponential growth requires economic benefit. Except at times of national crisis (such as a military emergency), public funding can only maintain an S curve when accompanied by exponential growth in tax revenue. Investment in refinements to devices or system are rarely an effective use of public funds, and should be left to commercial investment. This implies that public funding should not be used in phases 3, 4, and 5 (of the S curve) of innovation.

b) An S curve requires perceived economic impact. Investors use research as a tool to multiply money. Such multiplication is only possible when investors perceive potential for economic impact. Demonstrating potential for economic impact in order to attract investment is a reasonable use of public funds. Investment in new devices or systems can help bring a technology from rupture to early expansion. Funding for work in phase 2 has impact when it when is serves to document, develop or demonstrate an innovation in phase 1.

c) The most powerful innovations are often accidental and a nearly impossible to predict. They are also hard to recognize As a result, innovations with the greatest potential for impact are in what might be called "fundamental science". Hence, public funding should be concentrated in phase 1. However, not all fundamental sciences have the same potential for impact and rupture.

We note that these prescriptions are perhaps too dogmatic. For example, attempting to build a demonstration of a new technology (phase 2) can often trigger the creation of new enabling innovation (phase 1) out of desperation.

The most powerful innovations are those that result in new theories and methods. Theories and methods enable the development of new more effective and more efficient ways of doing existing things and, more importantly, they enable the development of ways of doing new things. Such innovations require a proper

ontological foundation. Public funding of the creation of an ontological foundation and the discovery of theories and invention of analytical methods is easily justified when then these activities enable a rupture in performance in a domain with substantial economic impact. The difficulty is that the time of rupture, and the increment in exponential growth rate are hard to predict.

4. Applications and Potential Markets

In developing this research roadmap, we began by generating a set of application scenarios, the so called "research dreams". These scenarios were used to develop and to illustrate the ontological concepts, and to illustrate the theories and methods that must be developed. Within each applications domain, one can identify the key issues, be they theoretical, algorithmic, or implementation/technological support.

The authors of each of these sections was askt to provide up to 4 pages describing:

1. What exists today?
2. What is the currently demand?
3. What is the long-term dream?
4. What are the fundamental research issues?

4.1 Autonomous (Mobile) Systems and Robotics

Autonomy of systems, in particular robots, is required to obtain automatic responses to situations and flexibility in solving tasks. Autonomy demands a sense to perceive and understand the environment and to react to it. Service robots exhibit today the largest degree of autonomy. While most systems operate in laboratories, a few systems are on the market.

Mobile systems can operate in in-door environments using laser range sensors for applications such as cleaning and goods delivery in Hospitals (e.g. SINAS or HelpMate). Small robots use sonar sensors for carpet cleaning (e.g. Kärcher, Electrolux). Open sensing problems are related to reliability in more complex environments and navigation in offices, e.g. seeing thin table legs. Navigation in populated areas is in development (e.g. for wheelchairs). Out-door navigation has been successful to steer cars on highways or to navigate mobile platforms in planetary environments. The first out-door automated guided vehicle (AGV) is used for automated container transport at harbours and uses millimetre wave radar (ACFR). Industrial in-door use of AGVs is in logistics to deliver goods for manufacturing, often on fixed paths and in few cases with active obstacle avoidance. In all these systems navigation is based on a 2D representations of the environment.

Existing systems using vision employ fast 2D recognition capable of determining size and location of a known object type. Often a network of cameras is used to control the handling systems on a production line. The set-up of the tasks can be done by integrators with a graphical user interface where ease of use is an important factor for moderate volume markets (2000 units pa). A first attempt on extending 2D methods is shown for space robotics, where clearly defined objects simplify visual data extraction. The basic control method is tele-operation with added visual servoing capabilities. Coupling tele-operation with visual sensor information is also used in tele-surgery, space or underwater systems. Industrial automation extended the standard 2D approach to 3D for cases where a very small correction of part location is needed and the visual search space is limited to a fixed area on the large part (e.g. ISRA).

Another important function to bring robots to more applications is the ease of use. Present interfaces rely on teach panels, a space mouse or similar input devices. Human gaze direction can be obtained (e.g. EyeBlink) and first human tracking systems enable user interaction in designed environments (e.g. MIT, NTT, Siemens). The person speaking is identified and a about ten gestures can be interpreted. In a specific domain, such as the American Sign Language, a full alphabet can be detected.

Besides navigation, service tasks and autonomy typically require a manipulation capability. The existence of the humanoid robots in Japan (e.g., Honda, Sony) creates new potential for usage, though presently these robots are pre-programmed. Work in America is leading towards artificial creatures (e.g. Breazeal, MIT, but also Sony) that imitate behaviours for learning and can physically interact with humans. Emphasis is on the control of many behaviours and the observation of human emotions.

Complimentary products evolve at rapid speed in the toy and edutainment market (Sony, Honda, …). Presently simple sensory modes and motion behaviours are used, e.g., dog following a coloured ball. Robot soccer has become an entertaining domain

to study distributed system and their coordination. Again, colourful patches are of basic need and robustness of vision as well as processing speed impose limits.

Current demand

Reliable navigation in-door as well as out-door

Several aspects block a wider application of autonomous systems. Maps should be learned and not predefined. Recognition of hallways and full boards in shops is good but needs to be extended to cope with chairs, tables and empty boards. This also requires reliable and sufficiently accurate object localisation.

Object localisation

Flexible manufacturing and service tasks follow the market demand of mass customisation. A roadblock to small lot sizes are versatile processes that can be used for a large variety of parts and the simple introduction of new variants. This creates a demand to move from present 2D perception towards locating and recognising objects with 3D visual sensing. It requires to learn objects from a few samples and to cope with geometrically varying work pieces and products. The extension is to handle and inspect the correct functioning of parts and products. It finally requires to cope with the large variety of lighting situations.

Intuitive Interaction Devices

For the applications listed above, but also to operate industrial robots, more natural and intuitive interfaces could highly relief the worker or person from learning a complex interface. Interfaces should be based on several sensing modalities (e.g., vision, speech, tactile) and exhibit high interpretative abilities to perceive the user's intention or to respond with clear questions to obtain unambiguous information.

Assistive devises for elderly and handicapped

A major application of object localisation is the aging population which waits for means to assist persons, e.g. lift heavy loads, stabilisation of walking, and the partial support in part handling. This requires the system ability to adapt to the user, learn

and adapt to new situations and tasks, and cooperate with other automation equipment (e.g. home automation). It also builds on using intuitive interaction tools.

Range cameras and high-dynamics cameras

Reconstruction from classical camera images (CCD or CMOS) with 8 bit pixel resolution is a hard problem. Progress in stereo provides additional depth information, though dependent on image quality and object texture. New range cameras (based on active sensing principles such as laser or time of flight) provide much more accurate depth information. Integration of range with colour information is expected to increase the robustness of approaches immensely, thought this requires cheaper range sensors together with a better dynamic range and signal to noise ratio of CMOS colour cameras. Simultaneously, further improvement of computing power is needed and can be expected.

Modular and expandable system design

Cost is still a major factor in automation systems. For the diversity of application areas it will be necessary to build reliable modules and to combine these modules on demand. If modules could be produced at large quantities and only the required modules need to be put together, the system costs will reduce significantly. To enable such as step, interfaces and standards are critical and need to be defined by the main players.

Distributed processing and knowledge acquisition

With the modular structure distributed processes are realised, which is a core solution to provide scalable systems. It then becomes necessary to provide solutions for the module/agent interaction and reaction to the given task. This also requires techniques for distributed acquisition of the relevant information and knowledge, either from other components or from large scale data basis such as common knowledge or the www.

Other Demands

Further demands to mechanics, sensors, cooperation and communication are documented for robotics are seen as technology driver in related areas (see Robotics Roadmap at www.euron.org).

Emerging Technologies

A future vision is an autonomous system or robot with the capability to be a person's "best friend". It can perceive the mental state of a human to support her/him. E.g., give elderly people impact to do to something, to excite them not to become passive. It is seen like the good companion or the old wise grandfather/mother. It will require solutions to the following tasks, which can be also exploited in other applications (see below):

- be a companion: cheer up, find someone to talk to (not the machine itself, also another person), help when needed, suggest help when found necessary, help when detected as really needed
- learn favourites and assist (pre-selection for shopping clothes, …)
- do household chores: clean up, after cooking, after party, after child, keep order at state wanted and appreciated as comfortable
- basic daily needs: cook, look at recipe and shop (find ripe fruits and vegetables, edible mushrooms), assist during cooking, detect ready meat on oven, detect hairs in soup, …
- extension to aid handicapped persons, be "eyes of blind people" (a person's best friend with additional capabilities)
- cognitive abilities with link to world-wide data bases including common sense reasoning

Related applications that require solving similar cognitive vision problems are, for example:

- understand hand writing, hand drawings
- sport aids: hit baseball, golf (+ golf caddy), play flipper, teach player correct motions, …
- the perfect sport referee: detect fouls, incorrect play in all games, offside in football, ...
- recognise forged paintings, writings

Some cognitive capabilities required:

- detect non-objects (danger, fear, emotions, …)
- learn how to react and to handle different emotional states

- understand relations between humans and objects
- understand relations between two persons
- reason about environment and its functions,
- recognise dangerous situations (screw loose, hose open, pot at edge) and errors/faults
- separate good and not good actions (good in the sense of good for the person and the community rules)
- learn alternative behaviours dealing with the same situation (there is not only one way to solve a problem), this ability helps to decrease failures, additionally if a system is assisting a person maybe preferences for one solution are demanded (learn priority of alternative solutions).

An extension of the person's best friend is to take care of several persons respectively to aid the society at all. A Community Care Taking System could aid society by having a broad understanding of humanity. It could be a system that brings people together and makes sure nobody is harmed. This is a visionary role of a robot (or system) as a "human protector", similar to Giskard in Asimov's novels. (Though mind reading is probably beyond our planning ability.)

Example applications, starting from simple and leading to more complex scenarios:

- detect child seat on front seat to not unlock airbag
- intelligent traffic control switching on most demand
- detect waste, separate waste
- detect landmines
- detect unusual behaviour of patients under observation in hospital
- detect drunken people getting into the car, car does not start
- detect situations before they lead to accidents and react accordingly
- find thieves (while they are doing it, better stop them before)
- detect and catch insects

Research Challenges

Representations of Objects, Activities and Contextual Relations

Representations are often tailored to one specific method. What is needed are representations that are flexible to be used in several processes (segmentation,

learning, recognition, reasoning, …) as well as can describe the contextual relationships. This includes **representations of objects, activities as well as relations between objects and persons and the contextual relations.** Little attention has received the task to represent the functional use of objects and, hence, to know about the related set of actions that can be performed. First representations of the function of objects might be achieved at short time, the detection of the present functional use of objects taking into account contextual relations requires to solve fundamental problems regarding object classification and reasoning. Representing contextual and functional relations will be critical to realise what is a dangerous situation for an adult or a child.

Fundamental Research Problems

Learn, acquire and exploit representations of objects and activities

Applications demand that the set of objects and activities can be enlarged and learned in short time. We will need methods to automatically extract a representation from few presentations.

This extends to develop efficient techniques that can learn and *evolve* over extended periods, where the ability to obtain and cluster concepts from learning is essential for classification and reasoning tasks. Objects or activities must be linked to the human/operator experience to obtain a common understanding (grounding problem). Learning includes the relations to context and memory to be prepared to solve future tasks more efficiently.

Reliable detection and localisation of objects and Reliable interpretation of activities

Though cue integration and invariants improved robustness of vision techniques, there is still a lack of techniques that operate under real-world conditions (all lighting conditions, cluttered settings, partly occlusions, many objects, different views, different materials). One possible way to obtain reliable systems is to work with more cues and higher quality sensors (see Sensors), a second road is self-evaluation and active control of perception (see Performance Evaluation). Most likely work in all

levels of vision is needed to certify the reliability and dependability of individual methods and a system.

Formal Definition of Perception Capability and Performance Evaluation

Performance evaluation is in its infancy. While first results help, it is still unknown how to self-evaluation processing and to make clear judgements of the (intermediate) results obtained. At the minimum the accuracy and certainty (believe) of results needs to be characterised. This is even more complex when taking into account simultaneous tasks and environments. Methods are needed to know if a module works or not works, and in the latter case to ask another system module for assistance. Performance evaluation is the key to enable a formal description of the perception capability, which is will then enable to build large complex systems.

Control and Attention

Another open problem is how an appropriate action (control) can be deliberatively composed using the evaluation/judgement and contextual knowledge or any other (intermediate) results or knowledge from image data or a data base. This includes attention, in particular in more complex systems where multiple points of attention need to be attended to.

Interpretation of uncertain input

Perceptual input requires methods to handle the inherent uncertainty as well as derive interpretations that lead to actions to improve the result, this might include a reasoning about visual input that could help to avoid incorrect interpretation of data as is often the case today.

Adaptation to User Preferences

This means to realise that different users have different ways of communication, which also means to interpret the actions and to understand the intention of the user, which goes far beyond any present one-to-one relationship of a user interface to which the user needs to adapt. It also includes to recognise emotions and to correctly interpret the meaning of the emotion, to recognise and interpret sign language (or any

gestures), to find differences between cultures, and to automatically translate gesture for another cultural environment

Scalability, "soft" real-time, and architectures

Cognitive vision and perception, and in particular handling several tasks with one system, requires several concurrent/parallel functionalities, hence it requires means to distribute functions/tasks in a scalable manner and using an appropriate architecture and framework. Wide usage will require to design components/processes which can be deployed on demand, linked dynamically, developed independently, reused, whose performance is describable in detail, and whose characteristics/performance (accuracy, reliability, …) are known and certified. Finally, techniques are needed to resolve conflicts if introduced by goals or actions. Scalability refers to the method level as well as the system level. Scalable methods will also arrive at useful systems that can operate at usable speed. Hard real-time might be needed in fixed rate industrial systems, but many service or autonomous systems require reliable behaviour at the cost of reasonable additional time (e.g. navigating around obstacles or taking a second look to reliable grasp and object).

Integration

To build the actual systems, the components need to be integrated. Several basic problems require attention including problems such as how to rapidly assemble a large number of components, how to evaluate and certify the performance of the complete system, which standards to agree upon to enable integration at larger scales, and what techniques do assure targeted control of these complex systems.

Multi-modal Perception

Vision alone will not capture all information, hence, multi-modal perception systems will be needed for robotics applications as well as intuitive interaction systems.

4.2 Industrial Inspection and Industrial Robotics

Industrial Robotics (from Patrick)

Automation is now commonplace in nearly every industry, e.g. car manufacturing, food industry, packaging. An important concern is to achieve the benefits of

automation without giving up too much flexibility. As a consequence hard automation is not appropriate because of variation of the objects and of changing tasks.

The primary goal of an increased level of automation is not simply to replace personnel with industrial robots. Practice has clearly shown that particularly monotonous and un-ergonomic tasks are error-prone when performed by humans. The substitution of these tasks by automated systems leads to increased process repeatability and product quality.

One can consider a '**Robot Zone'** that has been extending across a range from hard automation where robot use is optimal to situations requiring significant flexibility but where for example, a high degree of repeatability or cheap labour is not available. One way of improving flexibility is to provide robotic systems with a means of sensing their surrounding. Packaging of consumer goods is a significant area where the object shape cannot be controlled for the benefit of the robot, but fast and cost effective handling is required.

Existing systems using vision therefore include those employing fast (flyby) **2D recognition** capable of determining size and location of a known object type. Often a network of cameras is used to control the handling systems on a production line. The set-up of the tasks can be done by integrators with a graphical user interface where ease of use is an important factor for moderate volume markets (2000 units pa).

What is the current demand

(from Patrick)

The ability to learn from a very small sample remains an important characteristic as batch sizes get small, requiring powerful **generalisation** ability.

An increasing important issue is the 'extraction' of the knowledge of the quality expert who knows intuitively what is a good sample for a range of classification grades, and who knows the production context. New **interactive user interfaces** capable of capturing this knowledge have to be found and implemented.

Finally the results should be used to establish a **closed loop control** to actually improve the quality of the production, where traditional PID-type control is inadequate for modelling complex non-linear effects. This is only possible on the condition that the images are fully understood in the context of the particular production process.

In robotics, **t**he main issues for further applications in robotics include improving the sensing ability of the systems by expanding 2D recognition to **2,5D or full 3D** capability to deal with more difficult and varied objects.

Complex non-linear closed loop control is important for improved flexibility, allowing tracking and correction of complex moving and misaligned components. Mechanisms that support **safe interaction** between human and robots are also becoming of increasing importance.

Current Technology

In recent years manufacturing industry has seen increasing levels of automation. Handling systems, automatic packaging and transportation systems are state of the art in many plants. One gap being closed is that of optical quality control. Some of the required tasks may be adequately solved by traditional image processing systems using physical measures such as size and shape. Sometimes however the quality being described is a more complex set of features such as colour stability or **aesthetic appearance**.

These problems occur for various kinds of products (textile, ceramic tiles, carpets, etc), in fact everywhere where a complex design has to be assessed. As in most other automation application, the aim of this task is to simulate the human ability of adapting to a new job. This requires the emulation of the aesthetic assessment according to human pre-attentive perception.

The particular difficulties of the task are the tight time scales in the production lines and lean production, high production rates, limited number samples available to set up of the system, and unknown errors due to complex production methods. In addition to

these more general challenges, surface inspection requires significant flexibility to cope with permanently changing designs due to demanding end markets. A typical system of this kind would be the Ceravision system.

Emerging applications

The ability to learn from a very small sample remains an important characteristic as batch sizes get small, requiring powerful **generalisation** ability.

An increasing important issue is the 'extraction' of the knowledge of the quality expert who knows intuitively what is a good sample for a range of classification grades, and who knows the production context. New **interactive user interfaces** capable of capturing this knowledge have to be found and implemented.

Finally the results should be used to establish a **closed loop control** to actually improve the quality of the production, where traditional PID-type control is inadequate for modelling complex non-linear effects. This is only possible on the condition that the images are fully understood in the context of the particular production process.

4.3 Video Surveillance

(Rebecca, James, David?? - with jim)

The Vigilant Environment - A Research Dream
David Vernon

4.4 Man-machine interaction

An Interactive Cognitive Vision System which collaborates with its users in image interpretation tasks,
…adapting to cultural, social, and task-dependent contexts of usage.

4.5 Smart environments and ambient intelligence

(David, and Jim)

What exists Today

What is currently demanded?

What are the fundamental research issues?

What are the future dreams

4.6Mapping on demand

(Wolfgang)

User specifies a map and in real time gets a map drawn from multiple sources of information. Problem - How does the system specify what he needs.

What exists

Slow

Spectification requires an experienced user. (Lack of interaction language)

No interpretation capabilities (neither GIS nor Image Sequ).

What is current demand

Demand will increase based on location based services (from GPS). and widespread introduction of information technology.

Long term perspective

Technology will be made ubiquitous (Homes, cars, public spaces, ...

4.7 Indexing Photo databases and Content analysis of images

The following text on photographic content analysis has been extracted from [Smeulders00].

4.7.1 The driving force

Content-based image retrieval came around quickly. Most of the journal contributions survive 5 year or less. The impetus behind content-based image retrieval is given by the wide availability of digital sensors, Internet and the falling price of storage devices. Content-based retrieval will continue to grow in every direction: new audiences, new purposes, new styles of use, new modes of interaction, larger data sets, and new methods to solve the problems.

What is needed most is more precise foundations. For most of the papers in CBIR it is not clear what reproducible problem is being solved or whether the proposed method would perform better than an alternative. A classification of usage-types, aims and purposes would be very helpful here including hard criteria for distinguishing among domain types. In spite of the difficulties intrinsic to early years, it is now clear that content-based retrieval is no longer just old wine in new sacks. It will demand its own view of things as it is our belief that content-based retrieval in the end will not be completely decided upon within the field of computer vision alone. The man-machine interface, domain knowledge and database technology each will have their impact on the product.

4.7.2 The heritage of computer vision

An important obstacle to overcome before content-based image retrieval could take off was to realize that is was not necessary to solve the general image understanding problem in its entirety. It may be sufficient that a retrieval system presents similar images, similar in some user defined sense. Strong segmentation of the scene and complete feature descriptions may not be necessary at all to achieve the similarity ranking. Of course, the deeper one goes into the semantics of the pictures, the deeper also the understanding of the picture will have to be, but that could very well be based on categorizing pictures rather than on a precise understanding.

We see applications of content-based retrieval in three broad types: target search, category search and search by association. Target search builds on pattern matching and object-recognition. New challenges in content-based retrieval are the huge amount of objects among which to search, the incompleteness of the query specification and of the image descriptions, and the variability of sensing conditions and object states. Category search builds on object recognition and statistical pattern recognition problems. New challenges in content-based retrieval compared to the achievements of object recognition are the interactive manipulation of results, the usually very large number of classes, and the absence of an explicit training phase for feature selection and classification tuning. In the search by association the goal is unspecified at the start of the session. Here the heritage of computer vision is limited to feature sets and similarity functions. The association process is essentially

iterative, interactive and explicative. Therefore, association search is hampered most by the semantic gap. All display and relevance feedback has to be understood by the user so the emphasis must be on developing features transparent to the user.

4.7.3 The influence on computer vision

In reverse, content-based image retrieval offers a different look at traditional computer vision problems. In the first place, content-based retrieval has brought large data sets. Where the number of test-images in a typical journal paper was well under a hundred until very recently, a state of the art paper in content-based retrieval reports experiments on thousands of images. Of course, the purpose is different for computer vision and content-based retrieval. And, it is much easier to compose a general data set of arbitrary images rather than the specific ones needed in a computer vision application, but the stage has been set for more robustness. For one thing, to process a thousand images at least demands software and computational method be robust.

In the second place, content-based retrieval has run into the absence of a general method for strong segmentation. Especially for broad domains and for sensory conditions where clutter and occlusion are to be expected, strong segmentation into objects is hard if not impossible. Content-based retrieval systems have dealt with the segmentation bottleneck in a few creative ways. First, a weaker version of segmentation has been introduced in content-based retrieval. In weak segmentation the result is a homogeneous region by some criterion, but not necessarily covering the complete object silhouette. Weak segmentation leads to the calculation of salient features capturing the essential information of the object in a nutshell. The extreme form of the weak segmentation is the selection of a salient point set as the ultimately efficient data reduction in the representation of an object, very much like the focus-of-attention algorithms for an earlier age. Weak segmentation and salient features are a typical innovation of content-based retrieval. It is expected that salience will receive much attention in the further expansion of the field especially when computational considerations will gain in importance. The alternative to work around strong segmentation is to do no segmentation at all. Global features, such as wavelets and histograms, have been very effective. When the image was recorded with a photographic purpose it is likely that the center of the image means something

different than the surrounding parts of the image, so using that division of the picture could be of help too. Using no segmentation at all is likely to run dry on semantics at the point where characteristics of the target are not specific enough in a large database to discriminate against the features of all other images.

In the third place, content-based retrieval has revitalized the interest in color image processing. This is due to the superior identification of tri-valued intensities in identifying an object, as well as due to the dominance of color in the perceptive aspect of images. And, as content-based is user-oriented, color cannot be left out. The purpose of most image color processing here is to reduce the influence of the accidental conditions of the scene and the sensing (i.e. the sensory gap) by computing sensing and scene invariant representations. Progress has been made in tailored color space representation for well-described classes of variant conditions. Also, the application of local geometrical descriptions derived from scale space theory will reveal viewpoint and scene independent salient point sets thus opening the way to similarity of images on a small number of most informative regions or points.

Finally, the attention for invariance has been revitalized as well with many new features and similarity measures. For content-based retrieval, invariance is just one side of the coin, where discriminatory power is the other. Little work has been reported so far to establish the remaining discriminatory power of properties. This is essential as the balance between stability against variations and retained discriminatory power determines the effectiveness of a property.

4.7.4 Similarity and learning

Similarity is an interpretation of the image based on the difference between two elements or groups of elements. For each of the feature types a different similarity measure is needed. For similarity between feature sets, special attention has gone to establishing similarity between histograms due to their computational efficiency and retrieval effectiveness. Where most attention has gone to color histograms, it is expected that histograms of local geometric properties and texture will follow. Being such a unique computational concept, the histogram is receiving considerable attention from the database community for upgrading the performance on very large data sets. This is advantageous in the applicability on applying retrieval on very

broad domains. To compensate for the complete loss of spatial information, recently new ways were explored as described above.

Similarity of hierarchically ordered descriptions deserves considerable attention, as it is one mechanism to circumvent the problems with segmentation while maintaining some of the semantically meaningful relationships in the image. Part of the difficulty here is to provide matching of partial disturbances in the hierarchical order and the influence of sensor-related variances in the description.

We make a pledge for the importance of human based similarity rather than general similarity. Also, the connection between image semantics, image data, and query context will have to be made clearer in the future. Similarity-induced semantics and the associated techniques for similarity adaptation (e.g. relevance feedback) are a first important step, but more sophisticated techniques, possibly drawing from machine learning, are necessary.

Learning is quickly gaining attention as a means to build explicit models for each semantic term. Learning is enabled by the sharp increase in data sets, and machine power to form categories from captions, from partially labeled sets, or even from unlabeled sets. Learning is likely to be successful for large, labeled data sets on narrow domains first, which may be relaxed to broader domains and less standardized conditions when data sets to learn from grow very big. Obviously, learning from labeled data sets is likely to be more successful than unsupervised learning first. New computational techniques, however, where only part of the data is labeled, or the data is labeled by a caption rather than categories open new possibilities. It is our view that, in order to bring semantics to the user, learning is inevitable.

4.7.5 Interaction

We consider the emphasis on interaction in image retrieval as one of the major changes with the computer vision tradition. Putting the user in control and visualization of the content has always been a leading principle in information retrieval research. It is expected that more and more techniques from traditional information retrieval will be employed or reinvented, in content-based image

retrieval. Text retrieval and image retrieval share the need for visualizing the information content in a meaningful way as well as the need to accept a semantic response of the user rather than just providing access to the raw data.

User interaction in image retrieval has, however, some different characteristics from text retrieval. There is no sensory gap and the semantic gap from keywords to full text in text retrieval is of a different nature. No translation is needed from keywords to pictorial elements. In addition to the standard query types, six essentially different image based types have been identified in this paper. Each requires their own user interface tools and interaction patterns. Due to the semantic gap, visualization of the query space in image retrieval is of great importance for the user to navigate the complex query space. While currently 2- or 3-dimensional display spaces are mostly employed in query by association, target search and category search are likely to follow. In all cases, an influx of computer graphics and virtual reality is foreseen in the near future.

As there is no interactivity if the response time is frequently over a second. The interacting user poses high demands on the computational support. Indexing a data set for interactive use is a major challenge as the system cannot anticipate completely on the user's actions. Still in the course of the interaction the whole query space i.e. the active image set, the features, the similarity, and the interpretations can all change dynamically.

4.7.6 The need for databases

When data sets grow in size, and when larger data sets define more interesting problems, both scientifically as well as for the public, the computational aspects can no longer be ignored.

The connection between content-based image retrieval and database research is likely to increase in the future. Already the most promising efforts are interdisciplinary but, so far, problems like the definition of suitable query languages, efficient search in high dimensional feature space, search in the presence of changing similarity measures, are largely unsolved.

It is regrettable that less work across the computer vision and database disciplines has been done yet, with a few notable exceptions. When truly large data sets come into view of hundreds of thousands of images, databases can no longer be ignored as an essential component of a content-based retrieval system. In addition when interactive performance is essential, storage and indexing must be organized in advance. Such large data sets will have an effect on the choice of features, as the expressive power, computational cost and hierarchical accessibility determine their effectiveness. For very large data sets a view on content integrated with computation and indexing cannot be ignored. When speaking about ``indexing" in the computer vision the emphasis is still on what to index whereas the emphasis from the database side is on how to index. The difference has become smaller recently, but we believe most work is still to be done. Furthermore, in dealing with large feature vector sizes, the expansion of query definitions and query expansions in a useful manner for a variety of user aims is still mostly unanswered.

For efficiency, more work on complete sets of feature calculations from compressed images is needed.

4.7.7 The problem of evaluation

It is clear that the evaluation of system performance is essential to sort out the good and the not-so-good methods. Up to this point in time a fair comparison of methods under similar circumstances has been virtually absent. This is due to the infancy of content-based retrieval but also to objective difficulties. Where interaction is a necessary component in most systems, it is difficult to separate out the influence of the data set in the performance. Also, it may be the case that some queries may match the expressive power of the system whereas others, similar at first glance, may be much harder. Searching for a sunset may boil down to searching for a large orange disc at about the center of the image. Searching for lamp, which may seem similar to the general audience, is a much harder problem as there are a variety of designs behind a lamp. The success of the system heavily depends on the toolset of the system relative to the query. In addition, it is logical that a large data set is composed of several smaller data sets to get a sufficiently big size. Then the difficulty is the internal coherence of the large data set with respect to the coherence of its constituents. When a data set is composed of smaller data sets holding interior

decorations, prairie landscapes, ships and pigeons, it is clear that the essential difficulty of retrieval is within each set rather than among them and the essential size of the data set is still one quarter. There is no easy answer here other than the composition of generally agreed upon data sets or the use of very, very large data sets. In all cases the vitality of the content-based approach calls for a significant growth of the attention to evaluation in the future.

A reference standard against which new algorithms could be evaluated has helped the field of text recognition enormously; see http://trec.nist.gov. A comprehensive and publicly available collections of images, sorted by class and retrieval purposes, together with a protocol to standardize experimental practices will be instrumental in the next phase of content-based retrieval. We hope that a program for such a repository will be initiated under the auspices of a funding agency.

At any rate, evaluation will likely play an increasingly significant role. Image databases, with their strong interactional component, present very different problems from the present ones which will require borrowing concepts from the psychological and social sciences.

4.7.8 The sensory gap and invariance

A major bottleneck is the difference between the physical scene and the observation of the scene. The observation is affected by accidental imaging circumstance, the viewpoint to the scene, the aspects of interaction between light and material, the limited resolution of the sensory system, and several other factors. This sensory gap between the object in the world and the information captured from a recording causes unwanted variance in the description of an object. To counteract the accidental aspects of the scene, a cognitive vision system will represent the image in many diverse invariant representations. We consider the transformation of sensory responses to invariants an important and inevitable information reduction stage in a general vision system. The resulting directly observable quantities are an essential part of the cognitive vision domain. As many physical parameters each may or may not affect the image formation process, a large variety of invariants can be deduced from visual data. Computer vision has partly solved the problem of invariant transformations.

4.7.9 The semantic gap and cognitive vision

A critical point in the advancement of content-based retrieval is the semantic gap, where the meaning of an image is rarely self-evident. Use of content-based retrieval for browsing will not be within the grasp of the general public as humans are accustomed to rely on the immediate semantic imprint the moment they see an image, and they expect a computer to do the same. The aim of content-based retrieval systems must be to provide maximum support in bridging the semantic gap between the simplicity of available visual features and the richness of the user semantics.

One way to resolve the semantic gap comes from sources outside the image by integrating other sources of knowledge in the query. Information about an image can come from a number of different sources: the image content, labels attached to the image, images embedded in a text, visual ontologies, and so on. We still have very primitive ways of integrating present knowledge in order to optimize access to images. Among these, the integration of natural language processing and computer vision deserves attention.

4.8 Film, TV and Entertainment

Content Analysis of Broadcast Video

The management of the huge amount of broadcast material currently being produced would benefit from automatic indexing and retrieval systems, able to analyse and identify the content of cinema film and TV. Typical applications needs include searching an interested scene in a news program and filter out the ‘uninteresting’ sequences of the studio images.

Existing systems are capable of detecting shots and extracting key frames using syntactic edge information and colour histograms, and limited face recognition. Players include Tecmath/Blue Order, Virage and a few research systems. They are useful in a limited range of situations in management and production.

Another important area of application is in sports event where a viewer may wish to obtain further information about the event - for example in a football match the

distance of a player to the goal or the precise position of the offside line. Current systems like VIZ-RT (D) and Epis FAST from Symah Vision (F) offer such functionality on a manual basis needing a manual calibration step without awareness of the different actors (ball, players, referees) and or using footage from fixed cameras.

Modification of Images

Often there is the need or wish not only to analyse and interpret an image but to modify it. Again the technology is driven by the advertising industry. To show the right adverts to the right group of people is essential in this business. Targeted advertisements can be inserted on-line into broadcast sports program and this is already available using systems such as Symah Vision and Viz-RT.

Well-established applications exist in the digital restoration of damaged film material. Image processing algorithms are used to remove scratches, flicker or damages caused by mould and dust. Most of these algorithms have to work motion-compensated to avoid introduction of digital artefacts. The quality is directly related to the quality of the estimated motion. Typical software based products are Diamant, Restore and MTI, hardware based products include Snell&Wilcox Archangel.

Content Production

There is increasing use of visual reconstruction in content production to remove unwanted elements from the real-scene such as wires and combine real imagery with computer generated characters. Camera tracking systems include products from RealViz and 2d3, which reconstruct camera motion from the scene using un-calibrated multiple view geometry. Reconstruction of camera movement allows synthetic elements to be introduced into the scene with the correct perspective to match real scene elements. Other applications include separation of foreground and background elements such as wires and props in post-production, products include MoKey from Imagineer Systems.

Human motion capture based on visual tracking and reconstruction of marker positions from multiple camera views is widely used in film and game production to reproduce realistic character movement. Current leading systems from Vicon,

MotionAnalysis perform on-line reconstruction of marker position and transformation into skeletal motion parameters at rates of over 100-200Hz. This technique is also widely used in conjunction with force plates for biometric analysis of human motion such as walking gait in clinical diagnosis.

Multiple camera studio systems using large numbers of cameras have been used to produce special effects shots such as freezing dynamic scenes such as Snell&Wilcox FlowMo. Motion prediction between frames is used to interpolate a smooth sequence of inbetween frames to synthesise a smooth virtual camera shot.

Current demand

Content Analysis

There is a demand for systems to manage the vast quantity of media archive footage and provide efficient retrieval. Systems are required to automatically annotate archive footage, combining cues from visual and audio content. Annotation is required to provide efficient indexing for retrieval of clips pertaining to specific high-level subjects such as a news event. Systems should be able to analyse scene dynamics to characterize the temporal as well as spatial behaviour and detect specific salient events. Techniques are required to convert spatio-temporal content into high-level semantic cues for annotation and retrieval.

To provide more complex statistics such as ball possession for a team or individual player, number of goals and fouls for fans and clubs, new approaches are necessary. Automatic detection and reconstruction of dynamic events is required. Another very important aspect, especially in sport programs is advertisement monitoring. The question how often and how long a brand is visible and at which size and quality is of major interest for all advertising companies. In the DETECT project, both automatic ball and player tracking and brand monitoring are being realised.

Other needs emerge with interactive television where, a viewer may click on a player or actor and be provided with the name and additional information such as web-pages, advertising and products, which requires person and object tracking and identification.

Modification of Images

The idea of digital cinema restoration can be further developed. Automatic film coloration and motion compensated conversion into different formats (e.g. from interlaced to progressive or HD to SD) are desirable. Both of this applications need high-quality object based motion information also clearly indication occluded regions. HDTV has been adopted

for digital production and distribution in film. This is emerging standard for broadcast requiring standardization techniques for HDTV coding and transmission.

Content Production

Digital actor modeling has been used in several successful film productions but is a highly skilled labour intensive process to create and animate realistic digital character models. People are the central element of most entertainment and communication content. Vision based capture is required to reproduce highly realistic digital actor models for broadcast and interactive games content production. Tools are required for marker-less reconstruction of actor shape, appearance and movement. Ongoing research led by the broadcast and animation industry on EU IST projects such as ORIGAMI and MELIES aims to produce prototype studio production tools for digital actors. Further research is required to facilitate production of video realistic sequences of digital actors which allow the producer to modify camera viewpoint, scene illumination and even actor movement in post-production. Higher level interpretation of human movement is required to achieve semantic understanding of human movement and produce natural human-computer interaction with digital characters for interactive entertainment and personalized characters for on-line help services. Other potential applications of automatic scene understanding for content production include on-line annotation and synthesis of novel views for sports events.

What advances in Cognitive Vision are needed

Central to the demands of applications outlined above are a number of fundamental advances in cognitive vision:

- Automatic high-level annotation of dynamic scenes to identify scene subject and context.
- Event detection to identify semantically important temporal structure.
- Action understanding to interpret the causality of events to provide structured descriptions and interpret sequences of events.
- Fusion of information available from audio and visual sequences for annotation
- Fusion of multiple visual cues to recognise and interpret image structure
- Indexing of video archives based on high-level cues.
- Illumination invariant analysis of natural scenes
- Specific tools for modelling and interpretation of the spatio-temporal behaviour of people for scene understanding, content production and interaction.

These advances are required to enable high-level interpretation of sequences of images. Automatic interpretation of dynamic scenes is a generic requirement for many potential applications of computer vision in entertainment and communication.

4.9 Aerial and Satellite Image Analysis

(James Ferryman).

4.10 Aerospace

One of the main goals in the area of aerospace is safer flying systems. Whilst the technical feasibility of vision guided control has been demonstrated, there are no operating systems in civil aviation. The reliability requirements of such systems are extremely high. The SLATS project (**Safer Landing and Taxiing**) underway with support from the DTI, in collaboration with the Civil Aviation Authority, aims to build the safety case for the use of a dual redundant (vision and GPS) system by providing better aircraft localisation, and so to improve civil aircraft approach, to decrease runway occupancies and to reduce incidents that damage landing gear.

Emerging applications

In the military, long-term (10 years) development programmes exist to demonstrate these capabilities to meet the new defence needs of the 21st century. The direction of development is towards greater autonomy, decentralised intelligent systems and decentralised decision-making with sensing in an unknown environment.

The US Department of Defence has developed a four element reference architecture for data fusion which includes picture compilation (**what can I see?**), situation assessment (**where am I?**), threat classification (**what does it mean?**) and resource allocation (**what should I do about it?**). This may fulfil some of the requirement for cognitive vision. However the system-building aspects and the lack of well-characterised modules remain challenges at present. In particular the ontologies underlying module interaction and the subsequent negotiation have yet to be fully explored.

Research systems are in development to carry out constrained tasks with some civil applicability – for example exploring a single story building and building a map – for disaster recovery.

4.11 Medical imaging and life sciences

Life sciences

The discovery and development of new drugs is a long and complicated process with various stages of tests. The pharmaceutical and biotechnology industries already make extensive use of images in the latter stages of clinical trials (MRI, X-ray/CT) that are examined by medically trained staff. Increasingly images are being analysed in the earlier stages of the process. Drug development frequently makes use of animal models of disease – for example in a study to examine the effects of an anti-anxiety drug, the **behaviour of rat** may be observed. Existing systems, such as those from Noldus, are able to track the animal in an image sequence and record its positions, even if it's shape varies considerably during the movement. Manual annotation is required to link the movement to specific behaviours. Another application area is in the **study of living cells**. Cells change their appearance over time and after being stimulated or exposed to substances. Workstations supporting specific application algorithms are now appearing eg Arrayscan and Kineticscan from Cellomics.

Another of the challenges of the task is to be able to link changes in cells and animal behaviour to information in the expanding **genetic and other databases** and thus obtain new insights.

Emerging applications

Although image processing is used to extract data from a scene, the interpretation of the images is mostly still in the hands of a skilled human observer with clear limitations in terms of repeatability, experiment duration and cost where **skilled staff**

are limited. To automate this time consuming task an understanding of the scene is needed. The movement of the animal, a rat for example, has to be **identified as posture** such rat grooming, as well as **novel postures** that might otherwise be missed. **Social interactions** are also difficult using simple tracking systems.

A final problem is the developed of **user interfaces** that allow interaction with skilled staff and to present data in a form which is understandable in biological terms.

5. Fundamental Research Problems

The fundamental research problems should be drawn from the applications described in the previous section. However, the Cognitive Vision ontology group has done an excellent first cut at defining these issues. We take their description of Cognitive Vision Research Problems as a starting point for this section. To the extent possible, the research issues proposed in the previous section will be clustered around these problems. If necessary, we will define new categories of research problems.

5.1 Model Learning

cognitive vision function	required capability	application sector
learning	· Generalising from a small set (good example/no bad examples) · Linking low level features to higher level rules · Learning 'faces' or similar object	**Surveillance** **Industrial inspection** Film/TV/entertainment **Life Science**

Todays systems require the system designer to specify the model of reality that the system will use. This si a fundamental limitation (closed universe). To operate inan open universe, a system must be able to automatically acquire knowledge of 1 to 7 below. (Hillary or Gerhard)

1.Activity/Behaviours/Processes/Dynamics
2.Classification/Category (Approaches,Applications,General)
3.Context/Scenes/Situations
4.Function
5.Objects/Parts
6.Parameters(Approaches,Applications,General)
7.Task Control

Issues

1.Learning Control
2.Validation

3.Types of Learning

1.Case-based

2.Reinforcement

3.Supervised

4.Unsupervised

5.2 Knowledge Representation

5.3 Recognition, Categorization and Estimation

5.4 Reasoning about Structures and Events

cognitive vision function	required capability	application sector
reasoning about events and structures	Complex non-linear closed loop control Understand high level 'rules of the game'	Industrial inspection Stock photo databases Film/TV/ entertainment Industrial robotics Aerospace

5.5 Architecture and Visual Process Control

5.6 Interaction with the environment

5.7 Performance Evaluation

5.7.1 Evaluation of Methods and Algorithms

(James, Patrick Courtney, Adrian Clark...

Benchmarks

Performance Metrics

5.7.2 Self-monitoring, auto-description, auto-criticism, self diagnosis,

(Wolfgang Foerstner, Jim)

used for systems control

5.8 Self Diagnosis

Vision systems are nearly always embedded systems and are often the bottle neck of an automation chain. Their performance therefore is crucial for their usefulness within the embedding system. Moreover, vision systems mostly are federated systems, consisting of a set of basic vision modules which are performed in a chain or parallel.

Using the above mentioned ontology, we may interprete vision modules as observers or observation modules

The situation is sketched in fig.~\ref{fig:system-perf}. An observation module in a first instance uses observations, possibly including measurements, preferably with measures on their quality and produces observations, possibly measurements, again preferably with measures on their quality, which allows chaining based on the exchange of evaluated observations (cf. fig. \ref{fig:federated-module}).

Independent of the output type of an observation module, e. g. reading bar codes, detecting passengers, tracking objects, or determining ego-motion, the observation module may be successful and produce one of these outputs, or it may fail and not deliver its output. Acceptable failure rates modules may vary between 25 \% e.~g. in automatic image retrieval, and 0.0001 \% in visual control of production processes.

Many systems allow an efficient reaction to {\em self-diagnosis} of the observation modules. Self-diagnosis is the {\em internal evaluation of the moduel} based on redundant information within module. Often such redundancy is intentionally introduced into the design of the module only to enable self-diagnosis.

The evaluation itself {\em may include expected causes} for the performance and {\em recommendations of remedies} in case of a bad evaluation. Both, causes and recommendations may be encoded in or be learned by the system. In favourable situations fuailures may be {\em classified and corrected} for.

An increase in efficiency of an observation module with self-diagnosis could be achieved by eliminating failures in case the system actually predicts a failure based on internal self-diagnosis. A module with self-diagnosis therefore can be interpreted as a classifier with respect to its ability to predict success or failure, independent of the type of its output within the embedding module.

Decisions to use certain modules within a federated observation system need to be made subject to the success of the modules. This could require active checking. However, using an {\em algorithm for self-diagnosis} could produce {\em values characterizing the internal success or failure}, again preferably {\em with quality measures}. Together with the quality of the output observations, the self-diagnosis can be used by the control module.

The distinction between the quality of the output observations and the result of the self-diagnosis algorithm is motivated by the inability or often reduced ability of the control module to interpret the details of the output characteristics of the observation module.

Characterizing the performance of a system therefore obviously may also refer to characterizing the performance of its self-diagnostic abilities, on top of the performance of the output of the vision system as such.

```
\begin{figure}[htbp]
  \centering
  \includegraphics[width=1.0\textwidth]{system-perf}
  \caption{\em Components of an observation module with
    self-diagnostic capabilities. Input and output data may be
    measurements or observations.  Observation modules may be
    chained to yield federated observation modules including an
    internal control on their use, cf. fig. \ref{fig:federated-module}.}
  \label{fig:system-perf}
\end{figure}
```

```
\begin{figure}[htbp]
   \centering
   \includegraphics[width=1.0\textwidth]{federated-module}
   \caption{\em Example for federated module: Input data for federated
     module and outcome of individual modules are used to trigger
     control module, which itself is calling individual modules. The
     selfdiagnosis refers to the total outcome of the federeated
     module, possibly using the self-diagnosis capabilities of the
     individual modules into account. }
   \label{fig:federated-module}
```

6. Recommendations

What are the enabling technologies that may stimulate progress.

In summary, the following recommendations appeared as a result of the discussions:

- ⇨ Cognitive vision techniques will bring benefits to broad range of application areas and in particular, learning and reasoning.
- ⇨ At the current state of development suitable end user interface must be considered
- ⇨ A generic person detection module would be of benefit it many areas
- ⇨ A baseline system architecture would be of benefit
- ⇨ System design methodologies need to be developed and the system life cycle considered

Many of the currently discussed cognitive vision applications focus on the consumer as the end user. It is worth pointing out that there are many business and professional applications that would benefit from the technology. They are likely to be less cost sensitive and more rational in their requirements compared to consumer needs. However they have significantly different ease of use and reliability requirements, and the organisational business processes remain as barriers to be understood.

Annexes

Annex 1. A Glossary of terms for cognitive computer vision.

2.3 The ECVision Cognitive Vision Ontology

The contents of this section move to fundamental research issues.

(People directly involved in its development are: Bob Fisher, Wolfgang Förstner, Annett Faber and Hanns-Florian Schuster.)

Top level - Corresponds to different theory sets.

1. Model Learning (Survey Result)
2. Knowledge Representation (Survey Result)
3. Recognition, Categorization and Estimation (Survey Result)
4. Reasoning about Structures and Events (Survey Result)
5. Visual Process Control (Survey Result) & interaction with the environment
6. Emerging Topics'
7. Case Studies

The Hierarchy. (Need to find a more synthetic way to present this).

1.Model Learning (Survey Result)

1.Specific approaches to learning these different types of content (See also Knowledge Representation->Content for "what" things that are learned and Recognition, Categorization and Estimation->Specific Approaches for "how" things might be recognized.)

1.Activity/Behaviours/Processes/Dynamics

2.Classification/Category (Approaches,Applications,General)

3.Context/Scenes/Situations

4.Function

5.Objects/Parts

6.Parameters(Approaches,Applications,General)

7.Task Control

2.Issues

1.Learning Control

2.Validation

3.Types of Learning

1.Case-based

2.Reinforcement

3.Supervised

4.Unsupervised

2.Knowledge Representation (Survey Result)

1.Content (See also Model Learning->Specific Approaches for learning different types of content and Recognition, Categorization and Estimation->Specific Approaches for "how" things might be recognized.)

1.Activity/Behaviour/Processes/Dynamics

2.Classification/Category (Approaches,Applications,General)

3.Context/Scene/Situations

4.Function

5.Objects/Parts

6.Ontologies

7.Parameters

8.Task Control

2.Issues

1.Indexing

2.Storage

3.Style

1.Appearance-based

2.Embodied

3.Generative

4.Geometric

5.Logical

6.Ontological

7.Probabilistic

8.Procedural

9.Relational/Graphical

3.Recognition, Categorization and Estimation (Survey Result)

1.General Approaches

1. Appearance

2. Feature Sampling

3. Geometric/Structural

4. Physical Models

5. Property

6. Temporal (discrete or continuous)

2.Issues

1.Accuracy

2.Generic Classes

3.Labeling/Localization

3.General Techniques

1.Alignment

2.Attention

3.Search(Approaches,Applications,General)

4.Figure/Ground

5.Grouping/Perceptual Organization(Approaches,Applications,General)

6.Labeling(Approaches,Applications,General)

7.Parameter Estimation and Optimization

4.Specific Approaches to recognizing things (See also Knowledge Representation->Content for "things" that are learned and Model Learning->Specific Approaches for "learning" different types of content).

1.Activity/Behaviours/Processes/Dynamics (Approaches,Applications,General)

2.Class/Category(Approaches,Applications,General)

3.Context/Scenes/Situations

4.Functions

5.Objects/Parts(Approaches,Applications,General)

6.Parameters

4.Reasoning about Structures and Events (Survey Result)

1.Content
    1.Appearance/Visibility
    2.Objects & Spatial structures and their organistion
    3.Tasks/Goals
    4.Events & temporal structures and their organisation
2.Issues
    1.Performance
    2.Prediction
    3.Self-analysis
    4.Uncertainty
3.Methods
    1.Constraint Satisfaction
    2.Hypothesize and Verify
    3.Logical
    4.Model-based
    5.Rule-Based
    6.Statistical

5.Visual Process Control (Survey Result)
1.Decision Making
    1.Probabilistic
    2.Rule Based
    3.Soft Control
2.Issues
    1.Active Sensing
    2.Goal Specification
    3.Planning
    4.Process Control & Monitoring
    5.Speed of Response
3.Paradigms
    1.Central/Distributed
    2.Covert Control
    3.Reactive

6.Emerging Topics (Survey Result)

1.Vision & Language Fusion

7.Case Studies

A.2. Principal Research Groups in Cognitive Vision

A2.1 European Union member states

A2.2 European Union associated states

A2.3 European Union associated states

References

References

[Simon 87] H. A. Simon, "Sciences of the Artificial", ….

Allen 83

J.F. Allen: Maintaining Knowledge About Temporal Intervals. In: Communications of the ACM 26 (11), 832-843, 1983

Allen et al. 90

J. Allen, J. Hendler, A. Tate (Eds.): Readings in Planning. Morgan Kaufmann, 1990

Badler 75

N. I. Badler: Temporal Scene Analysis: Conceptual Descriptions of Object Movements, TR 80, Department of Computer Science, University of Toronto, 1975

Buxton 03

H. Buxton: Learning and Understanding Dynamic Scene Activity: A Review. Image and Vision Computing 21 (2003), 125-136

Charniak and Goldman 91

E. Charniak, R. Goldman: A Probabilistic Model of Plan Recognition. AAAI, 1:160-165, 1991

Escrig 98

M.T. Escrig, F. Toledo: Qualitative Spatial Reasoning: Theory and Practice, IOS, 1998

Fernyhough et al. 98

J. Fernyhough, A.G. Cohn, D. Hogg: Building Qualitative Event Models Automatically from Visual Input. Proc. ICCV-98, IEEE Computer Society, 1998, 350-355

Getoor et al. 01

L. Getoor, N. Friedman, D. Koller, B. Taskar: Learning Probabilistic Models of Relational Structure. Eighteenth International Conference on Machine Learning (ICML), Williams College, June 2001.

Hayes 79

P. Hayes: The Naive Physics Manifesto. In D. Michie (Ed.): Expert Systems in the Microelectronic Age, Edinburg University Press, 1979

Huber and Durfee 95

M.J. Huber and E.H. Durfee: Deciding When to Commit to Action During Observation-based Coordination. Proceedings First International Conference on Multi-Agent Systems, 1995, 163 - 170

Lutz 02

C. Lutz: Description Logics with Concrete Domains - A Survey. In Advances in Modal Logics Volume 4. World Scientific Publishing Co. Pte. Ltd., 2002

Mitchell 97

T. Mitchell: Machine Learning. McGraw Hill 1997

Neumann 89

B. Neumann: Natural Language Description of Time-Varying Scenes in: Semantic Structures, D. Waltz (Hrsg.), Lawrence Erlbaum, 167206, 1989

Pearl 00

J. Pearl: Causality. Cambridge University Press, 2000

Reiter 96

R. Reiter: Natural actions, concurrency and continuous time in the situation calculus. In Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth

International Conference (KR'96) , Cambridge, Massachusetts, U.S.A. November 5-8, 1996

Reiter and Mackworth 87
R. Reiter, A. Mackworth: The Logic of Depiction, TR 87-23, Dept. Computer Science, Univ. of British Columbia, Vancouver, Canada, 1987

Selfridge 55
O.G. Selfridge: Pattern Recvognition and Modern Computers. Western Joint Computer Conference, 1955, 91-93

Vila 94
L. Vila: A Survey on Temporal Reasoning in Artificial Intelligence. AI Communications 7 (1), 4-28, 1994

Weld and de Kleer 90
D.S. Weld, J. de Kleer (Eds.): Readings in Qualitative Reasoning about Physical Systems. Morgan Kaufmann, 1990

[Schumpeter 42] J. A. Schumpeter, Capitalism, Socialism and Democracy. New York: Harper & Brothers, 1942. Revised 2nd Edition, 1947. Enlarged 3rd edition, 1950.

[Utterback 94] J. M. Utterback , Mastering the Dynamics of Innovation, Harvard Business School Press, 1994.

[Mowery 98] D. C. Mowery and N. Rosenberg, "Paths of Innovation", Cambridge University Press, 1998.

Gladstone 00] M. Gladstone, "The tipping point: Or how little things can make a big difference.", New York: Little Brown and Company, 2000.

[Rogers 95] Rogers, E. M. Diffusion of innovations (4 th Edition). New York: The Free Press, 1995.

[Dent 93] Dent, H.. The Great Boom Ahead. Hyperion Press, Nov. 1993.

[Morris 69] American Heritage Dictionary, W Morris, Ed., Houghton Mifflin Co., 1969

[Cao 97] Cao, T. Y., Conceptual Developments in 20th centure field theories, Cambridge Univ.Press, 1997

D. C. Van Essen, C. H. Anderson, and D. J. Felleman, Information processing in the primate visual system: An integrated systems perspective, Science 255, 419-423, 1992.

L. J. M. Florack, Image Structure, PhD Thesis Utrecht University, 1991.

D. H. Foster and S. M. C. Nascimento. Relational colour constancy from invariant cone-excitation ratios.
Proc. R. Soc. London B 257, 115-121, 1994.

J. M. Geusebroek et al., Color invariance, IEEE Trans. Pattern Anal. Machine Intell. 23, 1338-1350, 2001.

L. Van Gool et al., Vision and Lie's approach to invariance, Image Vision Comput.. 13, 259-277, 1995.

D. Knill, W. T. Freeman, and W. S. Geisler (editors), Special issue JOSA-A on Bayesian and Statistical Approaches to Vision, American Optical Society, to appear, 2003.

J. J. Koenderink, The structure of images, Biol. Cybern. 50, 363-370, 1984.

D. Marr, Vision, Freeman and Co., 1983.

J. Mundy and A. Zisserman (editors), Geometric Invariance in Computer Vision, Springer-Verlag, 1992.

R. Nozick, Invariances: The Structure of the Objective World, Harvard University Press, 2002.

B. O'Shaughnessy, Consiousness and the World, Oxford University Press, 2002.

C. Schmid et al., Evaluation of interest point detectors, Int. J. Comput. Vision.. 37, 151-172, 2000.

A. W. M. Smeulders et al., Content based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Machine Intell. 22, 1349-1379, 2000.

A. W. M. Smeulders et al., Content based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Machine Intell. 22, 1349-1379, 2000.