Available online at www.sciencedirect.com
SCIENCE DIRECT®
Image and Vision Computing xx (2006) 1–14
image AND vision COMPUTING
www.elsevier.com/locate/imavis

Cognitive vision: The case for embodied perception

David Vernon *

LIRA-Lab, University of Genova, Viale Causa 13, 16145, Genova, Italy

Received 17 May 2004; received in revised form 29 June 2005; accepted 5 August 2005

Abstract

This paper considers arguments for the necessity of embodiment in cognitive vision systems. We begin by delineating the scope of cognitive vision, and follow this by a survey of the various approaches that can be taken to the realization of artificial cognitive vision systems, focussing on cognitive aspects. These range from the cognitivist symbolic representational paradigm, through connectionist systems and self-organizing dynamical systems, to the enactive cognition paradigm. We then consider various arguments for embodiment, beginning with paradigm-specific cases, and concluding with a paradigm-independent argument for embodied perception and cognition. We explore briefly different forms of embodiment and their relevance to the foregoing viewpoints. We highlight some of the key problems associated with embodied cognitive vision, including the phylogeny/ontogeny trade-off in artificial systems and the developmental limitations imposed by real-time environmental coupling. Finally, we conclude by considering some aspects of natural cognitive systems to see how they can provide insights to help in addressing these problems.

Keywords: Cognitive vision; Embodiment; Cognitivist systems; Dynamical systems; Enactive systems

1. The scope of cognitive vision

The term *cognitive vision* has recently been introduced to encapsulate an attempt to achieve more robust, resilient, and adaptable computer vision systems by endowing them with cognitive capabilities. A cognitive vision system should be able to engage in purposive goal-directed behaviour, it should be able to adapt robustly to unforeseen changes of the visual environment, and it should be able to anticipate the occurrence of objects or events [1–3]. The characteristic of anticipation and prospective behaviour in a cognitive vision system is very important as it requires the system to operate across a variety of time-scales, extending into the future, so that it is capable of more than reactive behaviour (which can be quite complex in its own right).

Some authors in discussing cognitive aspects of systems go even further than this. For example, it has been suggested that a cognitive system should be able to view a problem in more than one way and to use knowledge about itself and the environment so that it is able to plan and modify its actions on the basis of that knowledge [4]. Others suggest that a cognitive computer system should—as well as being able to reason, to learn from experience, to improve its performance with time, and to respond intelligently to things it's never encountered before—be able to explain what it was doing and why it was doing it [5]. This would enable it to identify potential problems in following a current approach to carrying out a task or to know when it needed new information in order to complete it.

Cognitive vision is in essence a combination of computer vision and cognition.[1] Consequently, to make sense of cognitive vision, we must first address the issue of cognition. This is where the trouble begins. Unfortunately, the term cognition has several interpretations, each of which is dependent on the very disparate underlying models, and the discipline of cognitive science is itself going through something of a metamorphosis [7]. The following attempts a very brief overview of the broad range of approaches that can be brought to bear on cognition; for a more extensive treatment, see [1].

* Tel.: +33 4 949 929 06.
E-mail address: vernon@ieee.org

0262-8856/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.imavis.2005.08.009

[1] See [6] for an alternative viewpoint on the possible nature of cognitive vision.

2. A survey of cognition paradigms

There are several quite distinct approaches to understanding and synthesis of cognitive systems, including physical symbol systems, connectionism, artificial life, dynamical systems, and enactive systems [7,8]. Each of these makes significantly different assumptions about the nature of cognition, its purpose, and the manner in which cognition is achieved. Among these, however, we can discern two broad classes: the *cognitivist* approach based on symbolic information processing representational systems; and the *emergent systems* approach, embracing connectionist systems, dynamical systems, and enactive systems, and based to a lesser or greater extent on principles of self-organization.

2.1. Symbolic information processing representational cognitivist models

Cognitivism asserts that cognition involves computations defined over symbolic representations, in a process whereby information about the world is abstracted by perception, represented using some appropriate symbol set, reasoned about, and then used to plan and act in the world. This approach has also been labelled by many as the *information processing* approach to cognition [9–14]. The discipline of cognitive science is often (erroneously) identified exclusively with this particular approach [14]. It is, however, by no means the only paradigm in cognitive science and there are indications that the discipline is migrating away from its stronger interpretations [7].

For cognitivist systems, cognition is representational in a strong and particular sense: it entails the manipulation of explicit symbolic representations of the state and behaviour of an objective external world [15] to facilitate appropriate, adaptive, anticipatory, and effective interaction, and the storage of the knowledge gained from this experience to reason even more effectively in the future. Vision in particular and perception in general are concerned with the abstraction of faithful spatio-temporal representations of the external world from sensory data. Reasoning itself is symbolic: a procedural process whereby explicit representations of an external world are manipulated and possibly translated into language.

In most cognitivist approaches concerned with the creation of artificial cognitive systems, the symbolic representations are the product of a human designer. This is significant because it means that they can be directly accessed and understood or interpreted by humans and that semantic knowledge can be embedded directly into and extracted directly from the system. However, it has been argued that this is also the key limiting factor of cognitivist vision systems: these designer-dependent representations are the idealized descriptions of a human cognitive entity and, as such, they effectively bias the system (or 'blind' it [15]) and constrain it to an domain of discourse that is dependent on and, a consequence of, the cognitive effects of human activity. This approach works well as long as the system does not have to stray too far from the conditions under which these descriptions were formulated. The further one does stray, the larger the 'semantic gap' [16] between perception and possible interpretation, a gap that is normally plugged by embedding programmer knowledge or enforcing expectation-driven constraints [17] to render a system practicable in a given space of problems.

One can see how this approach usually goes hand-in-hand with the fundamental assumption that 'the world we perceive is isomorphic with our perceptions of it as a geometric environment' [18]. The goal of cognition, for a cognitivist, is to reason symbolically about these representations in order to effect intelligent, adaptive, anticipatory, goal-directed, behaviour, and the goal of cognitive vision is to provide these symbolic representations in the first place.

The vast majority of computer vision systems, both cognitive and classical, adopt an essentially cognitivist position, especially with regard to their approach to representational issues. Since real perceptual systems work with inherently uncertain, time-varying, and incomplete information, computer vision systems are increasingly turning to the use of machine learning to improve the resilience of these systems (e.g. [19]). However, this does not alter the fact that the representational structure is still predicated on the descriptions of the designers. The significance of this will become apparent in later sections.

An example of the use of explicit symbolic (or conceptual) knowledge in cognitivist systems can be found in [20]. This is a model-based cognitive vision system, developed for the interpretation of video sequences of traffic behaviour and the generation of a natural language description of the observed environment. It proceeds from signal represententations to symbolic representations through several layers of processing, including gradient-based optical flow, edge detection, 3D model fitting, Kalman filter based computation of object vehicle trajectories. These trajectories are categorized as elementary vehicle movements or manoeuvers. Finally, vehicle behaviour is represented by situation graph trees (SGT) based on these manoeuvers. Automatic interpretation of this representation of behaviour is effected by translating the SGT into a logic program (based on fuzzy metric temporal Horn logic). Note that the flow of control between the sub-symbolic and symbolic levels is bi-directional so that, for example, the behaviour representational level can provide input to improve the performance of the Kalman filter tracking during periods of occlusion (Kalman filters are normally driven only by sensory data). See also [21–25] for related work.

The cognitivist assumptions are also reflected well in the model-based approach described in [26,27] which uses Description Logics, based on First Order Predicate Logic, to represent and reason about high-level concepts such as spatio-temporal object configurations and events.

Probabilistic frameworks have been proposed as an alternative (or sometimes an adjunct [26]) to these types of deterministic reasoning systems. For example, a cognitive vision system for interpreting the activities of expert human operators is described in [28–30]. It exploits dynamic decision networks (DDN)—an extension of Bayesian belief networks to incorporate dynamic dependencies and utility theory [31]—for

recognizing and reasoning about activities, and both time delay radial basis function networks (TDRBFN) and hidden markov models (HMM) for recognition of gestures. Although this system does incorporate learning to create the gesture models, the overall symbolic reasoning process, albeit a probabilistic one, still requires the system designer to identify the contextual constraints and their causal dependencies (for the present at least: on-going research is directed at removing this restriction) [28-30].[2] Recent progress in autonomously constructing and using symbolic models of behaviour from sensory input using inductive logic programming is reported in [32].

The dependence of cognitivist approaches on designer-oriented representations is also well exemplified by knowledge-based systems such as those based on ontologies. For example, see [33] which describes a framework for an ontology-based cognitive vision system that focusses on mapping between domain knowledge and image processing knowledge using a visual concept ontology incorporating spatio-temporal, textural, and colour concepts.

An adaptable system architecture for observation and interpretation of human activity that dynamically configures its processing to deal with the context in which it is operating is described in [34] while a cognitive vision system for autonomous control of cars is described in [35].

A cognitive framework that combines low-level processing with high-level processing using a language-based ontology and adaptive Bayesian networks is described in [36]. The system is self-referential in the sense that it maintains an internal representation of its goals and current hypotheses. Visual inference can then be performed by processing sentence structures in this ontological language. It adopts a quintessentially cognitivist symbolic representationalist approach, albeit that it uses probabilistic models, since it requires that a designer identify the 'right structural assumptions' and prior probability distributions. The authors say the model represents an approach to solving the symbol grounding problem and the frame problem (see also Section 6.4).

2.2. Emergent systems

Emergent systems, embracing connectionist, dynamical, and enactive systems, take a very different view of cognition. Here, cognition is a process of self-organization whereby the system is continually re-constituting itself in real-time to maintain its operational identity through moderation of mutual system–environment interactions and co-determination [37]. Co-determination implies that the cognitive agent is specified by its environment and at the same time that the cognitive process determines what is real or meaningful for the agent. In a sense, co-determination means that the agent constructs its reality (its world) as a result of its operation in that world. This has significant implications for the nature of perception and cognitive vision. 'Perceiving is not strictly speaking in the animal or an achievement of the animal's nervous system, but rather is a process in an animal–environment system' [14]. Co-determination is one of the key differences between the emergent paradigm and the cognitivist paradigm, wherein an objective reality common to all cognitive agents is assumed. For emergent systems, vision provides appropriate sensory data to enable effective action [37] but it does so as a consequence of the system's actions. In the emergent paradigm, cognitive vision is functionally dependent on the richness of the action interface [38].

2.2.1. Connectionist models

One of the original motivations for work on emergent systems was disaffection with the sequential, atemporal, and localized character of symbol-manipulation based cognitivism [8]. Emergent systems, on the other hand, depend on parallel, real-time, and distributed architectures. One of the key features of emergent systems, in general, and connectionism, in particular, is that 'the system's connectivity becomes inseparable from its history of transformations, and related to the kind of task defined for the system' [8]. Whereas in the cognitivist approach the symbols are distinct from what they stand for, in the connectionist approach, 'meaning relates to the global state of the system' [8]. Indeed, the meaning is something attributed by an external third-party observer to the correspondence of a system state with that of the world in which the emergent system is embedded.

Connectionist approaches are for the most part associative learning systems in which the learning phase is either unsupervised (self-organizing) or supervised (trained). For example, hand-eye coordination can be learned by a Kohonen neural network from the association of proprioceptive and exteroceptive stimuli [39,40]. As well as attempting to model cognitive behaviour, connectionist systems can self-organize to produce feature-analyzing capabilities similar to those of the first few processing stages of the mammalian visual system (e.g. centre-surround cells and orientation-selective cells) [41]. An example of a connectionist system, which exploits the co-dependency of perception and action in a developmental setting can be found in [42]. This is a biologically motivated connectionist system that learns goal-directed reaching using colour-segmented images derived from a retina-like log-polar sensor camera. The system adopts a developmental approach: beginning with innate inbuilt primitive reflexes, it learns sensori–motor coordination. Radial basis function networks have also been used in cognitive vision systems, for example, to accomplish face detection [29].

2.2.2. Dynamical models

Dynamical systems theory is very general and can be deployed to model many different types of systems in such diverse areas as biology, astronomy, ecology, economonics, physics, and many more. It has been used to complement classical approaches in artificial intelligence [43] and it has also been deployed to model natural and artificial cognitive systems [13,14,44]. Advocates of the dynamical systems approach to cognition argue that motoric and perceptual

[2] See [31] for a survey of probabilistic generative models for learning and understanding activities in dynamic scenes.

systems are both dynamical systems, each of which self-organizes into meta-stable patterns of behaviour. Perception-action coordination can also be characterized as a dynamical system.

A dynamical system defines a particular pattern of behaviour. The system is characterized by a state vector $\mathbf{q}$ and its time derivative $\dot{\mathbf{q}}$ is a function of the state vector, control parameters $\mathbf{p}$ and noise n. It is a self-organizing system because the system dynamics are defined by and only by the system state $\dot{\mathbf{q}} = \mathbf{N}(\mathbf{q}, \mathbf{p}, n)$.

In general, a dynamical system is an open dissipative non-linear non-equilibrium system: a system in the sense of a large number of interacting components with large number of degrees of freedom, dissipative in the sense that it diffuses energy (its phase space decreases in volume with time implying preferential sub-spaces), non-equilibrium in the sense that it is unable to maintain structure or function without external sources of energy, material, information (and, hence, open). The non-linearity is crucial: as well as providing for complex behaviour, it means that the dissipation is not uniform and that only a small number of the system's degrees of freedom contribute to its behaviour. These are termed *order parameters* (or *collective variables*). Each order parameter defines the evolution of the system, leading to meta-stable states in a multi-stable state space (or phase space). It is this ability to characterize the behaviour of a high-dimensional system with a low-dimensional model that is one of the features that distinguishes dynamical systems from connectionist systems [14].

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

Clark [7] has pointed out that the antipathy which proponents of dynamical systems approaches display toward cognitivist approaches rests on rather weak ground insofar as the scenarios they use to support their own case are not ones that require higher level reasoning: they are not 'representation hungry' and, therefore, are not well suited to be used in a general anti-representationalist (or anti-cognitivist) argument. At the same time, Clark also notes that this antipathy is actually less focussed on representations per se (dynamical systems readily admit internal states that can be construed as representations) but more on objectivist representations, which form an isomorphic symbolic surrogate of an absolute external reality.

It has been argued that dynamical systems allow for the development of higher order cognitive functions, such as intentionality and learning, in a straight-forward manner, at least in principle. For example, intentionality—purposive or goal-directed behaviour—is achieved by the superposition of an intentional potential function on the intrinsic potential function [14]. Similarly, learning is viewed as the modification of already-existing behavioural patterns that take place in a historical context whereby the entire attractor layout (the phase-space configuration) of the dynamical system is modified. Thus, learning changes the whole system as a new attractor is developed.

Although dynamical models can account for several non-trivial behaviours that require the integration of visual stimuli and motoric control, including the perception of affordances, perception of time to contact, and figure-ground bi-stability [14,45–48], the principled feasibility of higher-order cognitive faculties has yet to be validated. Rectifying this situation is one of the most important research issues in dynamical systems models of cognition and cognitive vision.

Dynamical approaches differ from connectionist systems in a number of ways [14,13,44]. Suffice it here to note that the connectionist system is often defined by a general differential equation, which is actually a schema that defines the operation of many (neural) units. That is, the differential equation applies to each unit and each unit is just a replication of a common type. This also means that there will be many independent state variables, one for each unit. Dynamical systems, on the other hand, are not made up of individual units all having the same defining equation and cannot typically be so decomposed. Typically, there will be a small number of state variables that describe the behaviour of the system as a whole.

2.2.3. Enactive systems models

Cognitivism, by definition, involves a view of cognition that requires the representation of a given objective pre-determined world [8,44]. Enaction [8,15,37,49–52] adopts a fundamentally different stance: cognition is a process whereby the issues that are important for the continued existence of the cognitive entity are brought out or enacted: co-determined by the entity as it interacts with the environment in which it is embedded. Thus, nothing is 'pre-given', and hence there is no need for representations. Instead, there is an enactive interpretation: a context-based choosing of relevance. In this sense, the philosophical ground of enaction is Husserlian phenomenology, in contradistinction to the objectivist realism of the cognitivist approach. Whilst this might sound vaguely out of place in a vision paper, and indeed irrelevant for those interested in engineering cognitive vision systems, it has very practical implications. It comes down to a simple choice of axioms upon which to build a cognitive vision system. Is the role of cognition to abstract objective structure and meaning through perception and reasoning? Or, is it to uncover unspecified regularity and order that can then be construed as meaningful because they facilitate the continuing operation and evolution of the cognitive system?

Enaction adopts the second stance, one that is actually more neutral, assuming only that there is the basis of order in the environment in which the cognitive system is embedded. From this point of view, cognition is exactly the process by which that order or some aspect of it is uncovered (or constructed) by the system. This allows that there are different forms of reality (or relevance) that are dependent directly on the nature of the dynamics making up the cognitive system and its space of

interaction with the environment. The advantage for cognitive vision systems is that the enactive approach focusses on the dynamics by which robust interpretation and adaptability arise.

The enactive systems research agenda stretches back to the early 1970 in the work of computational biologists Maturana and Varela and has been taken up by others, including some in the main-stream of classical AI [8,15,37,49–52].

The goal of enactive systems research is the complete treatment of the nature and emergence of autonomous, cognitive, social systems. It is founded on the concept of autopoiesis–literally *self–production*–whereby a system emerges as a coherent systemic entity, distinct from its environment, as a consequence of processes of self-organization. Three orders of system can be distinguished. First-order autopoietic systems correspond to cellular entities that achieve a physical identity through structural coupling with their environment. Second-order systems engage in structural coupling, this time through a nervous system that enables the association of many internal states with the different interactions in which the organism is involved. Third-order systems exhibit coupling between second-order (i.e. cognitive) systems, i.e. between distinct cognitive agents. These third-order couplings give rise to new phenomonological domains: language and a shared epistemology that reflects (but not abstracts) the common medium in which they are coupled. Such systems are capable of three types of behaviour: (i) the instinctive behaviours that derive from the organizational principles that define it as an autopoietic system (and that emerge from the phylogenic[3] evolution of the system), (ii) ontogenic behaviours that derive from the development of the system over its lifetime, and (iii) communicative behaviours that are a result of the third-order structural coupling between members of the society of entities. Linguistic behaviours are the emergent consequence of the third-order structural coupling of a socially cohesive group of cognitive entities.

A key postulate of enactive systems is that reasoning, as we commonly conceive it, is the consequence of reflexive use of the linguistic descriptive abilities to the cognitive agent itself [37]. This is significant: reasoning in this sense is a descriptive phenomenon and is quite distinct from the self-organizing mechanism (i.e. structural coupling and operational closure [37]) by which the system/agent develops its cognitive and linguistic behaviours. Since language (and all inter-agent communication) is a manifestation of high-order cognition, specifically co-determination of consensual understanding amongst phylogenically identical and ontogenically compatible agents, reasoning is actually a product of higher-order social cognitive systems rather than a generative process.

The emergent position is supported by recent results which have shown that a biological organism's perception of its body and the dimensionality and geometry of the space in which it is embedded can be deduced (learned or discovered) by the organism from an analysis of the dependencies between motoric commands and consequent sensory data, without any knowledge or reference to an external model of the world or the physical structure of the organism [53,54]. Thus, the perceived structure of the agent's environment could therefore be a consequence of an effort on the part of brains to account for the dependency between their inputs and their outputs in terms of a small number of parameters. There is in fact no need to rely on the classical idea of an objective a priori model of the external world that is mapped by the sensory apparatus to 'some kind of objective archetype'. The conceptions of space, geometry, and the world that the body distinguishes itself from arises from the sensori–motor interaction of the system, exactly the position advocated in [13]. Furthermore, it is the analysis of the sensory consequences of motor commands that gives rise to these concepts. Significantly, the motor commands are not derived as a function of the sensory data. The primary issue is that sensory and motor information are treated simultaneously, and not from either a stimulus perspective or a motor control point of view.

The enactive approach is mirrored in the ideas of self-maintenant system and recursive self-maintenant systems [55]. Here, autonomy is defined as the property of a system to contribute to its own persistence. Since there are different grades of contribution, there are therefore different levels of autonomy. Self-maintenant systems make active contributions to their own persistence but do not contribute to the maintenance of the conditions for persistence. Conversely, recursive self-maintenant systems do contribute actively to the conditions for persistence and can deploy different processes of self-maintenance depending on environmental conditions.

2.3. Hybrid models

Considerable effort has gone into developing approaches, which combine aspects of the emergent systems and cognitivist systems [3,38,56]. These hybrid approaches have their roots in strong criticism of the use of explicit programmer-based knowledge in the creation of artificially intelligent systems [57] and in the development of active 'animate' perceptual systems [58] in which perception-action behaviours become the focus, rather than the perceptual abstraction of representations. Such systems still use representations and representational invariances but it has been argued that these representations should only be constructed by the system itself as it interacts with and explores the world rather than through a priori specification or programming [38]. Thus, a system's ability to interpret objects and the external world is dependent on its ability to flexibly interact with it and interaction is an organizing mechanism that drives a coherence of association between perception and action. Action precedes perception and 'cognitive systems need to acquire information about the external world through learning or association' [3]. Hybrid systems are in many ways consistent with emergent systems while still exploiting programmer-centred (but not programmer-populated) representations (for example, see [19]).

Recent results in building a cognitive vision system on these principles can be found in [59–61]. This system architecture

[3] Phylogeny is concerned with the configuration of a systems as it evolves from generation to generation. Ontogeny is concerned with the development of the system and its structure within any one generation, i.e. over its life-time.

combines a neural-network perception-action component (in which percepts are mediated by actions through exploratory learning) and a symbolic component (based on concepts—invariant descriptions stripped of unnecessary spatial context—which can be used in more prospective processing such as planning or communication). A biologically motivated system, modelled on brain function and cortical pathways and exploiting optical flow as its primary visual stimulus, has demonstrated the development of object segmentation, recognition, and localization capabilities without any prior knowledge of visual appearance though exploratory reaching and simple manipulation [62]. This hybrid extension of the connectionist system [42] also exhibits the ability to learn a simple object affordance and use it to mimic the actions of another (human) agent. An embodied robotic system that can achieve appearance-based self-localization using a catadioptric panoramic camera and an incrementally constructed robust eigenspace model of the environment is described in [63].

2.4. A short critique

It is important to realize that the foregoing paradigms are not equally mature. Each approach has its own strengths and weaknesses, and its proponents and critics. The arguments in favour of dynamical systems and enactive systems are compelling but the current capabilities of cognitivist systems are actually more advanced. However, cognitivist systems are also quite brittle. It has been argued [64] that cognitivist systems suffer from three problems: the symbol grounding problem (see Section 6.4), the frame problem (the need to differentiate the significant in a very large data-set and then generalize to accommodate new data), and the combinatorial problem. These problems are one of the reasons why cognitivist models have difficulties in creating systems that exhibit robust sensori–motor interactions in complex, noisy, dynamic environments. They also have difficulties modelling the higher-order cognitive abilities such as generalization, creativity, and learning [64]. Enactive and dynamical systems should in theory be much less brittle because they emerge through mutual specification and co-development with the environment, but our ability to build artificial cognitive systems based on these principles is actually very limited at present. To date, dynamical systems theory has provided more of a general modelling framework rather than a model of cognition [64] and has so far been employed more as an analysis tool than as a tool for the design and synthesis of cognitive systems [65,64]. The extent to which this will change, and the speed with which it will do so, is uncertain. Hybrid approaches seem to offer the best of both worlds but it is unclear how well one can combine what are ultimately highly antagonistic underlying philosophies. Opinion is divided, with arguments both for (e.g. [7,60,66]) and against (e.g. [64]).

3. The case for embodiment

Having set the scene, we now turn to the issue of embodiment in cognitive systems and cognitive vision systems. Specifically, we wish to decide whether embodiment is a necessary condition of cognitive vision systems and, if so, what that means in practice. We begin by looking at the issue from both the cognitivist and the emergent perspectives, and then argue the case from a paradigm-independent viewpoint.

3.1. The cognitivist case for embodiment

From the perspective of the cognitivist paradigm, there is no case for embodiment, at least none for it as a mandatory requirement of cognition. Cognitivist systems do not necessarily have to be embodied. The very essence of the cognitivist approach is that cognition comprises computational operations defined over symbolic representations and these computational operations are not tied to any given instantiation. They are abstract in principle. It is for this reason that it has been noted that cognitivism exhibits a form of mind-body dualism [13,67]. Symbolic knowledge, framed in the concepts of the designer, can be programmed in directly and does not have to be developed by the system itself through exploration of the environment. As we have seen, some cognitivist systems do exploit learning to augment or even supplant the a priori designed-in knowledge and thereby achieve a greater degree of adaptiveness, reconfigurability, and robustness. Embodiment may therefore offer an additional degree of freedom to facilitate this learning, but it is by no means necessary.

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

3.2. The emergent case for embodiment

The perspective from emergent systems is diametrically opposed to the cognitivist position. Emergent systems, by definition, must be embodied and embedded in their environment in a situated historical developmental context [13].

To see why embodiment is a necessary condition of emergent cognition, consider what cognition means in the emergent paradigm. It is the process whereby an autonomous system becomes viable and effective in its environment. In this, there are two complementary things going on: one is the self-organization[4] of the system as a distinct entity, and the second is the coupling of that entity with its environment. ‘Perception, action, and cognition form a single process’ [67] of self-organization in the specific context of environmental perturbations of the system. This gives rise to the co-determination of the cognitive system and its environment and thereby to the ontogenic development of the system itself over its lifetime. This development is identically the cognitive process of establishing the space of mutually consistent couplings. Put

[4] The self-organization is achieved through an operationally closed network of activities characterized by circular causality [14] and possibly modelled by a dynamical system defined over a space of order parameters and control parameters.

simply, the system's actions define its perceptions but subject to the strong constraints of continued dynamic self-organization. The space of perceptual possibilities is predicated not on an objective environment, but on the space of possible actions that the system can engage in whilst still maintaining the consistency of the coupling with the environment. These environmental perturbations do not control the system since they are not components of the system (and, by definition, do not play a part in the self-organization) but they do play a part in the ontogenic development of the system. Through this ontogenic development, the cognitive system develops its own epistemology, i.e. its own system-specific knowledge of its world, knowledge that has meaning exactly because it captures the consistency and invariance that emerges from the dynamic self-organization in the face of environmental coupling. Thus, we can see that, from this perspective, cognition is inseparable from 'bodily action' [67]: without physical embodied exploration, a cognitive system has no basis for development.

Although this argument is compelling, it has one weakness: it requires you to accept the legitimacy of the emergent thesis. Many do not. If you do accept it, then the necessity of embodiment follows directly. Can one make an argument for embodiment that does not depend on the axioms of emergent cognition? We make an attempt in Section 3.3.

3.3. A paradigm-independent case

We begin with an assumption: that a cognitive system is an autonomous observer—an entity that sees and perceives. Its empirical knowledge of its environment is, as a consequence, contingent upon its operation in that environment and upon its perception and cognition. This knowledge, therefore, is descriptive: dependent on, and a consequence of, the system's cognitive activities. It is not a causal mechanism by which cognition is effected: it is the product of cognition, not the producer of cognition. It follows that the innate generative process of cognition cannot be based on such descriptions, otherwise, infinite regress ensues: cognition would be a function of description and description would be a function of cognition. Since descriptions are a defining feature of cognition and since descriptions are not intrinsic (and therefore cannot be directly instantiated by outside agencies), the system must be capable of creating its own descriptions. Furthermore, these descriptions must of necessity capture some essence of order and regularity—consistency and invariance—in the environment and of the system's interaction with that environment. Hence, a cognitive system must be capable of exploring and defining the space of interaction between itself and its environment. Thus, it must be embedded in the environment and an active part of that environment. That is, it must be embodied.

Two points should be noted about this argument.

First, we qualified the type of knowledge to be empirical. One can argue that theoretical knowledge, just like empirical knowledge, is also descriptive. However, in this case it is plausible that such knowledge is the product of a reflexive cognitive process involving the same linguistic deliberation that characterizes inter-agent communication, but in this instance turned back on itself in introspective discourse [37].

Second, we assumed the cognitive system is autonomous. This seems to be a natural thing to assume, since all natural cognitive systems display autonomous behaviour. It is a pivotal point, however, since it implies that the system is organizationally distinct from other cognitive systems and, therefore, does not *directly* share another cognitive system's components, processes, or knowledge (it may communicate with another system and because of that communication acquire knowledge, but that's a different matter altogether). If we abandon this assumption, all bets are off as it is not at all obvious that cognition in the absence of the robustness implied by autonomy is meaningful.

4. Shades of embodiment

If either of the two arguments above for embodied perception are valid then it is necessary to consider what exactly it is to be embodied. One form of embodiment, and clearly the type envisaged by proponents of the dynamical systems approach to cognition, is a physically active body capable of moving in space, manipulating its environment, altering the state of the environment, and experiencing the physical forces associated with that manipulation [67]. This 'strong' form of embodiment clearly satisfies the conditions of both arguments for embodiment and it seems to be a good place to begin because, having satisfied the boundary conditions, one can then focus on the core problem: the development of rigorous models of cognitive and perceptual processing. However, many computer vision and cognitive systems researchers have concerns about accepting this scenario as it seems to suggest that the only possible cognitive vision systems are ones that are part of robotic systems. This goes against much of the motivation for the creation of cognitive vision systems: resilience, robustness, re-configurability, open-ended improvement of performance, and especially automatic adaptability to unforseen operating conditions. Robotic applications are certainly not the only ones that can benefit for these capabilities. But yet we seem to be concluding that this is the only domain in which a cognitive system can be developed. There are two issues at stake here: first, is there a 'weaker' form of embodiment that still satisfies the needs of emergent systems, and second, even if there is not, does this necessarily imply that the only domain of application of cognitive systems is robotics?

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

If one looks closely at the emergent paradigm, one finds two cornerstones: the operational closure (or circular causality) of

system, and the structural coupling of the system with its environment. Operational closure by itself does not imply a need for embodiment: it is an organizational principle and applies to systems of many temporal and spatial scales. Coupling with the environment is a little trickier. The key requirement is that the mutual perturbations implied by the coupling, i.e. the mutual system–environment interactions, should be rich enough to drive the ontogenic development but not destructive of the self-organization [37]. It would seem then, that there is nothing in principle that requires the 'action' to be physical in any strong sense and, therefore, that it should be possible to develop an embodied cognitive vision system in any application that offers a suitably rich set of interactions. There is, however, an important caveat. In such a system, there is no guarantee that the resultant cognitive behaviour will be in any way consistent with human models or preconceptions of cognitive behaviour (but that may be quite acceptable, as long as the system performs its task adequately). If we want to ensure compatibility with human cognition, then it would seem that we do indeed have to admit the stronger version of embodiment and adopt a domain of discourse that is the same as the one in which we live: one that involves physical movement, forcible manipulation, and exploration, and perhaps even human form [68].

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

So, exactly what kinds of embodiment are possible? Ziemke has introduced a framework to characterize five different types of embodiment [69,70]. In order of increasing restriction, they are:

Structural coupling between agent and environment in the sense a system can be perturbed by its environment and can in turn perturb its environment.

Historical embodiment as a result of a history of structural coupling;

Physical embodiment in a structure that is capable of forcible action (this excludes software agents);

'Organismoid' embodiment, i.e. organism-like bodily form (e.g. humanoid robots); and

Organismic embodiment of autopoietic living systems.

A few notes about structural coupling are in order. First, it should be noted that the concept of structural coupling originates with Maturana and Varela [51] who also require that the system involved is an autopoietic system. This additional requirement is left implicit in Zimke's papers. Second, autopoiesis is a special type of self-organization (requiring self-specification and self-generation). An autopoietic system is a special type of homeostatic system (i.e. self-regulating system) in that the regulation applies not to some system parameter but to the organization of the system itself. This is reminscent of recursive self-maintenant systems [55], and the concept of circular causality [14]. A significant aspect of autopoiesis is that its function is to 'create and maintain the unity that distinguishes it from the medium in which it exists'.

Despite the current emphasis on embodiment, Ziemke argues that many current approaches in cognitive/adaptive/epigenetic robotics still adhere to the functionalist hardware/software distinction in the sense that the computational model does not in principle require an instantiation (cf. Newell and Simon [71]). Zimke suggests that this is a real problem because the idea of embodiment is that the morphology of the system is actually a key component of the systems dynamics. In other words, morphology not only matters, it is a constitutive part of the self-organization and the structural coupling with the environment. This tight relationship between system morphology and system dynamics (i.e. *cognition*) is frequently reflected too in biological cognitive systems (see Section 6).

There is another aspect to embodiment of a system. This is the environment in which the system is embedded: its physical and social context. In [72], 'cognition in context' is contrasted with 'cognition without context'. The latter is associated with the information processing cognitivist approaches, and is characterized as a process that can be divorced from the entity and the environment with which it is associated. The former is associated with embodied or situated cognition: variously described as 'cognition in the wild' [73], 'situated cognition', and 'natural cognition'.[5]

5. Implications

Apart from the issues of embodiment discussed in the previous section, there are other consequences of adopting an embodied emergent systems approach to the development of cognitive vision systems. We address two of these here.

The first issue is the trade-off between phylogenic configuration and ontogenic development. Phylogeny—the evolution of the system configuration from generation to generation—determines the sensory–motor capabilities that a system is configured with at the outset and that facilitate the system's innate behaviours. Ontogenic development—the adaptation and learning of the system during its lifetime—gives rise to the cognitive capabilities that we seek. Since, we do not have the luxury of having evolutionary timescales to allow phylogenic emergence of a cognitive system (we cannot wait around to evolve a cognitive system from nothing) we must somehow identify a minimal phylogenic state of the system. In practice, this means that we must identify and effect visuo-motor capabilities for the minimal behaviours that ontogenic development will subsequently build on to achieve cognitive behaviour. Put simply, we need to decide what visual processing capabilities are needed for a minimal emergent cognitive vision system. It is a major problem to accomplish

[5] Brooks, on the other hand, distinguishes between *situated* creatures or robots which are embedded in the world but do not deal with abstract descriptions, and *embodied* creatures or robots which possess a physical body and experience the world directly through the influence of the world on that body [68].

this without reverting to cognitivism: i.e. system identification based on representations derived from external observers. However, there is also a second question that is relevant here: what is the correct balance between phylogenic configuration and ontogenic development for a cognitive system in a particular environment? These are difficult questions and we will look at natural systems for some guidance in Section 6.

The requirements of real-time synchronous system–environment coupling and historical, situated, and embodied development have important implications. Specifically, the maximum rate of ontogenic development is constrained by the speed of coupling (i.e. the interaction) and not by the speed at which internal processing can occur [15]. Natural cognitive systems have a learning cycle measured in weeks, months, and years and, while it might be possible to condense these into minutes and hours for an artificial system because of increases in the rate of internal adaptation and change, it cannot be reduced below the time-scale of the interaction.

6. Learning from nature

In attempting to understand and build cognitive systems, it can help to look at how nature deals with cognition. For example, the study of biological systems can help resolve some of the issues surrounding the balance between phylogeny and ontogeny in developing cognitive skills.

In the particular case of cognitive vision systems, although some have argued that basing models on the human visual system has not been very effective [74], the trend today is to exploit new knowledge gained from research in the neurosciences based on, for example, neuroimaging studies using fMRI and PET. For instance, the human expert object recognition pathway has been modelled with multi-scale Gabor filters, feature detectors, non-accidental feature transforms, unsupervised clustering, and subspace projection [75]. Neuroimaging studies have also provided new answers to some long-standing problems in visual attention, a critical issue in cognitive vision [76]. For example, neuroscience has taught us that attention and control of eye position and movement are interlinked, neurophysiologically as well as functionally [77]. Visuospatial attention is significantly modulated by the position of the eye and 'attention cannot be directed towards spatial locations that are difficult for the eye to access' [78]. This co-dependency of perception and action is a recurrent theme in contemporary cognitive systems research and bolsters even further the case for embodied perception [65].

In the following, we revisit the phylogeny/ontogeny trade-off from the perspective of natural species, we look at some examples of innate phylogenically derived capabilities, and we consider some of the issues–such as motivation, imitation, and interation–that are relevant for ontogenic development.

6.1. The phylogeny/ontogeny trade-off: precocial and altricial species

Two types of natural species can be distinguished: precocial and altricial. Precocial species are those that are born or hatched with well-developed behaviours, skills, and abilities, which are the direct result of their genetic make-up (i.e. their phylogenic configuration). As a result, precocial species tend to be quite independent at birth. Altricial species, on the other hand, are born or hatched with poor or undeveloped behaviours and skills, and are highly dependent for support. However, in contrast to precocial species, they proceed to learn complex cognitive skills over their life-time (i.e. through ontogenic development).

Sloman and Chappell argue that, rather than view the precocial/altricial distinction as a simple dichotomy in phylogenic configuration and ontogenic potential, we should view the precocial and altricial as two ends of a spectrum of possible configurations: 'precocial skills can provide sophisticated abilities at birth. Altricial capabilities have the potential to adapt to changing needs and opportunities. So it is not surprising that many species have both' [79].

The challenge then is threefold: to identify the innate precocial skills (which need not come ready-made and may need tuning through reinforcement learning), to establish how altricial capabilities are developed, and to establish the right combination of both when designing systems. The next few sections provide some illustrations of the light that studies of natural systems have shed on these concerns.

6.2. Phylogeny: innate capabilities

The study of the capabilities of newborn human infants (neonates) can be instructive. For example, neonates have a repertoire of coordinated movements which are triggered by sensory stimuli. These movements are not solely action-related but also serve to establish a relationship between vision and proprioception [42]. Neonates also have the ability to control their gaze and direct it to significant sources of information. In addition, they also have an established link between the eye and the hand: newborn infants aim their extended arm movements towards an object upon which they are fixating [80,81] although visual feedback, especially that based on foveal vision, does not play a role [82,83]. Neonates have the ability to perform saccadic tracking of moving objects, using more saccades for smaller objects. As age increases, the number of saccades decreases. They can also perform smooth pursuit, a capability that improves with age but does not depend on object size [84]. Neonates do not have high visual acuity or stereoscopic vision at birth but develop it quickly: but by the fourth or fifth month, visual acuity has increased greatly and about two thirds have stereoscopic vision and depth perception [82]. In humans, colour plays a dual role of both aiding sensory processing, such as segmentation, and cognitive processing, such as representation and recall [85]. Studies in salamanders and rabbits, and extrapolated to humans, have shown that motion anticipation is accomplished at the retinal level and not at the cortical level where other motion processing is done [86].

Inspired by the behaviour of insects (honeybees), one robotic navigation system uses low resolution optical flow to effect a set of reactive navigation behaviours which allow

higher level navigation systems to focus on the goal-oriented tasks [87].

6.3. Ontogeny: modes of learning, and the importance of motivation and exploration

Precocial and altricial skills develop through different types of learning. Precocial skills, based on innate capabilities, are honed through continuous knowledge-free reinforcement-like learning. In this sense, precocial learning is somewhat akin to parameter estimation. On the other hand, altricial skills—which expoit precocial skills—develop through a different form of learning, driven not just by conventional reward/punishment cost functions (positive and negative feedback) but through spontaneous play and exploration which are not directly reinforced [79,88].

Sloman and Chappell argue that the goal of exploration is to discover 'discrete, re-usable, and (recursively) recombinable chunks of information' [79,88]. They note too that the variety of sensory and motor chunks that can be learned will depend crucially on

(1) the morphology, i.e. the physical structure and capabilities of the system (organism/robot);
(2) the richness of the system's environment during learning;
(3) the set of genetically determined internal operations whereby chunks of knowledge can be combined, both with reference to external action and, significantly, with respect to representations of internal actions.

In the same vein, Spelke [89] has suggested that complex cognitive skills may be based on the combination of cognitive capabilities that emerge early in human ontology and phylogeny. She calls these 'core knowledge systems': mechanisms for representing and reasoning about certain types of event and entities that are important in the ecology of the agent, such as inanimate objects that can be manipulated, people, and places, including the motion of these entities, their cardinality, and spatial and numerical relations between them. Spelke argues that since the core systems of infants are very similar to those of non-human animals, it makes sense therefore to study infants and animals to learn more about these core systems, even if non-human animals never develop complex cognitive skills and humans do.

The view that exploration is crucial to altricial or ontogenic development is supported by research findings in developmental psychology. For example, von Hofsten has pointed out that it is not necessarily success at achieving task-specific goals that drives development in neonates but rather the discovery of new modes of interaction: the acquisition of a new way of doing something through exploration [80,90]. In order to facilitate exploration of new ways of doing things, one must suspend current skills. Consequently, ontogenic development differs from learning in that (a) it must inhibit existing abilities, and (b) it must be able to cater for (and perhaps effect) changes in the morphology or structure of the system [82]. The inhibition does not imply a loss of learned control but an inhibition of the link between a specific sensory stimulus and a corresponding motor response.

In addition to the development of altricial skills through exploration (reaching, grasping, and manipulating what's around it), there are two other very important ways in which cognition develops. These are imitation [91,92] and social interaction, including teaching [93]. Unlike other learning methods such as reinforcement learning, imitation—the ability to learn new behaviours by observing the actions of others—allows rapid learning [92]. Metzoff and Moore [94,95] suggest that infants learn through imitation in four phases:

(1) body babbling, involving playful trial-and-error movements;
(2) imitation of body movements;
(3) imitation of actions on objects;
(4) imitation based on inferring intentions of others.

Neonates use body babbling to learn a rich 'act space' in which new body configurations can be interpolated although its significant that even at birth newborn infants can imitiate body movements [92].

In summary, cognitive skills emerge progressively through ontogenic development as it learns to make sense of its world through exploration, through manipulation, imitation, and social interaction, including communication [96]. Proponents of the enactive approach would add the additional requirement that this development take place in the context of a circular causality of action and perception, each a function of the other as the system manages its mutual interaction with the world: essentially co-development of action and perception, and co-determination of the system through self-organization in an ecological and social context.

6.4. The symbol grounding problem

If a cognitive vision system has (or develops) some form of symbolic representation of the world around it—and it seems that in some sense[6] that cognitive systems do develop symbolic representations—how does the representation acquires meaning? How do purely symbolic representations acquire semantic content? This is the so-called symbol grounding problem.

Harnad [97] suggests that symbolic representations must be grounded bottom-up in non-symbolic representations of two kinds: (a) iconic representations, which are derived directly from sensory data, and (b) categorical representations, based on the output of both learned and innate processes that detect invariant features of object and event categories from these sensory data. Higher-order symbolic representations can then be derived from these elementary symbols.

Sloman [79] has argued against this viewpoint. He notes that the internal symbolic representations that are the result of altricial learning (i.e. ontogenic development) are *attached* to the world through sensory perception rather than *grounded*.

[6] But not necessarily in the sense of the physical symbol system hypothesis [71].

The distinction is an important one. Symbol grounding implies that the meaning of a symbol is derived bottom-up by abstraction from direct sensory experience. The need for symbol grounding in this sense is a direct consequence of adopting the more conventional cognitivist approach to cognition (e.g. see [26,36]), exactly because cognitivism invokes this process of abstraction of isomorphic representations of the world. Symbol attachment is quite different. It arises through a rich process of structural coupling with the world. With symbol attachment, the symbols do not derive directly from the sensory data, they derive from the altricial learning (or ontogenic development), the process of developing new chunks of knowledge that are specific to the type of organism or system one is dealing with: its particular set of precocial skills, its altricial learning mechanism, the richness of its surrounding environment, and the particular morphology possessed by the system in its sensory and motoric apparatus. Thus, symbol grounding is required only if one adopts a cognitivist approach; symbol attachment is more neutral in the sense that it makes no strong claims about the isomorphism between world and respresentation, or the necessary uniqueness of these representations. It is also consistent with the enactive viewpoint and the concept of structural coupling and co-determination.[7]

6.5. Perception/action co-dependency

We have already remarked on the co-dependency of perception and action in biological systems. Perceptual development is determined by the action capabilities of a developing child and what observed objects and events afford in the context of those actions [80,98]. It is worth reinforcing this again, especially in the light of recent neurological evidence. For example, the presence of a set of neurons—mirror neurons—is often cited as evidence of the tight relationship between perception and action [99,100]. Mirror neurons are activated both when an action is performed and when the same or similar action is observed being performed by another agent. These neurons are specific to the goal of the action and not the mechanics of carrying it out [80].

In summary, the development of action and perception, the development of the nervous system, and the development (growth) of the body, all mutually influence each other as increasingly sophisticated and increasingly prospective (future-oriented) capabilities in solving action problems are learned [80].

An example of a system, which exploits this co-dependency in a developmental setting can be found in [42]. This is a biologically motivated connectionist system that learns goal-directed reaching using colour-segmented images derived from a retina-like log-polar sensor camera. The system adopts a developmental approach: beginning with innate inbuilt primitive reflexes, it learns sensori–motor coordination. Other biologically motivated work, modelled on brain function and cortical pathways and exploiting optical flow as its primary visual stimulus, has demonstrated the development of object segmentation, recognition, and localization capabilities without any prior knowledge of visual appearance though exploratory reaching and simple manipulation [62]. The system also exhibits the ability to learn a simple object affordance and use it to mimic the actions of another (human) agent.

7. Conclusions

In this paper, we have identified cognitive vision and cognitive systems with two broad positions. For the first —cognitivist—position, cognition entails the manipulation of explicit representations of the state and behaviour of the external world to facilitate appropriate, adaptive, anticipatory, and effective interaction, and the storage of the knowledge gained from this experience to reason even more effectively in the future. Reasoning itself is symbolic: a procedural process whereby explicit representations of an external world are manipulated to infer likely changes in the configuration of the world arising from causal actions.

For the second–emergent–position, cognition is a process of self-organization whereby the system is continually re-constituting itself in real-time to maintain its operational identity through moderation of mutual system-environment interaction and co-determination, particularly over extended timescales. Reasoning (perhaps deliberation would be a more appropriate term in this context) is the consequence of recursive application of the linguistic descriptive abilities (developed as a result of the consensual co-development of an epistemology in a society of phylogenically identical agents) to the cognitive agent itself.

Considering the basis of each approach to cognition, we argued that cognitivist cogitive vision systems do not have to be embodied. On the other hand, emergent self-organizing cognitive vision systems do have to be embodied. In an attempt to break this apparent impasse, we constructed a paradigm-independent argument in support of embodied perception. This argument turned on the necessity for a cognitive system to be able to generate its own empirical knowledge.

Accepting the case for embodiment, we then looked at the nature of embodiment and what it means in practice. We concluded that, in principle and by both arguments for embodied perception, embodiment only implies an ability to interact with the environment (rather than physically and forcibly exploring it) but that in practice if we wish the system to be compatible with human cognition then physical embodiment involving movement, manipulation, and exploration is in fact necessary. However, we also concluded that, in such an eventuality, the resultant cognitive vision system could still be subsequently exploited in embedded passive settings.

We noted two further implications of embodied perception. First, the need to identify and implement the minimal visuo-motor skills required for subsequent ontogenic development

[7] Sloman also points out that the symbol grounding viewpoint is identical with the philosophy of concept empiricism, a philosophy that has been refuted by Kant. On the other hand, symbol attachment is consistent with the phenomonology of Husserl and Heidegger.

and the need to do so in a way that does not prejudice the self-organization of an emergent system but at the same time is sufficient to facilitate it. Second, we noted that embodied development imposes a hard limitation on the speed of development: since the system is dynamically locked to real-time interaction, development progresses at least no faster than the rate of interaction and this is constrained by the dynamics of the system's visual environment.

Finally, we saw how the study of natural systems can shed light on some of the problems posed by cognitive vision.

Acknowledgements

The ideas presented here owe much to discussions with Dermot Furlong, Trinity College Dublin, and with Giulio Sandini, University of Genoa.

Dagstughl seminar 03441 [101] on cognitive vision systems, organized by Henrik Christensen, KTH, and Hans-Hellmut Nagel, Karlsruhe University, and had a strong influence in shaping the the structure of the paper and in achieving what I hope is a balanced perspective.

This work was facilitated by the European Commission as part of the European research network for cognitive vision systems—*ECVision*—under the Information Society Technologies (IST) programme, project IST-2001-35454. It was also strongly influenced by work in the IST RobotCub project 004370. The paper drew on ideas discussed at the brainstorming sessions conducted in the development of the *ECVision* research roadmap [2].

Special thanks go to Colette Maloney, European Commission, for her unstinting and far-sighted support.

Finally, I would like to express my gratitude to the reviewers for their insightful and probing comments. These helped greatly in improving the paper.

References

[1] D. Vernon, The space of cognitive vision, in: H.I. Christensen, H.-H. Nagel (Eds.), Cognitive Vision Systems: Sampling the Spectrum of Approaches, LNCS (In Press), Springer, Heidelberg. pp. 7–26.

[2] P. Auer, et al., A Research Roadmap of Cognitive Vision, ECVision: European Network for Research in Cognitive Vision Systems, 2005, http://www.ecvision.org/research_planning/ECVisionRoadmapv5.0.pdf

[3] G.H. Granlund, Does vision inevitably have to be active?, in: Proceedings of the SCIA99, Scandanavian Conference on Image Analysis. 1999

[4] E. Hollnagel, D.D. Woods, Cognitive systems engineering: new wind in new bottles, International Journal of Human–Computer Studies 51 (1999) 339–356.

[5] R.J. Brachman, Systems that know what they're doing, IEEE Intelligent Systems 17 (6) (2002) 67–71.

[6] H.-H. Nagel, Reflections on cognitive vision systems in: J. Crowley, J. Piater, M. Vincze, L. Paletta (Eds.), Proceedings of the Third International Conference on Computer Vision Systems, ICVS 2003, LNCS 2626, Springer, Berlin, 2003, pp. 34–43.

[7] A. Clark, Mindware—an Introduction to the Philosophy of Cognitive Science, Oxford University Press, New York, 2001.

[8] F.J. Varela, Whence perceptual meaning? a cartography of current ideas in: F.J. Varela, J.-P. Dupuy (Eds.), Understanding Origins—Contemporary Views on the Origin of Life, Mind and Society, Boston Studies in the Philosophy of Science, Kluwer Academic Publishers, Dordrecht (Hingham, MA), 1992, pp. 235–263.

[9] D. Marr, Artificial intelligence—a personal view, Artificial Intelligence 9 (1977) 37–48.

[10] J. Haugland, Semantic engines: an introduction to mind design in: J. Haugland (Ed.), Mind Design: Philosophy, Psychology, Artificial Intelligence, Bradford Books, MIT Press, Cambridge, MA, 1982, pp. 1–34.

[11] S. Pinker, Visual cognition: an introduction, Cognition 18 (1984) 1–63.

[12] J.F. Kihlstrom, The cognitive unconscious, Science 237 (1987) 1445–1452.

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

[14] J.A.S. Kelso, Dynamic Patterns—The Self-Organization of Brain and Behaviour, third ed., MIT Press, Cambridge, MA, 1995.

[15] T. Winograd, F. Flores, Understanding Computers and Cognition—a New Foundation for Design, Addison-Wesley, Reading, MA, 1986.

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

[17] J. Pauli, G. Sommer, Perceptual organization with image formation compatibilities, Pattern Recognition Letters 23 (7) (2002) 803–817.

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

[19] K. Okuma, A. Taleghani, N. de, J. Little, D. Lowe, A boosted particle filter: multitarget detection and tracking in: T. Pajdla, J. Matas (Eds.), Proceedings of the Eighth European Conference on Computer Vision, ECCV 2004, 3021 of LNCS, Springer, Berlin, 2004, pp. 28–39.

[20] H.-H. Nagel, Steps toward a cognitive vision system, AI Magazine 25 (2) (2004) 31–50.

[21] M. Arens, H.-H. Nagel, Quantitative movement prediction based on qualitative knowledge about behaviour, KI-Zeitschrift Künstliche Intelligenz, Special Issue on Cognitive Computer Vision, 2005. pp. 5–11.

[22] M. Arens, H.H. Nagel, Representation of behavioral knowledge for planning and plan recognition in a cognitive vision system in: M. Jarke, J. Koehler, G. Lakemeyer (Eds.), Proceedings of the 25th German Conference on Artificial Intelligence (KI-2002), Springer, Aachen, Germany, 2002, pp. 268–282.

[23] M. Arens, A. Ottlick, H.H. Nagel, Natural language texts for a cognitive vision system in: F.V. Harmelen (Ed.), Proceedings of the 15th European Conference On Artificial Intelligence (ECAI-2002), IOS Press, Amsterdam, 2002, pp. 455–459.

[24] R. Gerber, H.H. Nagel, H. Schreiber, Deriving textual descriptions of road traffic queues from video sequences in: V.H. F (Ed.), Proceeedings of the 15th European Conference on Artificial Intelligence (ECAI-2002), IOS Press, Amsterdam, 2002, pp. 736–740.

[25] R. Gerber, N.H.H., 'occurrence' extraction from image sequences of road traffic, in: Cognitive Vision Workshop, Zurich, Switzerland, 2002.

[26] B. Neumann, R. Möller, On scene interpretation with description logics, in: H.I. Christensen, H.-H. Nagel (Eds.), Cognitive Vision Systems: Sampling the Spectrum of Approaches, LNCS (In Press), Springer, Heidelberg, pp. 235–260.

[27] R. Möeller, B. Neumann, M. Wessel, Towards computer vision with description logics: some recent progress, in: Proceedings of the Integration of Speech and Image Understanding, IEEE Computer Society, Corfu, Greece, 1999, pp. 101–115.

[28] K. Sage, J. Howell, H. Buxton, Recognition of action, activity and behaviour in the actIPret project, KI-Zeitschrift Künstliche Intelligenz, Special Issue on Cognitive Computer Vision, 2005, pp. 30–34.

[29] H. Buxton, A.J. Howell, Active vision techniques for visually mediated interaction, in: International Conference on Pattern recognition, Quebec City, Canada, 2002.

[30] H. Buxton, A.J. Howell, K. Sage, The role of task control and context in learning to recognise gesture, in: Workshop on Cognitive Vision, Zürich, Switzerland, 2002.
[31] H. Buxton, Generative models for learning and understanding dynamic scene activity, in: ECCV Workshop on Generative Model Based Vision, Copenhagen, Denmark, 2002.
[32] A.G. Cohn, D.C. Hogg, B. Bennett, V. Devin, A. Galata, D.R. Magee, C. Needham, P. Santos, Cognitive vision: integrating symbolic qualitative representations with computer vision in: H.I. Christensen, H.-H. Nagel (Eds.), Cognitive Vision Systems: Sampling the Spectrum of Approaches, LNCS, Springer, Heidelberg, 2005, pp. 211–234.
[33] N. Maillot, M. Thonnat, A. Boucher, Towards ontology based cognitive vision in: J. Crowley, J. Piater, M. Vincze, L. Paletta (Eds.), Proceedings of the Third International Conference on Computer Vision Systems, ICVS 2003, LNCS 2626, Springer, Berlin, 2003, pp. 44–53.
[34] J.L. Crowley, Things that see: Context-aware multi-modal interaction, KI-Zeitschrift Künstliche Intelligenz, Special Issue on Cognitive Computer Vision.
[35] E.D. Dickmanns, Dynamic vision-based intelligence, AI Magazine 25 (2) (2004) 10–29.
[36] C. Town, D. Sinclair, A self-referential perceptual inference framework for video interpretation in: J. Crowley, J. Piater, M. Vincze, L. Paletta (Eds.), Proceedings of the Third International Conference on Computer Vision Systems, ICVS 2003, LNCS 2626, Springer, Berlin, 2003, pp. 54–67.
[37] H. Maturana, F. Varela, The Tree of Knowledge—The Biological Roots of Human Understanding, New Science Library, Boston, 1987.
[38] G.H. Granlund, The complexity of vision, Signal Processing 74 (1999) 101–126.
[39] M. Jones, D. Vernon, Using neural networks to learn hand-eye co-ordination, Neural Computing and Applications 2 (1) (1994) 2–12.
[40] B.W. Mel, MURPHY: A robot that learns by doing, in: Neural Information Processing Systems, American Institute of Physics, 1988, pp. 544–553.
[41] R. Linsker, Self-organization in a perceptual network, Computer (1988) 105–117.
[42] G. Metta, G. Sandini, J. Konczak, A developmental approach to visually-guided reaching in artificial systems, Neural Networks 12 (10) (1999) 1413–1427.
[43] R. Reiter, Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems, MIT Press, Cambridge, MA, 2001.
[44] T. van Gelder, R.F. Port, It's about time: an overview of the dynamical approach to cognition in: R.F. Port, T. van Gelder (Eds.), Mind as Motion—Explorations in the Dynamics of Cognition, Bradford Books, MIT Press, Cambridge, MA, 1995, pp. 1–43.
[45] J.J. Gibson, The Perception of the Visual World, Houghton Mifflin, Boston, 1950.
[46] J.J. Gibson, The Ecological Approach to Visual Perception, Houghton Mifflin, Boston, 1979.
[47] W. Köhler, Dynamics in Psychology, Liveright, New York, 1940.
[48] W.H. Warren, Perceiving affordances: visual guidance of stairclimbing, Journal of Experimental Psychology: Human Perception and Performance 10 (1984) 683–703.
[49] H. Maturana, Biology of cognition, Research Report BCL 9.0, University of Illinois, Urbana, IL (1970)
[50] H. Maturana, The organization of the living: a theory of the living organization, International Journal of Man–Machine Studies 7 (3) (1975) 313–332.
[51] H.R. Maturana, F.J. Varela, Autopoiesis and cognition — the realization of the living Boston Studies on the Philosophy of Science, D. Reidel Publishing Company, Dordrecht, Holland, 1980.
[52] F. Varela, Principles of Biological Autonomy, Elsevier North Holland, New York, 1979.
[53] D. Philipona, J.K. O'Regan, J.-P. Nadal, Is there something out there? Inferring space from sensorimotor dependencies, Neural Computation 15 (9).
[54] D. Philipona, J.K. O'Regan, J.-P. Nadal, O. Coenen, Perception of the structure of the physical world using unknown multimodal sensors and effectors in: S. Thrun, L. Saul, B. Schölkopf (Eds.), Advances in Neural Information Processing Systems 16, MIT Press, Cambridge, MA, 2004.
[55] M.H. Bickhard, Autonomy, function, and representation, Artificial Intelligence, Special Issue on Communication and Cognition, 17 (3–4) (2000) 111–131.
[56] G.H. Granlund, Cognitive vision—background and research issues, Research report, Linköping University (2002).
[57] H.L. Dreyfus, From micro-worlds to knowledge representation in: J. Haugland (Ed.), Mind Design: Philosophy, Psychology, Artificial Intelligence, Bradford Books, MIT Press, Cambridge, MA, 1982, pp. 161–204. Excerpted from the Introduction to the second edition of the author's *What Computers Can't Do*, Harper and Row, 1979.
[58] D.H. Ballard, Animate vision, Artificial Intelligence 48 (1991) 57–86.
[59] G. Granlund, A cognitive vision architecture integrating neural networks with symbolic processing, KI-Zeitschrift Künstliche Intelligenz, Special Issue on Cognitive Computer Vision.
[60] G. Granlund, Organization of architectures for cognitive vision systems in: H.I. Christensen, H.-H. Nagel (Eds.), Cognitive Vision Systems: Sampling the Spectrum of Approaches, LNCS, Springer, Heidelberg, 2005, pp. 39–58.
[61] G. Granlund, A. Moe, Unrestricted recognition of 3D objects for robotics using multilevel triplet invariants, AI Magazine 25 (2) (2004) 51–67.
[62] G. Metta, P. Fitzpatrick, Early integration of vision and manipulation, Adaptive Behavior 11 (2) (2003) 109–128.
[63] M. Jogan, M. Artac, D. Skocaj, A. Leonardis, A framework for robust and incremental self-localization of a mobile robot in: J. Crowley, J. Piater, M. Vincze, L. Paletta (Eds.), Proceedings of the Third International Conference on Computer Vision Systems, ICVS 2003, LNCS 2626, Springer, Berlin, 2003, pp. 460–469.
[64] W.D. Christensen, C.A. Hooker, Representation and the meaning of life, in: Representation in Mind: New Approaches to Mental Representation, The University of Sydney, 2000.
[65] G. Metta, D. Vernon, G. Sandini, The robotcub approach to the development of cognition: implications of emergent systems for a common research agenda in epigenetic robotics, in: Proceedings of the Fifth International Workshop on Epigenetic Robotics (EpiRob2005), (to appear)
[66] J.P. Crutchfield, Dynamical embodiment of computation in cognitive processes, Behavioural and Brain Sciences 21 (5) (1998) 635–637.
[67] E. Thelen, Time-scale dynamics and the development of embodied cognition in: R.F. Port, T. van Gelder (Eds.), Mind as Motion—Explorations in the Dynamics of Cognition, Bradford Books, MIT Press, Cambridge, MA, 1995, pp. 69–100.
[68] R.A. Brooks, Flesh and Machines: How Robots Will Change Us, Pantheon Books, New York, 2002.
[69] T. Ziemke, Are robots embodied? in: Balkenius, Zlatev, Dautenhahn, Kozima, Breazeal (Eds.), Proceedings of the First International Workshop on Epigenetic Robotics — Modeling Cognitive Development in Robotic Systems, 85, Lund University Cognitive Studies, Lund, Sweden, 2001, pp. 75–83.
[70] T. Ziemke, What's that thing called embodiment? in: Alterman, Kirsh (Eds.), Proceedings of the 25th Annual Conference of the Cognitive Science Society, Lund University Cognitive Studies, Lawrence Erlbaum, Mahwah, NJ, 2003, pp. 1134–1139.
[71] A. Newell, H.A. Simon, Computer science as empirical inquiry: Symbols and search, Communications of the Association for Computing Machinery, tenth Turing award lecture, ACM 1975 vol. 19 (1976) pp. 113–126.
[72] E. Hollnagel, The substance of cognitive modelling 2002. Available from: http://www.ida.liu.se/~eriho/CognitiveModels_M.htm
[73] E. Hutchins, Cognition in the Wild, MIT Press, Cambridge, MA, 1995.
[74] M. Shah, Guest introduction: the changing shape of computer vision in the twenty-first century, International Journal of Computer Vision 50 (2) (2002) 103–110.
[75] B.A. Draper, K. Baek, J. Boody, Implementing the expert object recognition pathway in: J. Crowley, J. Piater, M. Vincze, L. Paletta (Eds.), Proceedings of the Third International Conference on Computer Vision Systems, ICVS 2003, LNCS 2626, Springer, Berlin, 2003, pp. 1–11.

[76] J.K. Tsotsos, Cognitive vision need attention to link sensing with recognition in: H.I. Christensen, H.-H. Nagel (Eds.), Cognitive Vision Systems: Sampling the Spectrum of Approaches, LNCS, Springer, Heidelberg, 2005, pp. 27–38.
[77] L. Itti, C. Koch, Visual attention: insights from brain imaging, Nature Reviews 2 (2001) 194–203.
[78] L. Craighero, M. Nascimben, L. Fadiga, Eye position affects orienting of visuospatial attention, Current Biology 14 (2004) 331–333.
[79] A. Sloman, J. Chappell, The altricial-precocial spectrum for robots, in: IJCAI'05—19th International Joint Conference on Artificial Intelligence, Edinburgh. 2005. www.cs.bham.ac.uk/research/cogaff/alt-prec-ijcai05.pdf
[80] C. von Hofsten, An action perspective on motor development, Trends in Cognitive Science 8 (2004) 266–272.
[81] C. von Hofsten, Eye-hand coordination in newborns, Developmental Psychology 18 (1982) 450–461.
[82] G. Sandini, G. Metta, J. Konczak, Human sensori-motor development and artificial systems, 1997
[83] B. Sivak, C.L. MacKenzie, Integration of visual information and motor output in reaching and grasping: the contribution of peripheral and central vision, Neuropsychologica 28 (1990) 1095–1116.
[84] K. Rosander, C. von Hofsten, Development of gaze tracking of small and large objects, Experimental Brain Research 146 (2002) 257–264.
[85] K.R. Gegenfurtner, J. Rieger, Sensory and cognitive contributions of color to the recognition of natural scenes, Current Biology 10 (2002) 805–808.
[86] K.R. Gegenfurtner, The eyes have it, Nature 398 (1999) 291–292.
[87] J. Santos-Victor, G. Sandini, Embedded visual behaviours for navigation, Robotics and Autonomous Systems 19 (1997) 299–313.
[88] A. Sloman, J. Chappell, Altricial self-organising information-processing systems, in: International Workshop on the Grand Challenge in Non-classical Computation, York, 2005, www.cs.bham.ac.uk/research/cogaff/summary-gc7.pdf
[89] E.S. Spelke, Core knowledge, American Psychologist (2000) 1233–1243.
[90] C. von Hofsten, On the development of perception and action in: J. Valsiner, K.J. Connolly (Eds.), Handbook of Developmental Psychology, Sage, London, 2003, pp. 114–140.
[91] A. Billard, Imitation in: M.A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, MIT Press, Cambridge, MA, 2002, pp. 566–569.
[92] R. Rao, A. Shon, A. Meltzoff, A bayesian model of imitation in infants and robots in: K. Dautenhahn, C. Nehaniv (Eds.), Imitation and Social Learning in Robots, Humans, and Animals: Behaviour, Social and Communicative Dimensions, Cambridge University Press, MA, 2004.
[93] K. Dautenhahn, A. Billard, Studying robot social cognition within a developmental psychology framework, in: Proceedings of the Eurobot 99: Third European Workshop on Advanced Mobile Robots, Switzerland, 1999, pp. 187–194.
[94] A.N. Meltzoff, M.K. Moore, Explaining facial imitation: a theoretical model, Early Development and Parenting 6 (1997) 179–192.
[95] A.N. Meltzoff, The elements of a developmental theory of imitation in: A.N. Meltzoff, W. Prinz (Eds.), The Imitative Mind: Development, Evolution, and Brain Bases, Cambridge University Press, Cambridge, 2002, pp. 19–41.
[96] G. Sandini, G. Metta, D. Vernon, Robotcub: an open framework for research in embodied cognition, in: IEEE-RAS/RSJ International Conference on Humanoid Robots (Humanoids 2004), 2004, pp. 13–32.
[97] S. Harnad, The symbol grounding problem, Physica D 42 (1990) 335–346.
[98] E.J. Gibson, A. Pick, An Ecological Approach to Perceptual Learning and Development, Oxford University Press, Oxford, 2000.
[99] V. Gallese, L. Fadiga, L. Fogassi, G. Rizzolatti, Action recognition in the premotor cortex, Brain 119 (1996) 593–609.
[100] G. Rizzolatti, L. Fadiga, V. Gallese, L. Fogassi, Premotor cortex and the recognition of motor actions, Cognitive Brain Research 3 (1996) 131–141.
[101] H.I. Christensen, H.-H. Nagel, Report on dagstuhl seminar 03441: Cognitive Vision Systems 2003 Available from: http://www.dagstuhl.de/03441/Report/