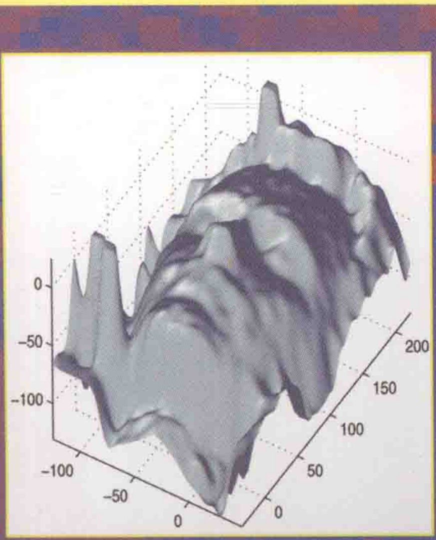
State-of-the-Art
Survey

LNCS 3948

Henrik I. Christensen
Hans-Hellmut Nagel (Eds.)

Cognitive Vision Systems

Sampling the Spectrum of Approaches



Springer

Henrik I. Christensen Hans-Hellmut Nagel (Eds.)

Cognitive Vision Systems

Sampling the Spectrum of Approaches

Springer

Volume Editors

Henrik I. Christensen
Royal Institute of Technology
Centre for Autonomous Systems
100 44 Stockholm, Sweden
E-mail: hic@nada.kth.se

Hans-Hellmut Nagel
Universität Karlsruhe
Fakultät für Informatik
Institut für Algorithmen und Kognitive Systeme
76128 Karlsruhe, Germany
E-mail: nagel@iaks.uni-karlsruhe.de

Library of Congress Control Number: 2006926926

CR Subject Classification (1998): I.4, I.2.9-10, I.2.6, I.5.4-5, F.2.2

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

ISSN 0302-9743
ISBN-10 3-540-33971-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-33971-7 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Markus Richter, Heidelberg
Printed on acid-free paper SPIN: 11414353 06/3142 5 4 3 2 1 0

Preface

During the last decade of the twentieth century, computer vision made considerable progress towards consolidation of its fundaments, in particular regarding the treatment of geometry for the evaluation of stereo image pairs and of multi-view image recordings. Scientists thus began to look at basic computer vision solutions – irrespective of the well-perceived need to perfect these further – as components which should be explored in a larger context.

In 2000, Horst Forster, Head of Division in the Information Society Directorate-General of the European Commission, through his contacts with many computer vision researchers throughout Europe, sensed their readiness to cooperate for the exploration of new grounds in a direction subsequently to become known as 'cognitive vision.' Horst Forster succeeded in convincing the European Commission to stimulate cooperation in this direction by funding a four-year program, which encountered an unexpectedly broad response. It has been a privilege for us to have had a glimpse at the unobtrusive, effective engagement of Horst Forster to advance scientific cooperation within the European Union.

It is a particular pleasure for us to thank Colette Maloney, who closely cooperated with Horst Forster throughout the past by accompanying the many projects funded under the cognitive vision programme. Her constant encouraging support, her practically instant response to a seemingly endless series of calls for help in organizational and financial matters, and her deep commitment to advancing scientific research in this topical area across Europe made collaboration with her a truly memorable experience.

As part of the efforts to further strengthen cooperation between research groups from different countries, a seminar was organized at Schloss Dagstuhl in Germany during October 26–30, 2003. Scientists active in related areas were invited from across the world. This seminar was co-sponsored by ECVision, the Cognitive Vision network of excellence under the leadership of David Vernon. The support from ECVision was instrumental to the organization of this seminar and the creation of this volume. Presentations and associated vivid discussions at the seminar were gradually transformed into a set of contributions to this volume. The editors thank the authors for their considerable efforts to draft, refine, and cross-reference these contributions.

The editors are grateful to Alfred Hofmann from Springer for agreeing to publish this book – and for his patience while we wrestled with the 'mechanics' to put it together.

All who participated in this seminar still remember the warm hospitality and quiet efficiency of the staff at Schloss Dagstuhl who thereby contributed significantly to turning this endeavor into a stimulating and successful event.

February 2006 Henrik I. Christensen and Hans-Hellmut Nagel

Contents

Part III Learning and Adaptation

Part IV Representation and Inference

Part V Control and Systems Integration

Part VI Conclusions

Part VII References, Subject Index, Author Index

1

Introductory Remarks

H.I. Christensen[1] and H.-H. Nagel[2]

[1] Kungliga Tekniska Högskolan
100 44 Stockholm, Sweden
`hic@nada.kth.se`

[2] Institut für Algorithmen und Kognitive Systeme,
Fakultät für Informatik der Universität Karlsruhe (TH)
76128 Karlsruhe, Germany
`nagel@iaks.uni-karlsruhe.de`

The notion 'cognitive vision system (`CogVS`)' stimulates a wide spectrum of associations. In many cases, the attribute 'cognitive' is related to advanced abilities of living creatures, in particular of primates. In this context, a close association between the terms 'cognitive' and 'vision' appears natural, because it is well known that vision constitutes the primate sensory channel with the largest spatiotemporal bandwidth.

Since the middle of the last century, technical means were gradually developed to record and process digitized image sequences. These technical advances created a seemingly unresistable challenge to devise algorithmic approaches which explain, simulate, or even surpass vision capabilities of living creatures. In this context, 'vision' is understood to refer to a set of information processing steps which transform the light intensity distribution impinging onto the transducer surface eventually into some kind of re-action, be it an observable movement, some acoustical communication, or a change of internal representations for the union of the depicted scene and the 'vision system' itself. The common understanding of 'vision' as a kind of information processing induces the use of the word 'system' in this context for whatever performs these processing steps – be it a living creature, a familiar digital computer, or any other alternative to realize a computational device.

The premises underlying such a view have been accepted to the extent that an attribute like 'cognitive' appears applicable to technical constructs despite the fact that it has been coined originally in order to characterize abilities of living creatures. Similar to the experience with other natural language terms referring to commonsense notions – like, e. g., 'intelligence' – scientific efforts to conceive an artifact, which could be considered equivalent to living creatures regarding its input/output relations, are accompanied by efforts to define precisely the notion involved, in our case 'cognitive vision'.

It should not come as a surprise that such endeavors result in a large spectrum of definitions. This observation can be attributed to the fact that complex abilities of living creatures involve many aspects, which have to be taken into account. It sometimes is useful to ask which among these aspects have been selected – or emphasized – in order to motivate a definition of the notion 'cognitive vision system'. Three aspects in

H.I. Christensen and H.-H. Nagel (Eds.): Cognitive Vision Systems, LNCS 3948, pp. 1–4, 2006.

particular appear frequently, either explicitly or implicitly, namely wide applicability, robustness, and speed. The first aspect mentioned implies that a 'true CogVS' can easily and reliably adapt to a wide variation of boundary conditions under which it is expected to operate. This implication rules out the possibility that a CogVS is endowed right from the start with 'all the knowledge' it might need in order to cope with new tasks. It is assumed instead that a CogVS can *learn* task-relevant spatiotemporal structures in its environment and can *adapt* its internal operational parameters in order to reliably estimate the current status of itself and of its environment. 'Robustness' implies that small variations of the environmental state, which are considered to be irrelevant for the execution of the current task, should not influence the performance. And 'speed' implies that the CogVS operates fast enough that task-relevant changes in the environment can be handled without endangering the desired performance level. This latter aspect became important once a 'vision system' had to provide sensory feedback for a mechanical system, in particular for the case of computer vision in the feedback loop of a moving or manipulating artifact.

Although such goals were propagated already rather early during the development of computer vision systems, it turned out that at most two of these three goals could be attained at the same time. If a system was claimed to be (more) widely applicable and robust, it was not fast enough. If it was robust and fast, it was not widely applicable (e.g. specialized machine vision systems for quality control in semi-automated manufacturing plants). And if a system approach was touted as fast and widely applicable, it usually was not robust – if it worked at all. Given our current understanding about the computational expenses required to even determine a small set of visual features reliably, this state of affairs is most plausible even almost up to present days. Ten or twenty years ago, when memory and processing capacity were smaller by three to four orders of magnitude compared to what is available at the same price today, many 'simplifications' or 'speed-ups' were simply a matter of necessity in order to be able to explore an experimental approach at all.

A frequently encountered argument in connection with a CogVS simply quotes that 'there is nothing new under the sun – in German: Alles schon dagewesen' (attributed to Rabbi Ben Akiba). As with the Delphi Oracle, the truth of such a statement can be 'proven' by choosing an appropriate point of view for the interpretation. Rather than burying the topic based on such an adage, it appears more fruitful to inquire in detail which changes or advances of the State-of-the-Art may justify to re-approach previously treated and subsequently abandoned problems. As mentioned already, the still exponential improvement of the price/performance ratio for digital memory and processors let it appear feasible that real-time processing of a video input stream does no longer compromise the quality of elementary signal processing steps to the extent that only rather brittle results could be expected. In addition, size, weight, and power consumption of today's computers and cameras allow to incorporate them into mobile experimental platforms (embodied computer vision systems). Advantages related to the fact that at least part of the system environment may 'serve as its own representation' removes many bottlenecks. A continuously updated state estimate can be used instead of time-consuming searches for the 'optimal currently appropriate hypothesis' about the state of the system and its environment.

The ability to experiment more – with the ensuing advantage of being able to study the gamut of influences in order to separate between genuinely important task conditions from mere disturbances (noise) – gradually provided a background of reproducible experiences. This in turn stimulated more theoretically oriented research which enabled to isolate critical conditions and to prepare measures to prevent them or to cope with them. Choosing a sufficiently large support area for the estimation of gray-value gradients and the exploitation of a search across a range of spatial scales may serve as examples. On top of such efforts, stochastic techniques are applied in order to cope with the unavoidable influences of noise.

In addition to these considerations, another aspect appears to become even more important. The extraction of local features (edge elements, corner points, texture elements, ...) and their aggregation to non-local descriptions (image segments, 3D-surface-facets in the scene) have matured to the point where it becomes possible to abstract from a quantitative geometric characterisation of relevant phenomena to a conceptual level of representation. At this latter level of representation, logic-based inference processes may be activated, which facilitate non-numerical consistency checks. The incorporation of such logic-based processes into a computer vision system offers three important advantages. It first allows to exploit general knowledge about spatiotemporal relations. This in turn allows to generate warnings when 'implausible' situations occur or to circumvent these altogether. And last, but not least, it simplifies the interface between the 'system' and its user – be it the developer or someone who has to supervise an operational system.

It increasingly appears justifiable, therefore, to speak about CogVS: current experimental systems begin to exhibit performance characteristics which start to reduce the qualitative difference to the performance of living creatures to a – still formidable – quantitative difference. This fact in turn opens the road for small, but effective gradual improvements of system performance. Growing familiarity with the 'real effects', which influence system performance, is likely to improve our understanding for the 'solutions' applied by living creatures. Eventually, the attribute 'cognitive' of a CogVS may become appropriate even in the sense that certain aspects of information processing by living creatures can be described in a suitable manner, quite apart from the view that 'cognitive' addresses processing at the conceptual level of representation.

The contributions collected in this volume originated in presentations at a Dagstuhl Seminar (# 03441, 26-31 October 2003) on Cognitive Vision Systems. It samples various views on what constitutes a CogVS and why. In order to preserve the wide spectrum of opinions and thereby to stimulate the debate about characteristics, means, and goals of building a CogVS, dogmatic decisions have been avoided regarding what is and is not a CogVS. The contributions have been grouped, however, with the aim to emphasize similarities in subgoals and/or approaches.

Part I (*Foundations of Cognitive Vision Systems*) collects contributions which address questions concerning the definition and overall structure of what appears to constitute a CogVS. Part II (*Recognition and Categorization*) is concerned with investigations which study the extraction and aggregation of features from video signals with the goal to establish a relation between a subset of these features and conceptual representations for observable bodies and behaviors recorded in the scene. Part III (*Learning and Adaptation*) concentrates on investigations which attempt to broaden the applicability of a system – or to reduce the necessity of interactive tuning phases – by machine learning approaches and by automatic parametric optimization. Part IV (*Representation and Inference*) collects contributions which study the exploitation of representations for inference processes, in particular inference processes based on (variants of) predicate logic. Part V (*Control and Systems Integration*) specifically addresses problem areas which become important due to the necessity to integrate a large and diverse set of processes into a coherent system. In this context, it becomes unavoidable to cope with limited resources – rather than having the system bogging down at unexpected times without reasons discernible from the outside. An attempt is made in a concluding section (Part VI) to condense the insights from this seminar into a small number of theses which could provide a starting point for future investigations.

Deliberately, no attempt is made at this point to condense even further the information formulated by authors of contributions to this volume in abstracts, introductory and concluding sections. Readers are invited to browse and most likely will find that the time spent doing this will have been worthwhile.

Part I

Foundations of Cognitive Vision Systems

2

The Space of Cognitive Vision

David Vernon

DIST, University of Genova, Italy
vernon@ieee.org

Abstract. Cognitive vision is an area that is not yet well-defined, in the sense that one can unambiguously state what issues fall under its purview and what considerations do not. Neither is there unequivocal consensus on the right approach to take in addressing these issues — there isn't a definitive universally-accepted scientific theory with 'gaps in understanding' that merely need to be plugged. On the contrary, there are clearly competing viewpoints and many poorly-understood issues (such as the point where vision stops and cognition starts). Depending on how you choose to view or define cognitive vision, there are many points of departure, some based squarely in artificial intelligence and image processing, others in developmental psychology and cognitive neuroscience, and others yet in cognitive robotics and autonomous systems theory. This paper is an attempt to sketch a framework within which the complete domain of cognitive vision can be set, a framework that embraces all of the possible approaches that can be taken and that highlights common concerns as well as fundamental differences between the approaches. Our goal here is to define cognitive vision in a way that avoids alienating any particular community and to state what the options are. While we will note in passing possible strengths and weaknesses of the various approaches, this paper will not attempt to argue in favour of one approach over another.

2.1 The Background to Cognitive Vision

It is nearly forty years since Roberts first published the results of his seminal attempts to construct a computer vision system [374]. Since then, computer vision has matured and undergone many stages in its evolution. From the blocks-world approaches of the sixties and early seventies [164, 201, 483, 419], to the knowledge-based and model-based approaches of the mid to late seventies [23, 171, 446, 54], the modular information processing approaches of the late seventies and early eighties with their strong emphasis on early vision [278, 30, 280, 283, 284, 282, 193, 281], the development of appearance-based vision in the nineties [81] – a decade that was perhaps distinguished more than anything by the creation of mathematically-sound robust early vision and the associated expansion of vision based on computational geometry [117, 175] – to the more recent probabilistic techniques and the increasingly-widespread use of machine learning [355]. On the way, computer vision has spawned a number of successful offshoots, such as machine vision for industrial inspection, the analysis of video data for remote monitoring of events, and the use of image analysis in the creation of special effects in the film

H.I. Christensen and H.-H. Nagel (Eds.): Cognitive Vision Systems, LNCS 3948, pp. 7–24, 2006.

industry. However, to date, the ultimate goal of creating a general-purpose vision system with anything close to the robustness and resilience of the human visual system remains as elusive as ever.

One of the more recent trends in computer vision research in the pursuit of human-like capability is the coupling of cognition and vision into cognitive computer vision. Unfortunately, it is apparent that the term cognitive computer vision means very different things to different people. For some, it means the explicit use of knowledge and reasoning together with sensory abstraction of data from a perceived environment; for others it implies the emergent behaviour of a physically-active system that learns to make perceptual sense of its environment as it interacts within that environment and as a consequence of that interaction. For others yet, it is a meaningless term in its own right and cannot be treated except as an intrinsic component of the process of cognition that, in turn, is an inherent feature of autonomous systems. Our goal here is to present all of these viewpoints in a single consistent framework:

1. To provide a definition of cognitive vision that is neutral with respect to possible approaches and to explain what capabilities might be provided by such a system;
2. To delineate the space of cognitive vision and characterize it in terms or dimensions that allow it to be mapped on to different approaches;
3. To highlight contentious and significant issues (*e.g.* the necessity for embodiment, the nature and need for representations, the nature and role of knowledge, the role of language, the inter-dependence of perception and action).

These are the issues to which we now turn.

2.2 Towards a Universal Definition of Cognitive Vision

There are several ways one can approach the definition of a discipline. One can take a functional approach, setting out the minimal tasks that a system should be able to carry out, or one can take an architectural approach, identifying the manner in which a system should be constructed and the functional modules that would be used in a typical system. Alternatively, one can adopt a behavioural but non-functional approach that identifies generic attributes, capabilities, and characteristics. A good definition should be neutral to any underlying model, otherwise it begs the research question and preempts the research agenda. Consequently, this rules out an architectural definition. A good definition should also be application-independent. This rules out a strictly functional definition, or at the very least necessitates that any functions be generic and typically common to all possible systems. Consequently, we will attempt to define cognitive vision using generic functionality (*i.e. capability*) and non-functional attributes.

We'll begin with the definition adopted by *ECVision* to date [12]:

> "Cognitive computer vision is concerned with integration and control of vision systems using explicit but not necessarily symbolic models of context, situation and goal-directed behaviour. Cognitive vision implies functionalities for knowledge representation, learning, reasoning about events & structures, recognition and categorization, and goal specification, all of which are concerned with the semantics of the relationship between the visual agent and its environment.'

Although this definition is useful, in that it focusses on many of the key issues, it depends a little too much on architectural issues (*e.g.* integration, control, functional modules) and it is not as neutral to underlying model(s) as perhaps it should be. That is, it strays from a definition of *what* cognitive vision is to a definition of *how* it is to be achieved. As we will see in Section 2.3, there are several competing approaches, not all of which are compatible with the one that is implicitly favoured in this definition. That said, however, it does provide us with a good starting point and the following is an attempt both to expand on it, drawing out the key issues even more, eliminating the model-dependent and architecture-specific components, and highlighting the generic functionalities and non-functional attributes.

> A cognitive vision system can achieve the four levels of generic functionality of a computer vision system:[1]
>
> 1. *Detection* of an object or event in the visual field;
> 2. *Localization* of the position and extent of a detected entity;
> 3. *Recognition* of a localized entity by a labelling process;
> 4. *Understanding* or comprehending the role, context, and purpose of a recognized entity.[2]
>
> It can engage in purposive goal-directed behaviour, adapting to unforeseen changes of the visual environment, and it can anticipate the occurrence of objects or events. It achieves these capabilities through:
>
> 1. a faculty for learning semantic knowledge (*i.e.* contextualized understanding of form and function), and for the development of perceptual strategies and behaviours;
> 2. the retention of knowledge about the environment, the cognitive system itself, and the relationship between the system and its environment;[3]
> 3. deliberation about objects and events in the environment, including the cognitive system itself.

This definition focusses on what constitutes a cognitive vision system, how it should behave, what it should be capable of achieving, and what are its primary characteristics. The first four points encapsulate generic functionality. The next set of issues deal with non-functional attributes, and the final three points suggest a way of spanning the space of cognitive vision.

The three non-functional characteristics of purposive behaviour, adaptability, and anticipation, taken together, allow a cognitive vision system to achieve certain goals, even in circumstances that were not expected when the system was being designed. This capacity for plastic resilient behaviour is one of the hallmarks of a cognitive vision system. The characteristic of anticipation is important as it requires the system to operate

[1] These four levels were suggested by John Tsotsos, York University, during the course of Dagstuhl Seminar 03441[73].

[2] Implicit in the fourth level is the concept of categorization: the assignment of an object or event to a meta-level class on some basis other than visual appearance alone.

[3] The distinction between environmental states, system states, and the environment-system relationship was introduced by Hans-Hellmut Nagel, Universität Karlsruhe, during the course of Dagstuhl Seminar 03441[73].

across a variety of time-scales, extending into the future, so that it is capable of more than reflexive stimulus-response behaviour.

The final three characteristics of cognitive vision — learning, memory, and deliberation — are all concerned with knowledge: its acquisition, storage, and usage. Knowledge is the key to cognitive vision. These three issues highlight the chief differentiating characteristics of cognitive vision *vis-à-vis* computer vision and, as we will see, allow us to define the space of cognitive vision in a way that is relevant to all the various approaches.

First, however, we must survey the different paradigms or approaches that attempt to model and effect these characteristics of cognitive vision.

2.3 A Review of Approaches to Cognition

If we are to understand in a comprehensive way what is meant by cognitive vision, we must address the issue of cognition. Unfortunately, there is no universally-accepted agreement on what cognition is and different research communities have fundamentally different perspectives on the matter.

Broadly speaking, we can identify two distinct approaches to cognition, each of which makes significantly different assumptions about the nature of cognition, the purpose or function of cognition, and the manner in which cognition is achieved. These are:

1. the *cognitivist* approach based on information processing symbolic representational systems;
2. the *emergent systems* approach, embracing connectionist systems, dynamical systems, and enactive systems.

Cognitivist approaches correspond to the classical and still prevalent view that 'cognition is a type of computation' which operates on symbolic representations, and that cognitive systems 'instantiate such representations physically as cognitive codes and ... their behaviour is a causal consequence of operations carried out on these codes' [360]. Connectionist, dynamical, and enactive systems can be grouped together under the general heading of emergent systems that, in contradistinction to the cognitivist view, argues against the information processing view of cognition as 'symbolic, rational, encapsulated, structured, and algorithmic', and argues in favour of one that treats cognition as emergent, self-organizing, and dynamical [447, 219].

2.3.1 Symbolic Information Processing Representational Cognitivist Models

Cognitive science has its origins in cybernetics (1943-53), following the first attempts to formalize what had to that point been metaphysical treatments of cognition. The intention of the early cyberneticians was to create a science of mind, based on logic. Examples of progenitors include McCulloch and Pitts and their seminal paper 'A logical calculus immanent in nervous activity' [294]. This initial wave in the development of a science of cognition was followed in 1956 by the development of an approach