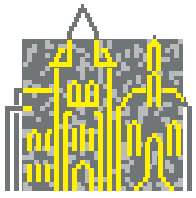
H. Christensen (Stockholm, S), H.-H. Nagel (Univ. Karlsruhe, D)
(Editors)

Cognitive Vision Systems

Dagstuhl Seminar 03441 – October 26 to October 31, 2003
Dagstuhl-Seminar-Report No. 400



SCHLOSS DAGSTUHL

INTERNATIONALES
BEGEGNUNGS-
UND FORSCHUNGSZENTRUM
FÜR INFORMATIK

ISSN 0940-1121

Herausgegeben von IBFI gem. GmbH, Schloss Dagstuhl, 66687 Wadern, Germany.

Das Internationale Begegnungs- und Forschungszentrum für Informatik (IBFI) Schloss Dagstuhl ist eine gemeinnützige GmbH. Sie veranstaltet regelmäßig wissenschaftliche Seminare, welche nach Antrag der Tagungsleiter und Begutachtung durch das wissenschaftliche Direktorium mit persönlich eingeladenen Gästen durchgeführt werden.

Gesellschafter:

- Gesellschaft für Informatik e.V. – Bonn
- TH Darmstadt
- Universität Frankfurt
- Universität Kaiserslautern
- Universität Karlsruhe
- Universität Stuttgart
- Universität Trier
- Universität des Saarlandes

Report on Dagstuhl Seminar 03441:

Cognitive Vision Systems

(26–31 October 2003)

H. I. Christensen[†] and H.–H. Nagel[‡]

[†] Numerical Analysis and Computer Science
Kungliga Tekniska Högskolan
100 44 Stockholm, Sweden
hic@nada.kth.se

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

Abstract. Early attempts to integrate AI and Computer Vision failed due to lack of robust vision techniques for the derivation of symbolic descriptions of the 'meaning' of images, and the lack of AI techniques to handle information with associated uncertainty. Over the last decade, significant progress has been achieved in Computational Vision, AI, and computer platforms.
Regarding Computational Vision, the basis in terms of generating a representation of the system environment through use of robust methods is not yet particularly strong. At the same time, the AI community has established new paradigms for handling uncertain information and scalable models. In parallel to these developments, the progress in the design and production of highly integrated circuits and computer programming systems has resulted in a system performance that facilitates real-time generation and processing of information even from video input streams.
The seminar discussed models for Cognitive Vision Systems (CVS) in terms of system layout and components. In addition, both Computer Vision and AI techniques as components of systems were presented. This seminar also involved discussions on the conceptual basis for Cognitive Vision and the feasibility of constructing computational systems that have 'cognitive' functionality.

1 Introductory Remarks

Prepared contributions during this seminar on 'Cognitive Vision Systems' stimulated considerable discussions centered around two views:

– How to delineate research efforts associated with a 'restricted' understanding of 'Cognitive Vision *Systems* (CVS)' against a 'broader' understanding, the latter being associated in particular with the adjective 'cognitive'?

– How to structure the research area to be subsumed under a 'restricted' understanding of a Cognitive Vision System?'

The first – more outwards oriented – view addresses the *definition* of this topical area, the second – more inwards oriented – view addresses its internal *structure*. The difference between these two views resurfaced repeatedly during the entire seminar. Most participants, though, gradually accepted the attitude that a considerable number of aspects needs to be taken into account in order to achieve a balanced view on this topical area.

2 On a Delineation of the Notion 'Cognitive Vision System'

Not surprisingly, the adjective 'cognitive' stimulated many associations, extending from neurophysiological and psychophysical research on the function of the human visual cortex via Cognitive Psychology and Cognition out to philosophical questions. *E. Granum* reminded the audience that one should not pick a more or less arbitrary part of what constitutes a human personality, analyse this part based on strictly technical notions, and then hope to be able to generalise any insights without creating difficulties for others inside and outside this area of scientific activities.

It was generally accepted that any CVS should not only be applicable to a wide range of tasks, but in addition should be able to adapt to its current environment. It was similarly agreed that 'vision' offers a very important, but by no means the only relevant sensory input channel, in particular that it is in general advantageous to evaluate the association between several signal modalities.

The survey by *D. Vernon* acknowledges these difficulties by an analysis of assumptions underlying different approaches to define a CVS. He suggests to cluster assumptions isolated by his analysis into four different types of approaches one of which is restricted to a kind of 'engineered-type' of CVS, to be separated by methods and claims from other, more broadly aspiring approaches (see his abstract in Section B.24). *G. Granlund* circumscribed his understanding of a CVS by a sort of 'constructive' approach: he distinguishes three phases in a kind of system-evolution. (i) A root system is endowed by inheritance or by engineering with certain basic capabilities, which then facilitate (ii) an 'exploratory phase' by repeatedly performing 'action-perception cycles' in order to generate concepts via associative learning. These concepts provide the basis for a subsequent (iii) refinement of the space of usable concepts through either 'passive' observation or a kind of language-based communication with the system environment. Arguments justifying this approach are sketched by Granlund in his abstract (see Section B.8).

H. Buxton emphasized that in particular the necessity to 'learn' in a rather broad sense links visual information processing across the divide between machines and living creatures.

J. Tsotsos illustrated by a two-pronged presentation the difficulties associated with the desire to be simultaneously precise and relevant to other groups

with a stake in Cognitive Vision. He first went back for about a quarter of a century to his early work on the algorithmic transformation of X-ray image sequences of ventricular motion into medically relevant conceptual characterisations. Subsequently, Tsotsos used the notion of 'attention' to illustrate the difficulties encountered by any attempt to start from first principles in order to derive predictions which can be confirmed or invalidated by current experimental methods. Obviously, attention becomes obligatory for any 'real' application which has to cope with *finite* resources.

3 On 'Engineered' Cognitive Vision Systems

A greater part of the presentations and discussions during this seminar was devoted to technical aspects of CVSs as realised by Computer (or Machine) Vision approaches. A major difference became discernible with respect to the question whether a (technical) CVS needs to be 'embodied' or not. Proponents of the first view require that any 'true CVS' needs to influence its environment directly by control of integrated actuators. Although this point of view appears 'natural' in case of exploratory system approaches or tasks which require mechanical manipulations of their environment, the possibility to admit purely 'observational' systems with cognitive attributes or properties should not be excluded. *H.-H. Nagel* presented an attempt to derive those aspects which should be specified either implicitly or – better – explicitly prior to any attempt to engineer a CVS. It was agreed during the discussion following this introductory presentation that resource limitations and the consequences, namely to incorporate kind of attention-related processes into any CVS specification, have to be studied, too, for engineered CVSs.

In order to facilitate the discussion of the large range of different topics raised during the seminar, a coarse CVS-layer-structure will be sketched first. A *signal handling* layer is assumed directly on top of a basic *sensory* layer. Whereas signals are evaluated in this second-lowest layer only within a *local* spatio-temporal environment in the signal domain, *non-local* aggregation processes are admitted in the next – so-called '*picture domain*' – layer. Processes in this layer are still restricted to handling representations in the 2D-image plane and temporal changes of such representations. Geometrical representations in the 3D-scene are handled in the next higher '*scene domain*' layer. The layer above this scene domain layer is assumed to comprise representations for elementary *conceptual* representations of bodies and phenomena in the recorded video stream. Further abstraction steps then comprise conceptual representations for aggregates of bodies and more complex relations between them. In addition, elementary representations for changes and trajectories of (parts of) bodies are aggregated into representations for movements and entire 'behaviours'.

Using these notions, contributions will be coarsely characterised in the sequel by an indication to which layers they refer. Roughly circumscribed system capabilities were used to cluster contributions into half-day sessions. Not sur-

3

prisingly, many presentations addressed more than one aspect and thus may be mentioned in several of the following subsections.

In the remainder, the notion '*class*' will be used if the members to be assigned to such a set do not differ from each other in any relevant aspect, apart from an identifier which allows to refer to a particular member ('*instance*') of this class. The associated 'classification' is equivalent to a *re*-cognition process in the conventional sense which is sometimes denoted also as 'Pictorial Pattern Recognition'. In contradistinction, the notion '*category*' will be reserved for those abstractions which subsume components with at least one essential difference between members of the category set – a difference, however, from which the categorisation process deliberately abstracts.

3.1 Approaches centered at the signal and picture domain layer

Unless a-priori knowledge is available about where in the image one has to search for what, the quest for efficiency and generality recommends to extract local, signal-level representations in a data-driven manner. As a consequence, the number of such representations ('features') is large in general with the necessity to study efficient methods for the aggregation of local signal descriptors into non-local ones which subsequently are to be exploited for the transition into the layer of elementary *conceptual* representations, i. e. for use by classification or categorisation processes. In this context, Franc & Hlavac (see abstract in Section B.7) study computationally tractable approximations of kernel methods.

Dickinson (see abstract in Section B.6) assumes that – more or less local – signal representations have been obtained already. Selected features then have to be aggregated into non-local representations at the picture domain level for classification or categorisation processes. This problem is formulated using a graph-matching approach with emphasis on many-to-many node matching.

An analogous problem is studied by *Kropatsch* and co-workers (see abstract in Section B.12) based on topological representations of planar subsets of the image plane, i. e. segmentation results. In this manner, planar shape and spatial relations in the 2D-image-plane are exploited to construct hierarchical representations which facilitate subsequent classification processes.

3.2 Linking the picture domain layer to the elementary conceptual representation layer

This subsection comprises various approaches which segment single images or image sequences in order to link selected segments to class or category names.

Such a process is studied by *Kittler* and *Ahmadyfard* (see Section B.11). Given the fact that the Probability Distributions Functions (PDFs) required for a Bayesian classification process have to be estimated in most cases from a finite 'learning sample', these PDFs may exhibit errors and thus can result in incorrect classifications. This is most likely to happen in *simultaneous* many-class-decision processes, in particular if the likelihood varies considerably between

various classes. Kittler and Ahmadyfard suggest to use a cascaded classifier approach, rejecting at each cascade step the least likely alternative(s).

Leonardis and co-workers (see Section B.15) study subspace methods which incrementally modify a tentative subspace basis obtained from a kind of Principal Component Analysis (PCA) in order to achieve a high breakdown point for subsequent classification.

The contribution by *D. Hall* (see Section B.10) may be considered as another example for a classification process. The signal level representation consists of first and second order derivatives at each pixel position, estimated by a scale-invariant subprocess. These signal descriptors are subsequently aggregated by several heuristics in order to obtain a picture domain representation which then is associated directly with the class name, i. e. an elementary conceptual description.

It is illuminating to compare this approach with the ones reported by Mohr and Leibe & Schiele, in particular with respect to the different representations used for local signal characteristics. *Mohr* (see Section B.17) first searches for 'interest points'. A next step attempts to pick up significant signal variations around the selected interest points, in particular texture-like characteristics. The resulting non-local representations at the picture-domain layer are used by a classifier trained in a preceding learning phase.

A good example for a categorisation approach is provided by the different variants of a 'car' which have to be detected and localized by the approach reported by *Leibe* and *Schiele* (see Section B.14). These authors rely on codebook-vectors acquired during a supervised 'learning phase'. The codebook can be considered as a 'signal level representation' of the object image to be searched for. Some processing steps convert this representation into one at the picture domain level and link this latter one directly to a particular concept (the category name) at the level of elementary concepts. No intermediate scene-level representation is derived by this approach.

V. Krüger (see Section B.13) reported the recognition and tracking of people in video sequences, based on probabilistic modelling. To capture variation over time, a Monte-Carlo approach is used to extract pose variations. The 'signal level' representation of variations is adequately captured by first order statistics for well-defined domains. The approach is well suited in particular for in-door environments where illumination variations are limited.

3.3 Approaches emphasizing investigations at the conceptual level

Given the principal capability to categorize individual non-local representations obtained at the picture-domain layer, the study of spatio-temporal relations between such entities becomes of interest: in other words, the context can be exploited within which the image of a body has to be found and recognized. Obviously, the number of considerations which have to be taken into account may grow rapidly with increasing realism of task and/or depicted scene. It thus appears advisable to *abstract* from details of particular instances, depending on the kind of context available. The transition from a geometrical representation –

5

be it at the picture-domain or the scene-domain level – to a conceptual representation offers two advantages. First, it realizes the abstraction step in a general manner, and second, it allows to exploit well established methods of formal logic in order to perform the required inference steps.

Cohn (see Section B.3) demonstrates these considerations for learning representations of road traffic behavior from image sequences. Similar considerations are touched by the presentation of *H. Buxton* (see Section B.2).

The potential of (variants of) formal logic in the context of CVSs has been nicely illustrated by *B. Neumann* (see Section B.20), amongst others by its use in an effort to formalise the interpretation of images from table-top settings.

B. Nebel (see Section B.19) provided an overview of topics from the area of formal logic, including problems of complexity analysis encountered during the search for an efficient implementation of inference processes, in particular in the context of machine-vision-based control of robots for (a simplified version of) 'soccer'-playing. This 'application' illustrates the transition from elementary to more complex conceptual representations, namely from categories to movements and behaviors. This latter aspect showed up prominently, too, in the research reported by *M. Thonnat* where she addressed, amongst other questions, the detection of illegal behavior in videos recorded by surveillance cameras in public spaces, for example at subway entrances (see Section B.22).

3.4 Studying interactions between a CVS and its environment

This topical area will be treated against the background of considerations discussed in particular by G. Granlund (see Section B.8) and the debate concerning 'embodiment' of CVSs – see the beginning of Section 3. A variety of related problems have been mentioned already in connection with the presentations of B. Nebel and were amply illustrated by *J. Little* (see Section B.16).

H. Niemann presented an approach which uses computer control of a robot carrying a video camera in order to obtain single video images for the detection and classification of any one from a set of six different known objects expected on an office table. In this example, binocular color image pairs had to be evaluated in order to determine the office table in space and to search for objects whose representation had been learned during a preceding learning phase. Control of the recording camera head comprises pan, tilt, zoom, and translation along a linear sledge. A semantic network comprised the context knowledge which had to be evaluated during the search and classification processes in order to record optimally suited images. These operations had been (partially) learned based on reinforcement learning (see Section B.21).

Research pursued at Bielefeld (see Section B.1), presented by *Ch. Bauckhage*, also implied interaction with the system environment, but with a different thrust. In this example, the CVS is expected to act as a kind of 'memory prosthesis', i. e. to support a human while the latter attempts to manipulate objects in the joint field-of-view of the human agent and the CVS. The CVS has to communicate with the human regarding what it recognizes, depending on questions posed by the human agent and potentially additional hints characterizing the object to

be looked for. Obviously, psychological problems regarding what a human may expect in such a situation and how he might react to the system's responses come into play, too, in this task context. This approach thus provides an example for straddling the boundary between a restricted and a broad understanding of CVSs.

Research pursued in the ActiPret project was presented by *M. Vincze* (see Section B.25). The project uses visual tracking and basic recognition to generate textual descriptions of actions such as loading a CD player. For a consortium like ActiPret, the integration of competences from a number of different institutions constitutes a crucial problem. Consequently, a critical aspect is the availability of software engineering methods for systems integration. A distributed – directory based – framework has been investigated that allows simple engineering of vision systems. The system has been tested in early experiments on monitoring human activities.

In research pursued at KTH (see Section B.4) vision is integrated into a mobile agent for manipulation of objects. Manipulation involved recognition of objects, servoing to bring the end-effector into proximity of the object, and finally integration with haptic sensing to allow pickup of the object. The CVS here integrates recognition to detect the object, servoing with a strong prior (the recognized model), and integration with other sensory modalities. Recognition is achieved using statistical learning. Servoing is based on a set of three different strategies (position, image, and hybrid) to facilitate efficient control.

J. Crowley (see Section B.5) discussed the integration of visual systems using a state-space approach, in which a supervisor controls an ensemble of visual processes. The methodology was illustrated in the context of intelligent rooms which mediate meetings. The system automatically detects people entering the room, the context of interaction between people participating in the meeting, and augments the room with visual aids. The various phases of meetings are encoded as a discrete event model that drives the vision system. Simple detection is used for handling images of people. A number of static gestures are used to interact with the system. Through utilisation of active cameras and projectors it becomes possible to detect and track meeting activities. At the same time, the methodology illustrates use of relatively simple structures for systems integration.

4 Discussion

A considerable part of an evening discussion was taken up by a 'post mortem analysis' of related earlier European Projects in which participants had been involved during the late eighties and early nineties as consortium members, coordinators, or reviewers. Although the goals formulated at that time were quite similar to current ones, unexpectedly large difficulties arose in the past with processes corresponding to the signal and picture domain layer sketched above. From hindsight, this did not appear as a surprise: today, about three orders of magnitude more computing power is available to university laboratories at

7

roughly the same cost. Corresponding processes today are more robust and allow faster as well as easier experimentation, but still represent a major hurdle to the realisation of widely applicable and reliable vision systems. Over a decade ago, efforts to study the conceptual level by construction of experimental systems thus were confronted with considerably greater obstacles than today.

5 Conclusions

Current experimental systems begin to process video signals reliably and fast enough at the signal and picture domain layer to facilitate more advanced experiments. This becomes discernible by the extended range of investigations concerned with classification and even categorisation tasks. The transition from the picture domain layer to elementary conceptual representations is performed routinely enough to allow an exploration of behavioral representations for developments in a scene recorded by one or more video cameras.

It thus appears possible to separate an 'engineering' type of CVS from a much broader, equally justifiable view on 'Cognitive Vision'. A metaphor possibly best captures the consensus which gradually emerged during this seminar: this separation should not be perceived in the form of a precise definition, but rather as a kind of 'fuzzy, semi-permeable wall' which allows selected ideas to migrate both inwards and outwards, possibly with some modifications depending on which direction an idea passes through this 'boundary region'. Such a view should allow to define an engineering approach towards a CVS without excluding, on the one hand, possibly stimulating ideas in a dogmatic manner or, on the other hand, 'washing out' the notion of a CVS to an extent which incapacitates the specification of a realisable systems approach. The latter clearly emerges as a challenge for the near future.

Acknowledgment

Our thanks go to the staff of Schloss Dagstuhl for their very smooth support of this seminar, both prior to and during the meeting. Numerous special wishes were handled very efficiently, making the stay during this week a memorable experience for all participants.

The coordinators gratefully acknowledge thoughtful comments by David Vernon on a draft version of this report as well as the quiet support by Ch. Köhler/ Freiburg and A. Ottlik/ Karlsruhe for taking notes during discussions, collecting source files of abstracts, and making these available to us after the meeting.

We want to acknowledge, too, that this seminar has been partially supported by the European Community under contracts No. IST-2000-29375 (CogVis), IST-2001-35454 (ECVision), and IST-2000-29404 (CogViSys).

8

A Final Program

A.1 Week schedule

	Monday	Tuesday	Wednesday	Thursday	Friday
AM	Cognitive Vision Systems	Recognition / Categorisation 1	Reasoning & Interpretation	Control	Interaction
PM	Learning	Recognition / Categorisation 2	Hike	Cognitive Vision	END
Evening				Panel	

A.2 Monday 27 October

Morning Session: 09.00 – 12.00
Cognitive Vision

- Cognitive Vision Systems: From Ideas to Specifications – *Nagel*
- The Structure of Cognitive Vision Systems – *Granlund*
- A Vision on Cognitive Vision – *Vernon*

Afternoon Session: 13.30 – 18.00
Learning

- Visual Learning, the Next Decade Challenge - *Mohr*
- Learning Visual Representations for Cognitive Vision Systems - *Buxton*
- Visual Learning and Recognition Using Subspace Methods - *Leonardis*
- Learning Qualitative Behaviour Descriptions from Visual Input - *Cohn*

A.3 Tuesday 28 October

Morning Session: 09.00 – 12.00
Recognition & Categorisation 1

- A Computationally Tractable Approximation of Kernel Methods for Learning – *Hlavac*
- A Framework for Object Class Detection – *Hall*
- Model Pruning in Object Recognition: A Theoretical Basis – *Kittler*

Afternoon Session: 13.30 – 18.00
Recognition & Categorisation 2

- Many-to-Many Feature Matching in Object Recognition – *Dickinson*
- Spatio-Temporal Configurations: Description & Retrieval – *Nebel*
- Appearance-based Object Categorisation – *Schiele*
- Interleaved Object Categorisation and Segmentation – *Leibe*

9

A.4 Wednesday 29 October

Morning Session: 09.00 – 12.00
Reasoning and Interpretation

- Knowledge-based Exploration of Scenes – *Niemann*
- Merging Relations, Probabilistic and Logic-based Representations for Dynamic Scene Interpretation – *Neumann*
- Hierarchies Relating Topology and Geometry - *Kropatsch*

Afternoon Session: Hike around the Wadern area

A.5 Thursday 30 October

Morning Session: 09.00 – 12.00
Control

- Visual Attention in a Cognitive Vision System - *Tsotsos*
- Integrating Video Information over Time - *Krüger*
- On Vision-based Communication - *Bauckhage*

Afternoon Session: 13.30 – 18.00
Cognitive Vision

- A Framework for Cognitive Vision or Identifying Obstacles to Integration - *Vincze*
- Towards Cognitive Vision: Knowledge and Reasoning for Image Analysis and Interpretation – *Thonnat*
- Presence of Vision – *Granum*

Evening Session: 20.00 –
Open Panel on CVS

A.6 Friday 31 October

Morning Session: 09.00 – 12.00
Interaction

- Collaboration with an Autonomous Agent - *Little*
- Integration of Cognitive Vision Systems - *Christensen*
- A Roadmap for Cognitive Vision – *Crowley*
- Summary / Wrap-Up *Christensen & Nagel*

THE END

10

B Abstracts of Contributions Presented at the Dagstuhl Seminar on 'Cognitive Vision' (27.10.03 – 31.10.03)

The abstracts are *ordered alphabetically* according to the (first) author's family name.

B.1 On Vision-based Communication

Christian Bauckhage, Universität Bielefeld

The cognitive vision project VAMPIRE investigates architectures and computational models for "Visual Active Memories". This term defines systems that evaluate, gather, and integrate contextual knowledge for visual analysis. Moreover, visual active memories can understand new scenes and situations, learn new spatio-temporal relations as well as new concepts and categories and they are scalable to new domains. The second important aspect considered in VAMPIRE are augmented reality techniques for "Interactive REtrieval". These should process and accomplish user queries and correspondingly revisualize past events and recognized objects. In short, the aim of VAMPIRE is to proceed towards memory prosthetic devices.

This talk will present first results achieved on the way to this goal. After techniques for object and action recognition in an office environment and approaches to system integration will have been discussed, the issue of how to evaluate complex cognitive systems will be raised. Experiences from earlier projects on advanced human-machine interaction will be reported and the promising potential of psychologically based usability experiments will be stressed.

B.2 Learning Visual Representations for Cognitive Vision Systems

Hilary Buxton COGS & Dept Informatics, University of Sussex, UK

I assume a framework for cognitive vision that is clearly purposive. Within this, we need to learn what kinds of things there are in the environment and more importantly, how to see them in the ongoing perception action cycle in order to reach our goals. First, then, I review some work on learning "what" models for objects, events, actions and behaviours and second, learning "how" to see, for which I propose a simple framework for task- based learning and perception that can be used in the design and development of Cognitive Vision Systems (CVS).

In particular, I use an approach that treats perception and learning as inference from data using generative, probabilistic models. The models form a family of increasing complexity from simple Principal Component Analysis (PCA) and Gaussian Mixtures (GM) that can support object recognition, to more complex temporal and decision theoretic models such as Dynamic Bayesian Nets (DBN) and Dynamic Decision Nets (DDN) that can support task control. There are both offline and online, incremental versions of the associated learning algorithms such as Expectation Maximisation (EM). Examples are given for using

these models in past projects e.g. VIEWS, current projects e.g. ActIPret, as well as speculations about future developments.

I present a basic ontology for CVS: 1) "what" knowledge -visual learning and memory, which involves the representation of objects and their behaviour, including recognition, categorisation, scene-context and expert ontology; 2) "how" knowledge -visual control and attention, which involves the visual skills of perception for tasks using expectation in the on-line system, including goals, task-context, priorities, resources and embodiment; and 3) cognition itself, which includes goal-based invocation of the perception-action cycle, learning of shared symbols for communication and self-reflection, development, social and affective issues, as well as creativity and consciousness.

B.3 Learning qualitative behaviour descriptions from visual input

Anthony G. Cohn, University of Leeds

In this talk I describe some work conducted at Leeds jointly between the Computer Vision and Knowledge Representation and Reasoning Groups, particularly in the context of the EU FP5 Project, CogVis, IST-2000-29375. A key focus of our work is to integrate quantitative and qualitative modes of representation; in particular we aim to exploit quantitative visual processing for tracking and motion analysis and exploit qualitative spatio-temporal representations to abstract away from unnecessary details, error and uncertainty. We exploit common-sense knowledge of the world as constraints on interpretations. Following Shanahan, interpretation is viewed as abducing an explanation of sensor data with respect to general and domain specific models. Crucial to our approach is our aim to learn as much as possible of these domain specific models.

The work I describe includes a system which learned traffic behaviours using qualitative spatial relationships among close objects travelling along 'typical paths' (which were also induced). In later work we induced the set of qualitative spatial relationships rather than taking these as given a priori. In another piece of work we have shown how it is possible to reason symbolically, using common-sense knowledge of continuity of physical objects, in order to refine ambiguous classifications from a statistical classifier. Finally, I describe ongoing work where we are attempting to learn symbolic descriptions of intentional behaviours such as those found in simple table top games involving dice or cards, using Inductive Logic Programming.

B.4 Integration of Cognitive Vision Systems

Henrik I. Christensen, NADA KTH Stockholm
hic@nada.kth.se

The study of cognitive vision systems involves the processes required to generate semantic models of the external environment so as to allow an artificial agent to operate and interact with the environment and other agents in that

environment. To facilitate this, the system must be endowed with competences for recognition, interpretation, open-ended acquisition of models at all levels, and integration of all processes into an operational unit that is embodied in a regular physical structure.

In this paper the general structure is discussed and the example of a mobile manipulation system is used to illustrate a number of current issues involved in the construction of such systems. The emphasis is here on component processes, and the integration of these into a system for fetch-and-carry, basic user interaction, and simple vision based manipulation.

B.5 A Research Roadmap for Cognitive Vision

James L. Crowley, INRIA Rhône–Alpes
crowley@imag.fr

In this talk, I described a distributed software model for context-aware perception of human activity. The basic building blocks in this model are perceptual modules, composed of a data transformation component and a control component. Modules are assembled into perceptual processes controlled by a reflexive process controller. Process controllers regulate computation, and provide a reflexive description of their internal state and capabilities. Explicit models of context are used to assemble federations of processes for observing and predicting activity. As context changes, the federation is restructured. Restructuring the federation enables the system to adapt to a range of environmental conditions and to provide services that are appropriate over a range of activities.

B.6 Many-to-Many Feature Matching in Object Recognition

Sven Dickinson, University of Toronto

One of the bottlenecks of current recognition (and graph matching) systems is their assumption of one-to-one feature (node) correspondence. This assumption breaks down in the generic object recognition task where, for example, a collection of features at one scale (in one image) may correspond to a single feature at a coarser scale (in the second image). Generic object recognition therefore requires the ability to match features many-to-many. In this talk, I will review our progress on three independent object recognition problems, each formulated as a graph matching problem and each solving the many-to-many matching problem in a different way. In the first problem, we define a low-dimensional, spectral encoding of graph structure and use it to match entire subgraphs whose size can be different. Next, we explore the problem of learning a 2-D shape class prototype (represented as a graph) from a set of object exemplars (also represented as graphs) belonging to the class, in which there may be no one-to-one correspondence among extracted features. Finally, in very recent work, we embed graphs into geometric spaces, reducing the many-to-many graph matching problem to a weighted point matching problem, for which efficient many-to-many matching algorithms exist.

B.7 A computationally tractable approximation of kernel methods

Vojtech Franc and Vaclav Hlavac, Czech Technical University Prague

We propose a technique for a training set approximation and its usage in kernel methods. The approach aims to represent data in a low dimensional space with possibly minimal representation error which is similar to the Principal Component Analysis (PCA). In contrast to the PCA, the basis vectors of the low dimensional space used for data representation are properly selected vectors from the training set and not as their linear combinations. The greedy algorithm has a low computational complexity and is capable to process huge data sets online. The proposed method was tested on training Support Vector Machines, Kernel Fisher Linear Discriminant and data approximation problem. The experiments show that the proposed approximation reduces significantly the complexity of the found classifiers (the number of the support vectors) while retaining their accuracy.

B.8 The Structure of Cognitive Vision Systems

Gösta Granlund, Linkping University

The purpose of cognitive systems is to produce a response to appropriate percepts. The response may be a direct physical *action* which may change the *state* of the system. It may be delayed in the form of a reconfiguration of internal models in relation to the interpreted *context* of the system. Or it may be to generate in a subsequent step a generalized *symbolic representation*, which will allow its intentions of actions to be communicated. As important as the percepts, is the dependence upon context.

A fundamental property of Cognitive Vision systems is the *extendability*. This requires that systems both acquire and store information about the environment autonomously – on their own terms. The distributed organisation foreseen for processing and for memory to allow learning, implies that later acquired information has to be stored in relation to earlier. The semantic character of the information implies a storage with respect to similarity, and the availability of a metric.

This has the consequence that the possibility for a designer to "push in" information into an operating system, is less than previously assumed. Such information has to go through the interpreting and organizing mechanisms of the system, as discussed above.

In the conventional structure of a vision system, a description usually in geometric terms, is first built up step by step. This description is then carried to a second unit where it is interpreted to generate appropriate actions. The problem with a description of an object or a scene is that it lacks an *interpretation*, i.e. links between actions that are related to the object and changes of percepts. The purpose of a Cognitive Vision system is to build up a model structure which relates the percepts emerging from an object to its states, or actions performed upon it.

In contrast, a system structure is proposed, where the first part of the system, step by step performs a mapping from percepts onto actions or states. The central mechanism is the *perception–action* feedback cycle, where in the learning phase, *action precedes perception*. The reason for this causal direction is that action space is much less complex than percept space. It is easy to distinguish which percepts change as a function of an action or a state change. Percepts shall be mapped directly onto states or responses or functions involving these. Symbolic representations are derived mainly from system states and action states.

Through active exploration of the environment, e.g. using perception-action learning, a system builds up *concept spaces*, defining the phenomena it can deal with. Information can subsequently be acquired by the system within these concept spaces *without interaction*, by extrapolation using passive observation or communication such as language.

A system can consequently acquire information with three different mechanisms, where each one is dependent upon the acquisition in the previous mechanisms:

- Copying at the time of system generation
- Active exploration, defining concept spaces through associative learning
- Communication, i.e. using language

This structure has been implemented for the learning of object properties and view parameters in a fairly unrestricted setting, to be used for subsequent recognition purposes. The presentation will give further details and properties of the proposed structure for Cognitive Vision systems.

B.9 Presence of Vision

Erik Granum, Computer Vision and Media Technology Laboratory,
Aalborg University
eg@vision.auc.dk

Presence, as the sense of "being there", is the theme of a current European Research initiative, where 10 projects are funded to develop theory and novel media for investigation of the concept of presence. Descriptions of the goals of this initiative were given as an example of a larger interdisciplinary research context in which the visual faculty plays an important role.

As a reference for a discussion of the content and progress in cognitive vision - a vision project of the early nineties - was briefly revisited (VAP, Vision as Process, 1989 - 1995). "Vision in context" was considered also at that time as relating vision to the environment in which is was operating as well as relating vision to other faculties and cognitive facilities within the operating (vision) system. In an attempt to get closer to a definition of "cognitive vision", Websters dictionary had been consulted. This revealed "cognitive" as what could be based on, and capable of being reduced to, empirical factual knowledge. Correspondingly, "cognition" was explained as "the act or process of knowing, including awareness and judgment".

The point was made, that some contributions to define the field – also given at this seminar – suggested a much more comprehensive content of a cognitive vision system, than the term cognitive apparently could account for. Hence the claimed characteristics like: embodiment, external action control, and action dependent vision, may be very relevant for a cognitive system, which is supported by the visual faculty, but not necessarily useful for description of a vision system itself. Rather, the above concepts emerge when vision is active within a cognitive system and in internal interaction with other human like functionalities, such as consciousness.

In conclusion it was suggested that the vision community joined up with other disciplines of relevance for the pursuit of the role of vision in a cognitive system. Doing it in our own were not likely to reveal the insight we are aiming at. We must be conscious and explicit about how vision is present in its context, and that is externally as well as internally.

B.10 A Framework for Object Class Detection

Daniela Hall, INRIA Rhône-Alpes

A successful detection and classification system must be able to compensate for intra-class variability and specific enough to reject false positives. We describe a method to learn class-specific features that are robust to intra class variability. These feature detectors enable a representation that can be used for a verification process. Instances of object classes are detected by a module that verifies the spatial relations of the detected features. We extend the verification algorithm in order to make it invariant to changes in scale. Because the method employs scale invariant feature detectors, objects can be detected and classified independently of scale. Our method has low computational complexity and can easily be trained for robust detection of different object classes.

B.11 Model pruning in object recognition: A theoretical basis

J. Kittler and A. Ahmadyfard, University of Surrey

We propose a multiple classifier system approach to object recognition in computer vision. The aim of the approach is to use multiple experts successively to prune the list of candidate hypotheses that have to be considered for object interpretation. The experts are organised in a serial architecture, with the later stages of the system dealing with a monotonically decreasing number of models. We develop a theoretical model which underpins this approach to object recognition and show how it relates to various heuristic design strategies advocated in the literature. The merits of the advocated approach are then demonstrated experimentally using the SOIL database. We show how the overall performance of a two stage object recognition system, designed using the proposed methodology, improves. The improvement is achieved in spite of using a weak recogniser for the first (pruning) stage. The effects of different pruning strategies are demonstrated.

B.12 Hierarchies relating Topology and Geometry

Walter G. Kropatsch, Yll Haxhimusa, Pascal Lienhardt, TU Wien

Cognitive Vision has to represent, reason and learn about objects in its environment it has to manipulate and react to. There are deformable objects like humans which cannot be described in simple geometric terms. In many cases they are composed of several pieces forming a 'structured subset of Rn or Zn'. We introduce the potential topological representations for structured objects: plane graphs, combinatorial and generalized maps. They capture abstract spatial relations derived from geometry and enable reconstructions through attributing the relations by e.g. coordinates. In addition they offer the possibility to combine both topology and geometry in a hierarchical framework: irregular pyramids. The basic operations to construct these hierarchies are edge contraction and edge removal. We show preliminary results in using them to hold a whole set of segmentations of an image that enable reasoning and planning actions at various levels of detail down to a single pixel in a homogeneous way. We further speculate that the higher levels map the inherent structure of objects and can be used to integrate (and 'learn') the specific object properties over time by up-projecting individual measurements. The construction of the hierarchies follows the philosophy to reduce the data amount at each higher level of the hierarchy by a factor ¿ 0 while preserving important properties like connectivity and inclusion.

B.13 Integrating Video Information Overtime

Volker Krüger, Aalborg University Esbjerg

(Joint work with Shaohua Zhou and Rama Chellappa at University of Maryland.)

vok@cs.aue.auc.dk

In this talk, I introduce a probabilistic approach for integrating video information over time and I use the face recognition application as an example application. In *still-to-video* recognition, the gallery consists of still images and the probe set consists of video sequences, a time series state space model is proposed to fuse temporal information in a probe video, which simultaneously characterizes the kinematics and identity using a *motion vector* and an *identity variable*, respectively. The joint posterior distribution of the motion vector and the identity variable is estimated at each time instant and then propagated to the next time instant. *Marginalization* over the motion vector yields a robust estimate of the posterior distribution of the identity variable. A computationally efficient *sequential importance sampling* (SIS) algorithm is used to estimate the posterior distribution. Empirical results demonstrate that, due to the propagation of the identity variable over time, a *degeneracy* in posterior probability of the identity variable is achieved to give improved recognition.

The gallery is generalized to videos in order to realize *video-to-video* recognition. An *exemplar-based learning* strategy is adopted to automatically select

video representatives from the gallery, serving as mixture centers in an updated likelihood measure. The SIS algorithm is applied to approximate the posterior distribution of the motion vector, the identity variable, and the exemplar index, whose marginal distribution of the identity variable produces the recognition result. The model formulation is very general and it allows a variety of image representations and transformations.

Experimental results using images/videos collected at UMD, NIST/USF and CMU with pose/illumination variations illustrate the effectiveness of this approach for both still-to-video and video-to-video scenarios with appropriate model choices.

B.14 Interleaving Visual Object Categorization and Segmentation

Bastian Leibe, Bernt Schiele, ETH Zürich
schiele@inf.ethz.ch

We present a method for object categorization in real-world scenes. Following a common consensus in the field, we do not assume that a figure-ground segmentation is available prior to recognition. However, in contrast to most standard approaches for object class recognition, our approach effectively segments the object as a result of the categorization. This combination of recognition and segmentation into one process is made possible by our use of an Implicit Shape Model, which integrates both into a common probabilistic framework. In addition to the recognition and segmentation result, it also generates a per-pixel confidence measure specifying the area that supports a hypothesis and how much it can be trusted. We use this confidence to derive a natural extension of the approach to handle multiple objects in a scene and resolve ambiguities between overlapping hypotheses with an MDL-based criterion. In addition, we present an extensive evaluation of our method on a standard dataset for car detection and compare its performance to existing methods from the literature. Our results show a significant improvement over previously published methods. Finally, we present results for articulated objects, which show that the proposed method can categorize and segment unfamiliar objects in different articulations and with widely varying texture patterns. Moreover, it can cope with significant partial occlusion.

B.15 Visual Learning and Recognition Using Subspace Methods

Ales Leonardis, University of Ljubljana

Visual learning is expected to be a continuous and robust process, which treats input data selectively. In the talk I present a method for subspace learning, which takes these considerations into account. I start with an incremental method, which sequentially updates the principal subspace considering weighted influence of individual images as well as individual pixels within images. This approach is further extended to enable determination of consistencies in the input data

and imputation of the values in inconsistent pixels using the previously acquired knowledge, resulting in a novel incremental, weighted, and robust method for subspace learning. I also present a new method which enables a robust calculation of the LDA classification rule, thus making the recognition of objects under non-ideal conditions possible, i.e., in situations when objects are occluded or they appear on a varying background, or when their images are corrupted by outliers. The main idea behind the method is to translate the task of calculating the LDA classification rule into the problem of determining the coefficients of an augmented reconstructive model (PCA). Specifically, we construct an augmented PCA basis which, on the one hand, contains information necessary for the classification (in the LDA sense), and, on the other hand, enables us to calculate the necessary coefficients by means of a sub-sampling approach resulting in a high breakdown point classification. The theoretical results are evaluated on the ORL face database showing that the proposed method significantly outperforms the standard LDA.

(This is a joint work with Danijel Skocaj and Sanja Fidler.)

B.16 Collaboration with an autonomous agent

James J. Little, University of British Columbia

We build visually guided mobile robots that collaborate with us in dynamic unstructured environments. By engineering these embedded agents we determine what visual information is relevant for a variety of tasks. When we build systems that interact with the world, with humans, and with other agents, we rely upon all of the aspects of cognitive vision, including: knowledge representation descriptions of the scene and its constituent objects models of agents and their intentions learning adaptation to the world and other agents reasoning about events and about structures interpretation of other agents' and users' interactions recognition and categorization

I will review the Robot Partners project at UBC which focuses on the design and implementation of visually guided collaborative agents, specifically interacting autonomous mobile robots. I will describe some of the capabilities of the robots that permit collaboration in variety of roles, and then will describe those roles and how they use the capabilities:

- Localization
- Navigation - Avoiding dynamic obstacles
- Local Spatial Context
- Face Recognition
- Facial Display
- Gestures and Expressions
- Interaction
- Associating Names with Visual Structures

The roles include Jose, the robot waiter, and Homer, the Human Oriented Messenger Robot. I will conclude with remarks on what are the next steps toward enabling the robots with cognitive vision.

B.17 Visual Learning, the next decade challenge

Roger Mohr, GRAVIR–INPG, Grenoble

Recent advances in computer vision allow to understand the major aspects of the image field. However when we consider the generic problem of recognition of the content of an image, it addresses many different problems like variability of appearance, variability of an object class, that no predefined model can capture it, and therefore only learning is the answer for collecting and structuring this highly complex set of possible data.

This presentation first list a set of hurdles to overcome for such a purpose and to reduce at bit the complexity of the task. For instance, invariant descriptors, if they are local, will allow robustness to occlusion and to some limited changes in geometry, in illumination. The variability of aspects due to intraclass changes can often be coped with mixture of probabilistic distributions of the features.

With this basis, the learning still stays difficult due to the curse of dimensionality of the learning space, together with the limited number of examples to learn from, due the high frequency of outliers in our noisy data. The positive point is that we usually have a large number of irrelevant data that allows to collect a large set of counter examples for our learning mechanism.

The presentation illustrates how combination of partial solutions from the computer vision world together with statistical robust approach solves some key problems for this learning case: sorting out features of interest in a unsupervised way, handling occlusion for recognition from the learned features ...

B.18 Cognitive Vision Systems: From Ideas to Specifications

H.–H. Nagel, Universität Karlsruhe (TH)

A Computer Vision System (CVS) is expected to map an image (sequence) into an appropriate description of the world section in the field of view of the recording camera(s). The resulting description is either to be communicated to animate or inanimate recipients, alternatively it may be exploited to influence the depicted world section directly via actuators coupled to the CVS. Human expectations with respect to what constitutes an appropriate description expand concurrently with the processing and storage capacities of computers. As a consequence, the algorithmic processes incorporated into a CVS become ever more complex and the diversity of CVSs expands continuously. This contribution addresses engineering aspects of CVSs and deliberately excludes the discussion of the question to which extent the methods underlying the design of a CVS may provide models for human vision. A structure is suggested for the processing steps devoted to the extraction of spatiotemporally local signal characterisations, their aggregation into non-local image descriptions, and the construction of spatiotemporal geometric descriptions in the scene domain. It is claimed that the association of these signal and geometric descriptions with conceptual descriptions makes the design and implementation of more versatile and encompassing

Cognitive Vision Systems feasible. In view of the complexity encountered in the course of such an endeavor, a set of aspects is proposed whose consideration should help to clarify the design options and to enforce precise specifications as a precondition for a viable design of a Cognitive Vision System.

B.19 Spatio-temporal configurations: Description and retrieval

Bernhard Nebel, Universität Freiburg
nebel@informatik.uni-freiburg.de

High-level descriptions of temporal as well as spatial configurations are important in natural language contexts and when one wants to describe general high-level patterns. For instance, in automatic mail sorting and in robotic soccer, it is important to describe possible configurations of spatial objects in qualitative terms. Furthermore, we often want to describe how such configurations change over time. In this talk, we specify a query language for spatio-temporal configurations, which are basically conjunctive queries using qualitative spatial and temporal relations. Interestingly, the complexity of the satisfiability problem of constraint-based reasoning is not relevant here. Nevertheless, the computational problem of evaluating conjunctive queries is – as is well known – NP-hard. However, if queries are only of a moderate size, this result does not affect the practical efficiency. After describing how to implement query processing, we give a demonstration of a system that is able to recognize and mark spatio-temporal configurations, which as one application has the task of identifying implausible scene interpretations.

B.20 Merging Relational, Probabilistic and Logic-based Representations for Dynamic Scene Interpretation

Bernd Neumann, Universität Hamburg

In this talk I examine the usefulness of different paradigms for model-based high-level scene interpretation. Frame-based representations of aggregates are presented as a convenient means to specify conceptual units embedded in taxonomies and partonomies, and comprising binary relations. It is shown that expressive description logic systems such as RACER can also define such concepts and in addition offer several inference services such as consistency checking and model construction. However, these services are not immediately useful for the scene interpretation process. In particular, an uncertainty management has to be introduced. It has been shown (Rimey, Koller et al., Mahoney and Laskey) that Bayes nets can be integrated with frame-based representations. By judiciously defining the nodes in high-level vision taxonomies and partonomies to correspond to statistically local phenomena, Bayesian inference can be used to control the interpretation process.

B.21 Knowledge-based Exploration of Scenes

H. Niemann, U. Ahlrichs, D. Paulus, Universität Erlangen

A system is presented which searches with an active camera for known objects, constrained to lie on a table, in an otherwise unknown office using color images. Both camera actions and image processing methods are represented as concepts of a semantic network. Image processing methods comprise depth computation to find a table, generation of object hypotheses in an overview image, and object verification in a close-up view. Camera actions are pan, tilt, zoom, and motion on a linear sledge. System actions, either image processing or camera actions, are initialized by a graph search based control algorithm which tries to compute the best scoring instance of a goal concept. The sequence of actions for computing an instance are determined by precedences which are either adjusted manually or computed from reinforcement-learning. Results are presented comparing the two approaches.

B.22 Towards Cognitive Vision: Knowledge and Reasoning for Image Analysis and Interpretation

M. Thonnat, INRIA–Sophia Antipolis

In this talk (see reference below), I will present a synthetic description of work we have done during the last decades on object categorization, program supervision and more recently on video understanding. First in the object categorization part, I will describe how to recognize the class of complex natural objects; Knowledge of the hierarchy of the object classes are represented explicitly in knowledge bases as well as inferences to interpret the results of image processing algorithms. The reasoning is a classification or categorization one. Second, in the program supervision part, I will describe how to use a library of vision programmes; knowledge of the library image analysis programmes and of their use are represented explicitly in knowledge bases. The reasoning is a planning and execution control one. Third, in the video understanding part, I will describe how to perform semantical interpretation of video sequences; two kinds of knowledge are made explicit, the 3D model of the environment to observe and the behaviours to recognize. The reasoning is an interpretation one. Finally I will conclude on future work, in particular on a cognitive vision platform.

M. Thonnat:
Towards Cognitive Vision: Knowledge and Reasoning for Image Analysis and Interpretation.
HDR thesis, French University Nice Sophia Antipolis, October 2003.

22

B.23 Visual attention in a cognitive vision system

John K. Tsotsos, York University, Toronto

The Selective Tuning Model is a proposal for modelling visual attention in primates and humans. Its strength comes from both: biological predictive power as well as computational utility. This presentation overviews the model, presents an implementation of attending to motion patterns and shows experiments in humans (psychophysics and brain imaging) that demonstrate its biological realism. The motion hierarchy successfully attends to motion patterns in image sequences and results in localization and labelling of those patterns. It is claimed that a cognitive vision system can greatly benefit from attentive processes.

B.24 A Vision on Cognitive Vision

David Vernon, Etisalat University UAE

The term cognitive vision has recently been introduced into the discipline of computer vision to cover attempts to capture the general-purpose robustness and resilience of natural vision systems. However, the form of cognition that the vast majority of cognitive vision systems invoke is based on cognitivism, a particular model of cognition. In this paper, we explore the cognitivist model and various alternative paradigms of cognition, setting out the assumptions each makes and teasing out the consequences of these assumptions. Based on this analysis, we reach two significant conclusions: (a) the cognitivist paradigm, though quite mature and effective when deployed in well-bounded circumstances, is incapable of achieving fully-developed general-purpose cognitive capability; (b) the alternative emergent paradigms are much less mature but offer the only principled way of achieving cognitive visual capability. These conclusions have important implications. The first is that a cognitive vision system must necessarily be embodied, capable of exploring and manipulating its environment. Second, the agent's perception of its world is a dependent on the richness of its motoric capabilities. Third, communication abilities, either linguistic or gestural, are learned as a consequence of multi-agent interaction; one can't directly implant knowledge in or extract knowledge from the agent's internal state. Fourth, since the system must develop (qua learn) in real-time through synchronous interaction with its environment, the rate at which it can develop cognitive capability is intrinsically limited. This poses a major dilemma for these approaches, requiring advanced phylogenic configuration before any ontogenic development can bring it to the level of useful cognitive behaviour. A major research issue is how to accomplish this without falling into the trap of conventional cognitivism: system identification based on representations derived from external observers.

23

B.25 A Framework for Cognitive Vision or Identifying Obstacles to Integration

Markus Vincze, TU Wien
Vincze@acin.tuwien.ac.at

Cognitive Vision Systems (CVS) attempt to provide solutions for tasks such as exploring the environment, making robots act autonomously or understanding actions of people. What these systems have in common is the use of a large number of models and techniques, e.g., perception-action mapping, recognition and categorisation, prediction, reaction and symbolic interpretation, and communication to humans. For this contribution these ingredients of a CVS are referred to as skills and are encoded in components. To arrive at the level of building a system it is considered essential to provide a framework that coordinates the components. Two principles organise the components: the service principle uses "yellow pages" to announce its capabilities and to select other components, and the hierarchy principle orders components along data abstraction levels and ascertains that system response is reactive. ActIPret will show the interpretation of a person handling tools involving skills such as tracking, object and gesture recognition, spatial-temporal object relationships and reasoning to extract the symbolic description. To move towards multi-task CVSs we invite researchers to exchange components and the framework.

24

Participants

- Bauckhage, Christian (Fraunhofer IAIS – St. Augustin)
- Buxton, Hilary (University of Sussex – Brighton)
- Christensen, Henrik Iskov (Georgia Institute of Technology)
- Cohn, Anthony G. (University of Leeds)
- Crowley, James Lawrence (INRIA Rhône-Alpes)
- Dickinson, Sven (University of Toronto)
- Granlund, Gösta (Linköping University)
- Granum, Erik (Aalborg University)
- Hall, Daniela (INRIA Rhône-Alpes)
- Hlavac, Vaclav (Czech Technical University)
- Kittler, Josef (University of Surrey)
- Köhler, Christian (Universität Siegen)
- Kropatsch, Walter (TU Wien)
- Krüger, Volker (Copenhagen Inst. of Technology Aalborg University)
- Leibe, Bastian (RWTH Aachen)
- Leonardis, Ales (University of Ljubljana)
- Little, Jim (University of British Columbia – Vancouver)
- Mohr, Roger (INRIA Rhône-Alpes)
- Nagel, Hans-Hellmut (KIT – Karlsruhe Institute of Technology)
- Nebel, Bernhard (Universität Freiburg)
- Neumann, Bernd (Universität Hamburg)
- Niemann, Heinrich (Universität Erlangen-Nürnberg)
- Ottlik, Artur (Universität Karlsruhe)
- Schiele, Bernt (MPI für Informatik – Saarbrücken)
- Thonnat, Monique (INRIA – Sophia Antipolis)
- Tsotsos, John K. (York University – Toronto)
- Utcke, Sven (Universität Hamburg)
- Vernon, David (Etisalat University – Sharjah)
- Vincze, Markus (TU Wien)