Henrik I. Christensen, David Hogg, Bernd Neumann:

Knowledge Based Computer Vision

Dagstuhl-Seminar -Report 196
08.12.1997 - 12.12.1997 (9750)

Editor: Thomas Hartkens
hartkens@informatik.uni-hamburg.de

Contents

Lessons learnt from the last thirty years of research.

Summarised by Henrik I Christensen

Centre for Autonomous Systems,
Numerical Analysis and Computing Science
Kungliga Tekniska Hogskolan
S-100 44 Stockholm, Sweden
hic@nada.kth.se

An evening session at the Dagstuhl seminar was reserved for a discussion of lessons learnt by some of the senior members of the community. The following persons were asked to look back on thirty years of research: Prof. C. Brown, Prof. A. Mackworth, Prof. J. Kittler, Prof. E. Dickmanns, Prof. R. Bajcsy, Prof. T. Matsuyama, Prof. H.H. Nagel, and Prof. J. Tsotsos. The text below is a brief summary of the major points raised during a most rewarding discussion.

1. Technology has to a large extent driven the progress and it has gradually enabled use/formulation of more advanced theories for problems that earlier were considered out of reach.

2. Committment to science is critical to achieve success. This at the same time requires organisation as progress otherwise might end up being incremental without major breakthroughs. Without organisation there is at the same time a risk that the same methods are re-invented at a later stage. For the evaluation of progress and results it is also essential to be explicit about ones reference. Computer vision includes both biological and engineering goals and it is important to specify towards which øeld the research is directed at it determines how one should evaluate the results. It was emphasised that it is essential to stick to a goal/topic even if it implies switching between funding agencies.

3. It is essential that the task / system is considered from the start. This also implies that a multi-disciplinary approach must be used. Computer vision should exploit existing theories and techniques from other sciences. This includes øelds like physics, biology, computer science, artiøcial intelligence, statistics and control engineering. At the same time there is a conÆict with the disciplines as each of these disciplines in turn require credit for progress, both due to funding constraints and due to interdisciplinary collaborations. This requires a delicate balance between other disciplines and computer vision.

4. Time/dynamics has only recently been recognised as an important aspect of building operational systems. This is partly due to the fact that construction of fully operational system only recently has become possible. In addition methods for description of dynamics at several dicœerent levels from control theory to temporal logic have only recently been integrated into a coherent framework.

5. The combination of dicœerent disciplines has only happened recently which in part is due to the fact that there has been a kind of 'religious' separation between øelds like geometry, pattern recognition, control theory and artiøcial intelligence. I.e., simple applications, for example in pattern recognition, were not considered computer vision. In the view of complete systems it is, however, now apparent that such systems can only be built when the disciplines are combined with proper use of a multi-disciplinary approach.

6. Computer vision should have clear goals and they should at the same time be meaningful to science in general. This is in particular important to motivate computer vision as a discipline. Other sciences have deøned golden standards and use hypothesis testing etc. as a basis for their work. Such standards should also be imposed on scientiøc journals and conferences as it will allow broader recognition of computer vision as a well-established science. This at the same time implies that performance characterisation becomes a critical topic for evaluation of developed techniques.

7. The issue of adequate computer power was discussed. It is not immediately obvious if we have enough computing power to solve current problems. A more important problem might, however, be adequate knowledge. Most systems developed today use little or no explicit knowledge. Another related problem is that almost no systems have a well characterised knowledge base, which implies that the systems can not be combined with methods for learning and/or adaptation.

Knowledge-Based Computer Vision: The Issuses

Bernd Neumann

Fachbereich Informatik
Universitt Hamburg
Vogt-Klln-Strae 30
22527 Hamburg, Germany
neumann@informatik.uni-hamburg.de

The basis thesis of this introduction to the workshop is: Computer Vision systems know too little. One way to check for the available knowledge is to ask for best possible predictions of a scene, e.g. for the expected evolvement of a garbage-collection scenario. One issue of the workshop should be to deal with neglected bodies of knowledge, in the view of the author:

1. high-level knowledge about large pieces of spatial and temporal context
2. real-life visual appearances and phenomena
3. knowledge about good processing strategies.

Another issue is how to obtain knowledge - by design or experience. Various forms of learning should lead tp large (re)usable knowledge bases. A third issue is how o exploit knowledge, although the "how" should be a secondary issue after the "what". Knowledge-based vision systems should try to harvest the advantages of the ideal knowledge-based system architecture, where ontological, domain-speciøc and problem-speciøc knowledge is separated and corresponding inference procedures are identiøed. This could improve reusability and also help to establish formal properties of a vision system, in particular soundness, completeness and tractability.

Image Interpretation as Model Construction: Symbolic Reasoning meets Model-based Computer Vision

Carsten Schrder

debis Systeme, Hamburg

I presented a formal, logic-based approach to image understanding which combined methods from two diœerent research areas, namely knowledge- based representation and model-based vision. After describing two well-known approaches I presented a concive deønition of the required solution of an image understanding problem. I then proposed to use on object-centered, KL-ONE-like description logic tailored to the representation needs in image understanding, sketched a tableau-like calculus for, ørst, checking the consistency of a given knowledge base, and, second, computing the required solutions, i.e. the models of the knowledge base. In particular, I showed how to plugg in methods known in the øeld of model-based vision into the calculus for veriøying hypotheses generated on the logical level with numerical optimization techniques in the image. I argued that shape classes of objects and geometric constraints on the shape, position, and orientation parameters of objects represent the glue between symbolic reasoning on the logical level on the numerical reasoning on the image level.

Interpretation of Aerial Images for the Reconstruction of Urban Scenes

W. Frstner

Institut fr Photogrammetrie, Universitt Bonn
Nussallee 15, D-53115 Bonn, Germany
wf@ipb.uni-bonn.de

Making maps from aerial images, a continuous task in GIS, requires large bodies of knowledge. The talk presents a ørst step in extracting complex buildings from multiple aerial images. It is an instance of a successful path from the pixels to the labeled 3d-structure. The goal is to generalize the approach.

Basis is a multi-layer representation of a building, each layer adapted to the available (abstract) sensor, which aggeregates pieces of evidence. The expected appearance is made explicit in image models, which allow image centred reasoning, especially grouping.

The problems to be solved are:

- deøning a (building) model which is generic enough to cover a large percentage of real buildings with their geometric description and their meaning in a GIS or planning context,
- ønding several oaths from the original 2d-data or 3d-range data to the ønal interpretation in order to establish experience for automatic control of the interpretation process,
- tracking uncertainty about observables, reasoning and model parts in order to provide consistent self diagnosis.

References (www.ipb.uni-bonn.de/ipb/lit/lit.html)

1. 1995 Braun et al.: Models for Photogrammetric Building Reconstruction, Computer and Graphics, Vol. 19, No. 1, pp. 109˘118, 1995

2. 1996 Englert/Guelch: One-Eye Stereo System for the Acquisition of Complex 3D Building Descriptions, GIS, 4/1996

3. 1997 Brunn et al.: A Multi-Layer Strategy for 3D Building Acquisition, Proc. of IAPR-TC7 Workshop, Graz, 1997, Oldenbourg Verlag,

4. 1997 Lang/Frstner: Surface Reconstruction of Man-Made Objects using Polymorphic Mid-Level Features and Generic Scene Knowledge, Zeitschr. f. Photogrammetrie und Fernerkundung, 6/1997

5. 1997 Fischer et al: Integration of 2D and 3D Reasoning for Building Reconstruction using a Generic Hierarchical Mocel, in Foerstner/Pluemer: Semantic Modeling for the Acquisition of Topographic Information from Images and Maps, Birkhaeuser, 1997

Cooperative Distributed Vision

Takashi Matsuyama

Department of Electronics and Communication
Kyoto University
Sakyo, Kyoto 606, Japan
e-mail: tm@kuee.kyoto-u.ac.jp

This talk gives an overview of our øve years project on Cooperative Distributed Vision (CDV, in short). From a practical point of view, the goal of CDV is summarized as follows:

Embed in the real world a group of network-connected Observation Stations (real time image processor with active camera(s)) and mobile robots with vision, and realize

1. robust, Æexible, and real time dynamic real world scene understanding, and

2. versatile image media capturing, generation, and editing.

Applications of CDV include real time wide area surveillance, remote conference and lecturing systems, interactive 3D TV and intelligent TV studio, navigation of mobile robots and disabled people, cooperative mobile robots, and so on.

The aim of the project is not to develop these speciøc application systems but to establish scientiøc and technological foundations to realize CDV systems enough capable to work persistently in the real world.

From a scientiøc point of view, we put our focus upon Integration of Perception, Action, and Communication. That is, the scientiøc goal of the project is to investigate how these three

functions should be integrated to realize intelligent systems; we believe that intelligence does not dwell solely in brain but emerges from active interactions with environments through perception, action, and communication.

Technological developments by the project include versatile and high precision vision sensors, real time vision hardwares and softwares, robust and

Æexible vision algorithms, communication protocols for cooperation and so on.

In this talk, we proposed a functional dependecy model for an Active Vision Agent (AVA, in short), which perceives, makes actions, and communicates with each other to fuløll a given task. We emphasize that in deøning communication between AVAs, the discrimination between vacuous and embodied AVAs is crucial and that multiple communication links can be established between embodied AVAs.

References

1. URL: `http://vision.kuee.kyoto-u.ac.jp/CDVPRJ/`

The Role of Attention in Knowledge-Based Vision Systems

John K. Tsotsos

John K. Tsotsos

Dept. of Computer Science
University of Toronto
Toronto, Ont. Canada M5S 3G4
tsotsos@cs.toronto.edu
http://www.cs.toronto.edu/Λtsotsos/jkt.html

This presentation provides a 20-year retrospective of my research on the topic of knowledge based vision systems and attention. The development of the ALVEN system a computer vision system to interpret the dynamics of the human left ventricle from X-Ray image sequences, began in 1976 (Tsotsos 1985). This work was quite successful; yet the computer vision community at the time was being so strongly by the Marr philosophy that espoused that top-down inÆuences did not play a role in vision that the particulars of the model were quickly forgotten. Unfortunately, Marr was eventually proved wrong on this point (see Tsotsos 1990). Nevertheless, I set out to satisfy my own belief that attention was a critical component of perception. This led to a series of papers on the subject proving that purely data-driven perception is an intractable problem (Tsotsos 1989, 1990), and further, the size of problems humans routinely solve cannot be solved by the brain in a purely data-driven manner. Attention is one of the mechanisms that can lead to a tractable solution.

This strong theoretical conclusion formed the basis of the Selective Tuning Model for visual attention (Tsotsos et al. 1995). The model was developed in such a way so to not only have strong computational utility but also predictive power for primate perception. To date a number of predictions are gathering signiøcant positive evidence (see my web pages for details).

Moving back towards knowledge-based vision armed with this powerful new model, we embarked on a project to develop a visually-guided robot to

assist physically-disabled children, PLAYBOT (Tsotsos et al. in press; also see the web pages). This robot required a new control mechanism to overcome the inherent intractability of the existing control methods (Tsotsos 1995) for such a complex task; there was a need to enable intermediate presentations, goals, hierarchies and attentive mechanisms in order to seamlessly integrate deliberate with reactive control depending on vision as primary sensor; the S* strategy as developed and is now being tested (Tsotsos 1997).

The overall lessons drawn from the 20 years of research are that if we seek to solve human-like problems and achieve human-like performance (or better) both attentive mechanisms and knowledge of the domain must be utilized to their fullest.

References

1. Tsotsos, J.K., "The Role of Knowledge Organization in Representation and Interpretation of Time-Varying Data: The ALVEN System", Computational Intelligence, Vol. 1, No. 1., Feb. 1985, p16 - 32.

2. Tsotsos, J., "The Complexity of Perceptual Search Tasks", Proc. International Joint Conference on Artiøcial Intelligence, Detroit, August, 1989, pp1571 - 1577.

3. Tsotsos, J.K., "Analyzing Vision at the Complexity Level", Behavioral and Brain Sciences 13-3, p423 - 445, 1990.

4. Tsotsos, J.K., Culhane, S., Wai, W., Lai, Y., Davis, N., NuÆo, F., "Modeling visual attention via selective tuning", Artiøcial Intelligence 78(1-2),p 507 - 547, 1995.

5. Tsotsos, J.K., "On Behaviorist Intelligence and the Scaling Problem", Artiøcial Intelligence 75, p 135 - 160, 1995.

6. Tsotsos, J.K., "Intelligent Control for Perceptually Attentive Agents: The S* Proposal", Robotics and Autonomous Systems 21-1, p5-21, July 1997.

7. Tsotsos, J.K., Verghese, G., Dickinson, S., Jenkin, M., Jepson, A., Milios, E., NuÆo, F., Stevenson, S., Black, M., Metaxas, D., Culhane, S., Ye,

Y., Mann, R., "PLAYBOT: A visually-guided robot to assist physically disabled children in play", Image & Vision Computing Journal, Special Issue on Vision for the Disabled, (in press).

Learning spatio-temporal models

D.C. Hogg

School of Computer Studies
Univerity of Leeds
dch@scs.leeds.ac.uk

The aim of our work is to devise ways for learning spatio-temporal models from passive observation of video sequences. Such models are intended for use in the interpretation of image sequences. We describe two approaches, both of which have been implemented and preliminary results obtained within plausible application domains.

The ørst approach uses a qualitative spatial calculus over semantic regions of the image plane. Both the semantic regions and events represented in the calculus are learned automatically through observation of a scene over extended periods.

The second approach uses a very simple characterisation of the relationship between moving and stationary objects to build a probabilistic model capturing some aspects of typical behaviours. Outliers of this model are labelled as atypical (and therefore interesting) events.

References

1. Fernyhough, J., Cohn, A.G. and Hogg, D.C. (1998), Building Qualitative Event Models Automatically from Visual Input, Proc. IEEE International Conference on Computer Vision, Bombay.

2. Heap, A.J. and Hogg, D.C. (1998), Wormholes in Shape Space: Tracking through Discontinuous Changes in Shape, Proc. IEEE International Conference on Computer Vision, Bombay.

3. Morris, R.J. and Hogg, D.C. (1998), Statistical Models of Object Interaction, Proc. Of Workshop on Visual Surveillance, Bombay.

The Acquisition and Use of Interaction Behaviour Models

Neil Johnson, Aphrodite Galat, David Hogg

School of Computer Studies
The University of Leeds
neilj@scs.leeds.ac.uk

In recent years many researchers have become interested in the development of techniques to allow a more natural form of interface between the user and the machine, utilising interactive spaces equipped with cameras and microphones where such techniques can be developed and tested (see, for example, [1]). In achieving this goal, it is essential that the machine is able to detect and recognise a wide range of human movements and gestures, and this has been a principal avenue of research (see, for example, [2, 3, 4, 5, 6, 7, 8]). We wish to investigate the provision of natural user-machine interaction from a dicœerent angle, allowing the machine to acquire models of behaviour from the extended observation of interactions between humans, and using these acquired models, to equip a virtual human with the ability to interact in a natural way. We describe a novel approach to interaction modelling, using a relatively simple interaction for our experiments - shaking hands.

Training data is acquired by automatically locating and tracking individuals within a corpus of typical interactions. Interactions are modelled by means of a previously developed, statistically based, modelling scheme which allows behaviours to be learnt from the extended observation of image sequences [9]. Interaction is represented as the joint behaviour of object silhouettes just as Kakusho et al. consider joint behaviour in their recognition of social dancing [5]. The model is enhanced to enable the extrapolation of realistic future behaviours.

Having learnt a generative interaction model from the observation of image sequences containing individuals performing simple interactions, interaction with a virtual human is investigated. The model provides information about how an interaction may proceed in the form of a Markov chain, and interaction with a virtual human is achieved by following a route through this

chain such that, as the interaction proceeds, the real human's silhouette continually matches half of the joint silhouette represented within the model. In a Bayesian approach to interaction tracking, multiple interaction hypotheses are stochastically propagated through the model by a method based on Isard and Blake's CONDENSATION [10].

References

1. A.P. Pentland. Machine Understanding of Human Motion. Technical Report 350, MIT Media Laboratory Perceptual Computing Section, 1995.

2. L. Campbell and A. Bobick. Recognition of Human Body Motion Using Phase Space Constraints. In Fifth International Conference on Computer Vision, pages 624-630, June 1995.

3. J.W. Davis and A.F. Bobick. The Representation and Recognition of Action Using Temporal Templates. Technical Report 402, MIT Media Laboratory Perceptual Computing Section, 1996.

4. D.M. Gavrila and L.S. Davis. Towards 3-D Model Based Tracking and Recognition of Human Movement: a Multi-View Approach. In Int. Workshop on Face and Gesture Recognition, pages 272-277, 1995.

5. K. Kakusho, N. Babaguchi, and T. Kitahashi. Recognition of Social Dancing from Auditory and Visual Information. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, pages 289-294, October 1996.

6. S. Nagaya, S. Seki, and R. Oka. A Theoretical Consideration of Pattern Space Trajectory for Gesture Spotting Recognition. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, pages 72-77, October 1996.

7. T.E. Starner and A. Pentland. Visual Recognition of American Sign Language Using Hidden Markov Models. In Int. Workshop on Face and Gesture Recognition, pages 189-194, 1995.

8. A.D. Wilson and A.F. Bobick. Conøguration States for the Representation and Recognition of Gesture. In Int. Workshop on Face and Gesture Recognition, pages 129-134, 1995.

9. N. Johnson and D. Hogg. Learning the Distribution of Object Trajectories for Event Recognition. Image and Vision Computing, 14(8):609-615, August 1996.

10. M. Isard and A. Blake. Contour Tracking by Stochastic Propagation of Conditional Density. In Proceedings of European Conference on Computer Vision, volume 1, pages 343-356, 1996.

Visual Models from Natural Language Description

Amitabha Mukerjee

Department of Mechanical Engineering
Indian Institute of Technology,
Kanpur 208016, India
amit@iitk.ernet.in

Variational Geometry to deøne shape classes are widely used for object models. In addition, dynamic scenes need to model temporal variation (trajectories), and changes in relative position, along with the agent tasks/intentions underlying them. Natural language is a human-oriented mechanism for specifying the constraints on the scene while maintaining a large degree of imprecision. However developing such models also call for a mechanism to display the model as a range of possible scenarios. In this work, continuum constraints are used to arrive at instantiation of NL descriptions such as "there is a bench to the left of the tree". The model has the following information about a speciøc domain (urban parks):

- ≱ geometric shape clases and articulations (e.g. humans = 18-DOF articulation)
- ≱ continuum potential øeld functions ("left of", "near", etc.)
- ≱ action models ("walk", "give to", "pick up") - these are trajecries in the conøguration space of the articulation agents
- ≱ camera viewpoint

These models are integrated to generate 3-D instantiations for verbal input such as "he gave the Æower to the woman". These are then displayed to the user as a graphics animation or a synthetic image sequence. The inverse process, from scene to concepted models, call for the same knowledge elements on a substrate of low-level object tracking.

Probabilistic Methods for Co-Operation and Navigation

Christopher Brown

Dept. of Computer Science
University of Rochester
Rochester NY 14627-0226
brown@cs.rochester.edu
http://www.cs.rochester.edu/u/brown/

Illustrating Professor Neumann's topic of using knowledge of appearances, Randal Nelson and Andrea Selinger of Rochester use appearance-based methods to recognize objects. Low-level features made up of a small number of connected edges are stored in a large associative database and accessed by a clever hashing scheme. Each object may be represented by as many as 100 (processed) views over an entire viewing sphere. The lookup algorithm incorporates an evidence-weighing mechanism to pick the best candidate from the catalog. The system's performance is impressive on catalogs up to size 24 objects. It deals well with multi-object clutter, background texture, and obscuration of part of the target object. Further, it generalizes in the sense of being able to recognize a previously unseen object from a category with instances that it has seen.

On a more preliminary note, at Rochester we are developing Bayes nets and hidden markov models for action- and plan-recognition applications. The domain is dynamic, real-time, cooperating agents pursuing joint plans, and in which re-planning may be needed to cope with surprizes, defections, etc. Bayes nets are to be used for object and situation recognition: we plan to maintain a dynamic version of an occupancy grid, in which state vectors describing a dynamic world are stored in a grid of øxed spatial resolution. This representation raises issues of how to propagate beliefs between cells in space and across time.

The group planning activities will feature multiple cooperating agents, ørst in simulation but ultimately (we hope) in the real world. Each agent must

sense whether the chosen plan is progressing well, or if it has changed, or if it should be changed. In the last two cases, the agent has the problem of choosing his role in a new plan. Ideally, a distributed set of intellligence agents can, simply by observation and using no explicit communication for negotiations, self-organize, form a coherent group plan, and assign roles in the plan to agents. Synchronization issues are vital here to assure that the plan choice and role-assignment process converge stably. We plan to use hidden markov models to recognize the roles being performed by agents who are following some plan, and Bayes nets to recognize actions.

Since much or most of this work is in its early stages, my goal is not to report, much less to recommend, these approaches, but rather to use the Dagstuhl format to inform myself on the possible positive and negative aspects of our proposed approach.

Design and Building Vision-Based Robots

Alan K. Mackworth

Laboratory for Computational Intelligence
Department of Computer Science
University of British Columbia
Vancouver, B.C., Canada V6T 1Z4
mack@cs.ubc.ca

Knowledge-based image interpretation needs to be re-interpreted. The traditional approach, based on the classic Good Old-Fashioned Artiøcial Intelligence and Robotics (GOFAIR) paradigm, proposes that domain-speciøc knowledge is used by the robot/agent at run-time to disambiguate the retinal array into a rich world representation. The argument is that the impoverishment and ambiguity of the visual stimulus array must be supplemented by additional knowledge. This approach has failed to make substantial progress for several reasons. One diŒculty is the engineering problem of building robots by integrating oœ-line knowledge-based vision systems with on-line control-based motor systems. Especially in active vision systems this integration is diŒcult, ugly and ineŒcient. I argue that, with a radical re-interpretation of 'knowledge-based', we can design, build and verify quick and clean knowledge-based situated robot vision systems.

We need practical and formal design methodologies for building integrated perceptual robots. The methodologies are evolving dialectically. The symbolic methods of GOFAIR constitute the original thesis. The antithesis is reactive Insect AI. The emerging synthesis, Situated Agents, has promising characteristics, but needs formal rigor and practical tools. The critiques and rejection by some of the GOFAIR paradigm have given rise to the Situated Agent approaches of Rosenschein and Kaebling, Brooks, Ballard, Winograd and Flores, Lavignon and Shoham, Zhang and Mackworth and many others.

The Constraint Net (CN) model is a formal and practical model for building hybrid intelligent systems as Situated Agents. In CN, a robotic system is modelled formally as a symmetrical coupling of a robot with its environment. Even though a robotic system is, typically, a hybrid dynamic system, its

CN model is unitary. Many robots can be designed as on-line constraint-satisfying devices. A robot in this restricted scheme can be veriøed more easily. Moreover, given a constraint-based speciøcation and a model of the plant and the environment, automatic synthesis of a correct constraint-satisfying controller becomes feasible, as Zhang and I have shown for a simple ball-chasing robot.

These ideas are illustrated by application to the challenge of designing, building and verifying active perception systems for robot soccer players with both oœ-board and on-board vision systems. This work is joint with Ying Zhang and many others in our laboratory.

Vision-Based Navigation

J Kittler, N Georgis

Centre for Vision, Speech and Signal Processing
School of Electronic Engineering, Information Technology
and Mathematics, University of Surrey,
Guildford GU2 5XH, United Kingdom
J.Kittler@ee.surrey.ac.uk

It has been demonstrated by psychophysical experiments that one of the modalities used for navigation by humans is view based. In computer vision terms the view based approach to navigation translates into the task of steering the vehicle towards the epipole in the stored model view. The computation of the epipole (as well as the detection of small obstacles) requires a very accurate estimation of the relative position of corresponding points in model and scene images. It is demonstrated that inaccurate matches give rise to epipole line instability. A novel correspondence analysis technique based on the Robust Hough Transform is presented. The technique exploits image intensity proøle information and a motion model voting scheme based on a robust kernel. It is shown that the accuracy in estimating the corresponding point displacement obtained with this technique leads to a very accurate estimate of the epipolar geometry which makes the view based approach feasible.

Knowledge representation for understanding dynamics scenes

Ernst D. Dickmanns

Institut fuer Systemdynamik und Flugmechanik (ISF)
UniBw Muenchen, Fakultaet fuer Luft- u. Raumfahrttechnik
D-85577 Neubiberg
Tel.: 089 6004 2077/3583; Fax: 089 6004 2082
Ernst.Dickmanns@unibw-muenchen.de

Modeling of processes along the time axis is considered as important as modeling of shapes and trajectories in 3-D space for eŒcient dynamic scene understanding. Parallel use of both diœerential and integral representations in both space and time have shown to be essential for recursive estimation and mission control. Knowledge about the world is aŒxed to object classes with properties of shape and dynamics. Through an ıanalysis by synthesisȷ approach, eŒcient feature extraction from image sequences is controlled by making predictions on the basis of measurement models. Prediction errors are used for adapting both parameters and state variables in the dynamical models.

These more conventional tools from systems dynamics, control engineering, and computer graphics are complemented by knowledge representation methods as developed in AI for situation assessment, adaptation and behavior decision on the higher system levels (e.g. state charts and tree representations); control implementation, however, is realized again by well proven control engineering methods. A four-layer system architecture results.

Experimental results are shown for autonomous road vehicles in public traŒc and for air vehicles maneuvering near the ground (video sequences).

References

1. Dickmanns, E.D.; Graefe, V.: a) ‚Dynamic monocular machine vision", b) ‚Application of dynamic monocular machine vision". J. Machine Vision Application, Springer-Int., Nov. 1988, pp 223-261

2. Dickmanns, E.D.; Christians, T.: ‚Relative 3D-state estimation for autonomous visual guidance of road vehicles". In: Kanade, T. e.a. (Eds): 'Intelligent Autonomous Systems 2', Vol. 2, Amsterdam, Dec. 1989, pp 683-693

3. Schick,J.; Dickmanns, E.D.: ‚Simultaneous Estimation of 3D Shape and Motion of Objects by Computer Vision". IEEE-Second Workshop on Visual Motion, Princeton, Oct. 1991

4. Dickmanns, E.D.; Mysliwetz, B.: ‚Recursive 3D Road and Relative Ego-State Recognition". IEEE-Trans. PAMI, Vol.14, No.2, Special Issue on 'Interpretation of 3D Scenes', Febr. 1992, pp 199-213

5. Dickmanns, E.D.; Fagerer, C.; Dickmanns, D.: ‚Visual Grasping with Long Delay Time of a Free Floating Object in Orbit". 4th IFAC Symposium on Robot Control (SY.RO.CO.'94), Capri, Italy, 1994

6. Dickmanns, E.D.: ‚Performance Improvements for Autonomous Road Vehicles". Int. Conf.on Intelligent Autonomous Systems (IAS-4), Karlsruhe, 1995

7. Fuerst, S.; Werner, S.; Dickmanns, D.; Dickmanns, E.D.: ‚Landmark navigation and autonomous landing approach with obstacle detection for aircraft". AeroSense '97, Conference 3088, Orlando FL, April 20-25, 1997

8. Dickmanns, E.D.: ‚Vehicles Capable of Dynamic Vision". 15th Int. Joint Conf. on Artiøcial Intelligence (IJCAI-97), Nagoya, Japan, August 23-29, 1997

Intention Recognition Based on a Fuzzy Metric Temporal Logic Approach to Image Sequence Analysis

H.-H. Nagel, M. Haag, K. H. Schfer (and others)

Institut fr Algorithmen und Kognitive Systeme (IAKS)
Fakultt fr Informatik
Universitt Karlsruhe (TH)
D-76128 Karlsruhe, Ge rmany

Fraunhofer-Institut fr Informations- und Datenverarbeitung (IITB)
Fraunhoferstr. 1
D-76131 Karlsruhe, Germany
hhn@iitb.fhg.de

Knowledge-based Computer Vision is studied by the design and implementation of a complete system which converts digitized image sequences into natural language text describing relevant developments in innercity road traŒc scenes recorded by a stationary B/W-video camera. A Signal Layer automatically detects, initializes, and tracks images of road vehicles, using a model-based approach [Kollnig &Nagel 97]. Fuzzy transductors (see [Kollnig & Nagel 93]) transform the geometric tracking results into elementary conceptual descriptions which are transmitted to the Inference Layer. An inference engine based on Fuzzy Metric Temporal Horn Logic (see [Schfer 96]) condenses the elementary into more abstract conceptual descriptions which are processed by the Natural Language Text Generation Layer ˘ see [Gerber & Nagel 96].

Explicitly represented knowledge comprises polyhedral object models (sedan, van, bus, truck with trailor, etc.), lane conøgurations at intersections and gas stations, illumination, as well as generic vehicle motion. The inference engine maps elementary conceptual descriptions to a path through a Situation Graph (see [Nagel 88, Nagel 91]) which represents schematic knowledge about the situated concatenation of elementary vehicle maneuvers into maneuver sequences, taken to represent intentions of a vehicle as an agent.

Results obtained for extended image sequences (several thousand frames) of various intersection and gas station scenes illustrate the viability of this approach ˘ see, e.Λg., [Haag et al.97], [Haag & Nagel 98]. This system provides a valuable tool for further research and methodological improvements.

References

1. R. Gerber and H.-H. Nagel: Knowledge Representation for the Generation of Quantiøed Natural Language Descriptions of Vehicle TraŒc in Image Sequences. Proc. IEEE International Conference on Image Processing (ICIPΛ'96), Lausanne/CH, 16˘19 September 1996, Vol.ΛII, pp.Λ805˘808.

2. M. Haag and H.-H. Nagel: Incremental Recognition of TraŒc Scenes from Video Image Sequences. In H.Buxton and A.Mukerjee (Eds.), Proc.Workshop on Conceptual Descriptions, 2 January 1998, Mumbai/India (to appear).

3. M. Haag, W. Theilmann, K. Schfer, and H.-H. Nagel: Integration of Image Sequence Evaluation and Fuzzy Metric Temporal Logic Programming. In ıKI-97: Advances in Artiøcial Intelligenceȷ, 21st Annual German Conference on Artiøcial Intelligence, 9-12 September 1997, Freiburg/Germany; G. Brewka, Ch. Habel, and B. Nebel (Eds.), Lecture Notes in Artiøcial Intelligence 1303, Springer-Verlag Berlin Heidelberg New York/NY 1997, pp.Λ301-312.

4. H. Kollnig and H.-H. Nagel: Ermittlung von begriØichen Beschreibungen von Geschehen in Strassenverkehrsszenen mit Hilfe unscharfer Mengen Informatik ˘ Forschung und Entwicklung 8 (1993) 186˘196 (in German).

5. H. Kollnig and H.-H. Nagel: 3D Pose Estimation by Directly Matching Polyhedral Models to Gray Value Gradients. International Journal of Computer Vision 23:3 (1997) 283˘302.

6. H.-H. Nagel: From Image Sequences towards Conceptual Descriptions. Image and Vision Computing 6:2 (1988) 59˘74.

7. H.-H. Nagel: La representation de situations et leur reconnaissance partir de s quences d'images ˘ The Representation of Situations and Their Recognition from Image Sequences. Proc.∀ ^{e} Congrs Reconnaissance des Formes et Intelligence Artiøcielle, Lyon-Villeurbanne/France, 25˘29 Novembre 1991, pp.Λ1221˘1229.

8. K. Schfer: Unscharfe zeitlogische Modellierung von Situationen und Handlungen in Bildfolgenauswertung und Robotik. Dissertation, Fakultt fr Informatik der Universitt Karlsruhe (TH), Juli 1996; published in: Dissertationen zur Knstlichen Intelligenz (DISKI) 135, inøx-Verlag, St.ΛAugustin/Germany 1996 (in German).

Interpretation of Image Sequences for Visual Surveillance

M. Thonnat

2004, route des Lucioles - B.P. 93
06902 Sophia Antipolis Cedex, France
Monique.Thonnat@inria.fr

In this talk we have presented the work we have done for the last three years in automatic image interpretation. The class of applications we are interested in is visual surveillance of real-world scenes with a øxed monocular color camera. First, we have proposed a general complete architecture of a system taking as input the image sequence and presenting alarms to a human operator. These alarms are generated based on the behavior analysis of mobile objects (like humans and vehicles) involved in human activities.

The interpretation system is based on a mobile region detection module, a tracking module and a scenario recognition module. In particular we have discussed the role of contextual information, describing the static environment, for behavior analysos and image procesing error recovery.

Then, we have presented results achieved in the framework of the european PASSWORDS Esprit project. Finally, current research performed in a new Esprit project, AVS-PV, for prevention of vandalism in metro stations has been described.

References

1. F. Bremond: Environnement de rsolution de problmes pour l'interprtation de squences d'images, PhD, Univ. Nice october 97

2. F. Bremond, M. Thonnat: Issues of representing context illustrated by video-surveillance applications, Inter. Journal of Human-Computer Studies Special Issue on Context, 1997

3. F. Bremond, M. Thonnat: Issues in representing context illustrated by scene interpretation applications, CONTEX'97,february 1997.

4. F. Bremond, M.Thonnat:Analysis of human activities described by image sequences FLAIRS'97, mai 1997.

5. F. Bremond, M. Thonnat: Tracking multiple non-rigid objects in a cluttered scene, SCIA'97, june1997.

6. F. Bremond, M. Thonnat: Recognition of scenarios describing human activities, International Workshop on Dynamic Scene Recognition from Sensor Data, juin 1997.

7. F. Bremond, M. Thonnat: Object tracking and scenario recognition for video-surveillance, IJCAI'97, august 1997

8. F.Bremond, M.Thonnat: A context representation for surveillance systems, ECCV Worshop on Conceptual Descriptions from Images, April 1996.

9. N. Chleq, C. Regazzoni, A. Teschioni, M. Thonnat: A visual surveillance system for the prevention of vandalism in the metro stations, EMMSEC'97,november 1997.

10. N. Chleq, M. Thonnat: Realtime image sequence interpretation for video-surveillance applications, ICIP , Lausanne September 1996.

A Hybrid Approach to Identifying Objects from Verbal Descriptions

Gerhard Sagerer

Technische Fakultt
Universitt Bielefeld, Germany
sagerer@techfak.uni-bielefeld.de

Human-computer interaction using means of communication which are natural to humans, like speech or gesture, is a challenging task. In the talk, we address to the problem of fusing the understanding of spoken instructions with the visual perception of the environment. The visual recognition of objects is realized via a hybrid approach attaching probabilistic formalisms, like artiøcial neural networks or hidden Markov models, to concepts of a semantic network. Additionally, an eŒcient processing strategy for image sequences is used propagating the structural results of the semantic network as an expectation for the next image. This method allows to produce linked results over time supporting the recogntion of events and actions. The identiøcation of objects from verbal descriptions is based on a Baysian network approach. The objects with the highest joint probability of being observed in the scene and being intented in the instruction are identiøed using the common qualitative representation for observed ıtypeȷ, ıcolorȷ and ıspatial relationȷ as well as the uttered ıtypeȷ, ıcolorȷ, ısizeȷ, ıshapeȷ and ıspatial relationȷ. The parameters of the Baysian network are estimated from the results of psycholinguistic experiments, from a WWW-questionnaire, and from the confusion matrices of the object recognition module.

Representations for Vision and Language

Hilary Buxton

Cognitive and Computing Sciences
University of Sussex
Falmer, Brighton, BN1 9QH, UK
hilaryb@cogs.susx.ac.uk

This paper contrasts two ways of forming conceptual descriptions from images. The ørst, called ımonitoringȷ, uses little top-level control, instead just following the Æow of data from images to interpretation. The second, called ıwatchingȷ, emphasizes the use of top-level control, and actively selects evidence for task-based descriptions of the dynamic scenes. Here we look at the eœect this has on forming conceptual descriptions.

First, we look at how motion verbs and the perception of events contribute to an eœective representational scheme. Then we go on to discuss illustrated examples of computing conceptual descriptions from images in our implementations of the monitoring and watching systems.

Finally, we discuss alternative approaches and conclude with a discussion of future work.

References

1. AIJ-78 Hilary Buxton and Shaogang Gong 'Visual surveillance in a dynamic and uncertain world'

2. AIJ-forthcoming Richard Howarth 'Interpreting a dynamic and uncertain world: task-based control'

Bayesian Networks for Building Recognition

Karl-Rudolf Koch

Institute of Theoretical Geodesy
Rheinische-Friedrich-Wilhelms University of Bonn
Nuallee 17, 53115 Bonn
koch@theor.geod.uni-bonn.de

Our ørst experience in knowledge based computer vision was gained with the interpretation of digital, multispectral, aerial images by Markov random øelds. The image analysis is performed at two levels, at the low level the image is segmented, to obtain regions, at the high level the regions get a meaning. The image model and the object model are expressed by Markov random øelds, i.e. labeling processes for the pixels at the low level and for the regions at the high level (Klonowski and Koch 1997). The advantage of this aproach is the simplicity of the semantic modeling, the disadvantage is a lacking Æexibility. An alternative to the Markov random øelds are the Bayesian networks. They can be constructed such that the same results are obtained as with Markov random øelds and they allow more Æexibility with respect to modeling. But the question, how to construct them, has to be answered. Dicœerent forms of Bayesian neworks were tried for the recognition of buildings. It turned out, the best result were obtained when the Bayesian networks were built up dynamically on the adjacency graph for the faces of the buildings which are visible in the digital image. The faces form the aspects which lead to the buildings. At the root of the Bayesian networks the nodes are situated with the data characterizing the faces. The aspects depend on the faces, on the vertices of the adjoining faces and on the form of the adjacency graph. Finally, the buildings with their frequency of their appearance as prior information depend on the aspects. First results for the recognition of buildings are presented in (Kulschewski 1997).

References

1. Klonowski, J. and K.R. Koch (1997) Two Level Image Interpretation Based on Markov Random Fields. In: Foerstner, W. and L. Pluemer

(Hrsg.), Semantic Modeling for the Acquisition of Topographic Information from Images and Maps. Birkhaeuser Verlag, Basel, 37-55.

2. Kulschewski, K. (1997) Building Recognition with Bayesian Networks. In: Foerstner, W. and L. Pluemer (Hrsg.), Semantic Modeling for the Acquisition of Topographic Information from Images and Maps. Birkhaeuser Verlag, Basel,196-210.

Hierarchical Probabilistic Object Detection

Ansgar Brunn

Institut fr Photogrammetrie
Universitt Bonn
Nussallee 15, D-53115 Bonn, Germany
ansgar@ipb.uni-bonn.de

In my talk I presented an approach for object detection in vectored datasets which have been acquired by various sensors in a reference system. Statistical variances of the observations are assumed to be known from physical models or empirical evaluations.

The approach consists of two main steps: The ørst step is the generation of a vector valued data pyramid similar to image pyramids. In the second step a probabilistic network is set up which consists of equally subnets for each element of the pyramid. The subnets condense all information available at each element of the pyramid using the Bayes' rule. The measurements are introduced via normal distributions. By top down forward propagation the probabilities of interest of each element on each level are computed. Following only the most interesting branches of the statistical net an exterior algorithm can lead immediately to the most interesting parts of the dataset.

Examples of this fast and robust approach have been taken from building detection using digital elevation models acquired by an airborne laser scanner.

Bayesian Inferencing on Brain MRI Images

Ruzena Bajczy, James C. Gee, Alexei Machado

GRASP laboratory, Computer and Information Science Department
Neurology department
University of Pennsylvania,Philadelphia,PA.19104.
bajcsy@central.cis.upenn.edu

For the past øfteen years (since 1982), we have been pursuing research at the GRASP Laboratory on the quantitative analysis of human brain morphology based on the idea of having available an apriori pictorial representation of the anatomy of the brain, i.e., an anatomy atlas, and an elastic matching procedure that can adjust the structures in the atlas to øt the measurements obtained via computed tomography, and later magnetic resonance imaging. We have, through the years, reøned the procedure so that it now produces satisfactory results in comparison with human analysis of the same data. Put it simply, we can warp one set of brain images into another set (a reference set), and if the ørst group represents normal subjects of the same gender and age, for example, acquired with the same imaging modality, then this procedure enables a map of anatomic variations within that population to be constructed.

To realize these analyses, a variety of equally important issues must be addressed, including: a measurement model for image matching; the construction and application of prior models for brain shape variation; the extraction of salient image features through classiøcation of voxels that are possibly of mixed tissue type; and the development of numerical algorithms that make practical the application of the methods. We shall present our approach to these issues, demonstrate their application in recent experiments, and discuss open problems that remain in the development of a comprehensive methodology for computational anatomy.

EŒcient Topological Indexing for 2-D Object Recognition

Sven J. Dickinson

Department of Computer Science and Center for Cognitive Science
Rutgers University
New Brunswick, NJ 08903
sven@cs.rutgers.edu

In this work, we develop a shape indexing mechanism that maps the topological part structure of a 2-D object's silhouette into a low-dimensional vector space. Based on an eigenvalue characterization of a shock tree, this topological signature allows us to eŒciently retrieve a small set of candidates from a database of models, to which we apply a recent shape matching algorithm (see ICCV '98). In order to build invariance to occlusion, deformation, scale, translation and rotation, local evidence is accumulated in each of the object's topological suspaces. We demonstrate the approach with a series of experiments.

This talk represents joint work with:

Ali Shokoufandeh
Department of Computer Science and Center for Cognitive Science
Rutgers University
New Brunswick, NJ 08903

and

Kaleem Siddiqi and Steven W. Zucker
Center for Computational Vision & Control
Yale University
New Haven, CT 06520-8285

The Appearance Manifold as a Formal Foundation for Vision

James L. Crowley

Professor, I.N.P. Grenoble
Projet IMAG-PRIMA
INRIA Rhones Alpes
655 Ave de l'Europe
38330 Montbonnot, France
http://pandora.imag.fr/Prima/jlc.html

For the last 30 years, computer vision has has relied on assemblies of ad hoc techniques. To mature, the øeld must adopt mathematical foundations that makes it possible analyse and predict the performance of systems and components. We propose such a foundation based on a representation of local appearance.

We begin with a criticism of edge based image description. Edges are shown to be a very partial description of visual information, appropriate only for polyhedral objections. Edge detection is unreliable for many of the interesting structures in an image such as corners and curved surfaces. Correcting for broken and missing edges generally leads to the use of grouping procedures which are ad hoc and have high computational complexity. The end result is a system which can work under laboratory conditions, but which is impossible to analyse.

We then examine the computer vision problem from the perspective of information theory. In such a framework, recognition is expressed as template matching and the basic operation is minimizing a distance metric. In computer vision this is known as SSD (Sum of Squared Dicœerences). We show that this approach allows an analysis of the probability of error. However, this approach has two major Æaws: 1) It requires searching over a very large set of templates, and, 2) it applies to a closed universe deøned by images which have been previously seen. In the remainder of the talk we address these two problems.

We replace search by table lookup using the concept of a window space (W-space). A W-space uses a neighborhood of M pixels as an orthogonal basis. In this representation, each neighborhood of an image is a point. An image is an ordered set of points which comprise a surface. Changing the illumination or viewing position deforms this surfaces, giving a manifold. We show how to reduce the memory requirements (the information capacity) of a W-space by projection to a linear subspace deøned by principal components analysis. We present examples of structural and statistical characterisations of appearance manifolds using the images from the Columbia data base. This approach provides object identity and pose by table lookup in a formally analysable manner.

We then show how to extend the method to an open universe by interpolation and extrapolation. We obtain invariance to 2-D rotation and to scale by interpolation between ølters and by the use of a Gaussian Pyramid. We obtain invariance to illumination by extrapolation using functional approximation. The result is a new approach to object representation and recognitions, and a new formalism for design and analysis of computer vision systems.

Model-based 3D Recognition and Localization from Single Perspective Views

Stefan Lanser

Forschungsgruppe Bildverstehen (FG BV), Prof. Radig
Technische Universitaet Muenchen
lanser@informatik.tu-muenchen.de

Most of this talk is about MORAL. Regardless of whatever other associations the reader might have, in this context MORAL stands for Munich Object Recognition And Localization. This is a system for the recognition and localization of task relevant (articulated) rigid polyhedral 3D objects in the context of autonomous mobile Systems (AMS). The recognition stage is view-based, the pose estimator performs a full perspective 2D-3D matching. The construction of associations linking (2D) model lines and corresponding image features, the construction of hypotheses containing the maximum consistent set of associations per view, as well as the ønal search for corresponding lines during the 3D pose estimation incorporate two kind of measurements: GEOMETRIC MEASURES (the dicœerence of orientation in the view-based recognition and the alignement error after the 2D-3D pose estimation), and TOPOLOGICAL MEASURES. For the latter simple, but robust local constraints are classiøed into a set of conøgurations which are quasi-invariant against changes of the camera pose. The presented approach is not able to handle classes of objects, but only instances. On the other side, the use of exact 3D models imposes powerful constraints allowing a robust recognition and accurate localization of objects. This enables MORAL to support a variety of manipulation and navigation tasks.

Towards an Active Visual Observer

Jan-Olof Eklundh

umerisk analys och datalogi, NADA
KTH
S-100 44 Stockholm, Sweden
joe@bion.kth.se

Studying seeing agents and perception-and-action from a general perspective is a far-reaching and diŒcult undertaking. Three aspects that are particularly important are:

- The systems perspective (which complements traditional reductionist analysis.
- Time, i.e. that processing as well as output is time-dependent (note the diœerence between natural and pictorial vision).
- The task dependence that it inherent in the perception-action cycle.

Our goal is to understand and develop principles of an active visual observer that has certain capabilities to navigate and manipulate objects in the world using vision. We believe that building such systems must be part of such an eœort, since we need to study the systems empirically.

In this context knowledge has to be included at all levels, from the early perceptual level to the behavioral and cognitive levels. In the presentation we consider aspects of knowledge at the early stages.

Our starting point is as follows: An active observer is at any moment involved in a set of tasks. These tasks prime a set of models and processes, which of course in turn change over time due to goal fuløllment, new tasks, new information and other changes of the world and in the obeserver.

Using this approach we show that simple modules for e.g motion, depth and shape estimation can be combined to address the øgure-ground segmentation problem and hence provide higher level methods for recognition and scene

understanding with appropriate and uncluttered information. The main problems instead become fusion of and selection between cues. We have in addition to traditional uncertainty based methods applied simple votings schemes for this. By voting we can combine information from cues of dicœerent types by assuming that coincidences are non-accidental. The knowledge aspect in this approach is that each module is highly specialized: it assumes knowledge about the world that may or may not be true. Subsequent processing resolves that problem. Experiments with a mobile robot øxating and "masking out" independently moving objects illustrate the approach. Another set of experiments deal with rapid detection of conspicuous planar surfaces.

Semantic Networks in Active Vision Systems - Aspects of Knowledge Representation and Purposive Control

Ulrike Ahlrichs

Chair of Pattern Recognition
University of Erlangen-Nuremberg
ahlrichs@informatik.uni-erlangen.de

The goal of this work is the visual exploration of a scene in real-time. We constrain ourselves to the following three technical restrictions: we use an explicit representation of knowledge about objects and their features, spatial relations between objects, and the steps which are necessary to solve a certain task. In order to exploit the represented knowledge, a problem independend modiøed A-star graph search algorithm should be applied [Nie90] and the knowledge should be represented as a semantic network based on the ERNEST-philosophy [Sag97]. In addition, the inÆuence of active vision on the representation and the control algorithm should be examined.

The application domain of the system is the analysis of oŒce scenes with an active camera where the camera's øeld of view is too narrow to get the whole information of a scene without camera movements.

In this talk a model of saccades developed by Takacs [Tak96] is introduced which distinguishes three diœerent types of saccades: attentive, conditional, and reÆex saccades. These diœerent types of saccades are used to model the camera movements. The saccades are represented via a semantic network which consists of concepts and links. The links describe three diœerent types of relation between concepts. They are referred to as specialization, part or concrete link. The semantic network contains the scene representation and actions like saccades which are connected by concrete links. The diœerent kinds of saccades are provided on diœerent levels of abstraction.

References

1. [Nie90] Niemann, H.; Sagerer, G.; Schroeder, S.; Kummert, F.: ERNEST: A Semantic Network System for Pattern Understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Bd. 9, 1990, S. 883-905.

2. [Sag97] Sagerer, G.; Niemann, H.: Semantic Networks for Understanding Scenes, Advances in Computer Vision and Machine Intelligence, Plenum Press, New York and London, 1997.

3. [Tak96] Takacs, B.; Wechsler, H.: Attention and Parrern Detection Using Sensory and Reactive Control Mechanism, International Conference on Pattern Recognition, Wien, 1996, S. 19-23.

The Dialogue with the Scene:Probabilistic Knowledge Based Active Vision

Joachim Denzler

Lehrstuhl fr Mustererkennung
Universitt Erlangen/Nrnberg, Germany
denzler@informatik.uni-erlangen.de

In speech recognition and understanding systems, for example automatic train time table inquiries, the dialogue manager is an important module. In the case of ambigue or missing information presented by the user, the system has to ask questions, to get the missing information or to resolve ambiguities. A similar problem arises in the øeld of active vision, especially for autonomous mobiles systems. Such systems have to purposively gather information from the environment and have to react on events in the scene, in order to solve a given task. Visual actions have to be chosen, for example, changes of the camera parameter (zoom, vergence, focus) or taking another view of an object to get a more robust classiøcation result.

To choose a suited action out of a set of possible ones, the system has to understand the scene and the meaning (semantic/pragmatic) of objects in the scene with respect to a given task. Recently, in speech understanding the concept of semantic attributes has been proposed [Haas97]. For a given area of application, each word in a spoken sentence, is mapped to a semantic attribute out of a set of possible ones. The mapping is done by a probabilistic assignment function, which can be estimated by a labled training set. In this training set, the assignment needs not to be known.

This concept is transferred to the problem of extracting the semantic/pragmatic of objects in a scene in a special task context. For example, for solving the task "Leave the room" the system has to look for obstacles and destinations. Having assigned a certain object to the semantic attribute 'obstacle', the system can react by tracking the obstacle and avoiding it. The reaction in most cases only depends on the assigned semantic attribute and not on the object itself. This means, that a set of basic visual and navigational actions can be connected with semantic attributes, independent to the

objects. The assignment, i.e. the probabilistic mapping of certain objects to semantic attributes is task dependent and can be learned during a training step.

We show, that this concept of semantic assignment øts well in the context of classical knowledge bases and processing, like semantic networks [Niemann90]. There, so called holistic instantiations at the object level have been proposed [Moratz94]. The probabilistic assignment in terms of holistic instantiation can be interpreted as a holistic instantiation at the semantic level of a semantic network.

In the future we will integrate relations between objects in the probabilistic assignment function. This can be done at the object level (a handle on the door means 'open it', a handle on a cup mean 'carry it'), at the semantic level (a rubbish bin is an obstacle, if it is on the motion plane), or at both levels (a rubbish bin is an obstacle, if it is on the Æoor).

References

1. J. Haas, J. Hornegger, R. Huber, and H. Niemann: Probabilistic semantic analysis of speech. In Paulus and Wahl, editors, Mustererkennung 1997, pages 524 ˘ 531, Berlin, September 1997. Springer.

2. R. Moratz, S. Posch, and G. Sagerer. Controlling multiple neural nets with semantic networks. In W. G. Kropatsch and H. Bischof, editors, Mustererkennung 1994, pages 288˘295, Wien, July 1994. Informatik Xpress 5.

3. H. Niemann, G. Sagerer, S. Schrder, and F.ΛKummert. Ernest: A semantic network system for pattern understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 9:883˘905, 1990.

Current Research Eœorts in Use of Function in Computer Vision

Louise Stark, Melanie Sutton, Kevin Bowyer

University of the Paciøc
3601 Paciøc Avenue Stockton, CA 95211 USA
lstark@uop.edu

One of the main goals of the work in function-based reasoning in computer vision is to develop alternative approaches to generic object recognition and manipulation. A vision system cannot be expected to hold an explicit model, or even a parameterized model of each object that may be encountered in the environment. Reasoning about function has been addressed at diœerent levels in a number of øelds including psychology, artiøcial intelligence, computer vision and robotics.

In psychology work has been done to try to understand how humans form category concepts. In AI it has long been recognized that object categorization can be aided by associating function to structure. Most work in AI starts with a symbolically labeled object. It is the task of computer vision and robotics researchers to make that association. To automatically symbolically label the structure, information must be extracted from sensor data. Diœerent approaches include deriving function from shape-based reasoning, deriving function from motion (i.e. observing an object in use), and also through object manipulation.

Recognition for Action: On the Use of Functional Knowledge

Ehud Rivlin

Computer Science Department,
Technion, Technion City,
Haifa 32000, Israel
ehudr@cs.technion.ac.il

Function-based recognition takes place when an object is evaluated in the context of action or activity. An object can suit a purpose, fuløll a function. If an agent recognizes this, it has in eœect recognized the object. To perform this type of recognition we need on one hand a deønition of the desired function, and on the other the means of determining whether the object can fuløll that function. To ønd out if an object can fuløll a function we need to perform various partial recovery tasks.

Objects can be categorized in terms of functionality, i.e. usefulness to a given agent in performing speciøc purposes or tasks. We ørst consider the class of tasks which involve a transfer of force between a user and a recipient object. A device that transfers force, i.e. that converts an input force into an output force, is called a machine. We show that the requirements of eŒcient force transfer and geometric simplicity lead to a description of a machine as consisting of primitive parts, deøned in terms of (qualitatively speciøed) properties, which can be mapped into functional requirements. Many common household objects, including furniture, appliances, and tools, can be regarded as ımachinesȷ. We show how such objects can be categorized.

To show the general applicability of these principles we demonstrate how they can be used in various domains. We show that documents can be analyzed as functional objects (information machines) which facilitate the transfer of information from the author, across both time and space, to the reader. We claim that in general, the role of a document can be characterized by the functions of the documents components. The function of each component can in turn be derived from its physical attributes and from its relationship with

other components. If a functional description of a document can be derived, it provides insight into the category of the document, and ultimately into strategies for automatic interpretation.

The same approach can be used for scene categorization and interpretation. Having prior knowledge about the scene type, functional hypotheses concerning the area in the current focus of interest is raised. A functional hypothesis consists of a functional model which describes the relationship of the functional area to its surrounding neighborhood. Roads permit passage of vehicles for dicœerent parts of a site. Parking areas provide storage of vehicles. There are certain relationships between the size of the site, its components (e.g buildings), and the size of the roads, of the nearby parking areas, etc. We show how such analysis can be applied to a parking lot as a functional area. We establish a connection between the dicœerent parts of the functional area, and evaluate them. This information is used for categorization, in which the functional area is classiøed as some class of a parking lot.

Function based recognition tries to achieve a mapping from function to form. When an agent has some action to carry out an appropriate object is searched for. Observing an acting agent trying to perceive what is the action taking place involve an inversion of this mapping. Since the mapping from function to form is many to many, we need the information provided by motion to enable us to infer what is the mapping that the acting agent did, exactly. We show how, given a model of an object, we can use the motion of the object, while it is being used to perform a task, to determine its function. Our analysis results in couple of motion descriptors, which are compared with stored descriptors that arise in known motion-to-function mappings to obtain function recognition. We show how these type of descriptions can be used to categorize well-designed objects (mechanisms) using a language which describes their dicœerent behaviors.

Knowledge in image-based scene representation

Vaclav Hlavac

(in collaboration with Tomas Werner and Tomas Pajdla)

Czech Technical University
hlavac@vision.felk.cvut.cz

For visualizing 3-D scenes described by a set of reference intensity images, image-based scene representation is believed to be an alternative to 3-D model reconstruction/rendering. We address these questions: Are image-based approaches as general as model-based ones? Can an arbitrary 3-D scene be visualized without reconstructing a consistent 3-D model? We believe it is important to clarify the power and usefulness of image-based approaches because many works on this topic appeared recently, yet none of them deal with this issue explicitly.

We show that the answer to both questions above is negative. We further argue that it is plausible to distinguish three groups of approaches to visualizing 3-D scenes: image interpolation, image extrapolation, and 3-D model reconstruction/rendering. Image interpolation is the simplest one yet its applicability is limited, 3-D model reconstruction/rendering is general but diŒcult.

We advocate image extrapolation as a trade-oœ useful in practice. It is able to visualize correctly the part of aΛgeneral 3-D scene that is visible from two reference images. In fact, image extrapolation can be also considered the reconstruction and rendering of a partial projective 3-D model.

Perceptual, Functional, Prototypical Knowledge: Which Knowledge for Which Application?

Giovanni Adorni, Stefano Cagnoni

Department of Computer Engineering
University of Parma
Viale delle Scienze
43100 Parma - Italy
adorni@ce.unipr.it

During the talk diœerent aspects of diœerent kind of knowledge have been discussed together with possible applications. Subjective Contours [4] and Structural Information Theory [2] are examples of tools to extrapolate "Perceptual Knowledge" used to recover hidden part of images due to occlusions and noise, before a higher level recognition process. Object recognition is basically a matching process, where the 3-D structures possibly inferred or reconstructed from the visual data are compared to a set of prototypes, each describing a class of objects. If the prototypes incorporate only structural (geometrical) knowledge, that is, they try to describe "What objects look like", it is computational expensive and diŒcult to cope with all the possible shapes of the objects belonging to the same class, without loosing to much discrimination power. Therefore, recognition can be approached from a "functional" point of view, deøning a framework within which the class prototypes can be described [6]. Therefore, "What objects are for" has been another topic of the discussion: functional decomposition of compound objects in elementary objects [5], together with a description of spatial relationships between such elementary objects [1], can then became an interesting tool for object recognition purposes [6]. A ørst application discussed has been a rule-based system, whose rules encode heuristic knowledge as well as perceptual knowledge and spatial relationships between elementary øgures [3]. A second application discussed has been the lane detection system on board of the Mob-Lab, a Mobile Laboratory developed during the European Prometheus

project for testing real-time computer vision aids for a safer vehicle guidance. Lane or road detection is performed by means of a cellular automata paradigm taking advantage of some geometrical knowledge about the lane or road to be detected [7]. The last application discussed has been a multi-agent system that allow a mobile agent (i.e., a mobile robot) to navigate in indoor environments solving navigation conÆicts, if any, with other physical agents through the use of road traŒc signs and rules. Robots move themselves following lines drawn on the Æoor by means of a ccd camera. The ccd camera is also used to recognize traŒc signs [8]. TraŒc signs recognized by the camera, information on the neighbor agents acquired through infrared and ultrasound sensors, and the knowledge about traŒc rules, are used to coordinate robots interactions [9]. TraŒc signs are recognized by means of a neural network approach taking advantage of the knowledge about the typical position of the signs inside the environment and of a focus of attention mechanism to reduce the amount of data to be processed and to avoid that patterns which have not been considered during the network training phase create unpredictable behavior and generate spurious responses [10].

References

1. M. DiManzo, G. Adorni and F. Giunchiglia, Reasoning about Scene Descriptions, Proceedings of the IEEE, Vol.74(7), pp.1013-1025, 1986.

2. G. Adorni and L. Massone, Coding Patterns, in ARTIFICIAL INTELLIGENCE II: Methodology, Systems, Application, Ph. Jorrand e V. Sgurev (eds.), Elsevier Science Publishers B.V. (North Holland), pp.395-402 (1987).

3. G. Adorni, L. Massone, G. Sandini and M. Immovilli, From Early Processing to Conceptual Reasoning: an Attempt to Fill the Gap, in Procs. 10th. IJCAI, pp.775-778, Milan, August 1987.

4. G. Adorni, L. Massone and M. Sambin, Subjective Contours: a Computational Approach, in Procs. 7th. International Congress of Cybernetics and Systems, pp.29-33, London, September 1987.

5. M. DiManzo, E. Trucco, F. Giunchiglia and F. Ricci, FUR: Understanding Functional Reasoning, Journal of Intelligent Systems, Vol.4, pp.431-457, 1989.

6. G. Adorni, Spatial Reasoning as a Tool for Scene Generation and Recognition, in: Human and Machine Vision - Analogies and Divergencies, V. Cantoni (ed.), Plenum Press, pp. 289-318 (1994).

7. G. Adorni, A. Broggi, G. Conte and V. D'Andrea, Real-time image processing for automotive applications, in: Real-time imaging: Theory, techniques, and applications, P.A. Laplante e A.D. Stoyenko (eds.), IEEE Press, New York, pp. 161-194 (1996).

8. G. Adorni, G. Destri and M. Mordonini, Indoor vehicle navigation by means of signs, In: Procs. 1996 IEEE Intelligent Vehicles Symposium, pp. 76-81, Tokyo, September 1996.

9. G. Adorni, M. Mordonini and A. Poggi, A multi-agent system for mobile robots coordination, In: Procs. 1997 IEEE Conference on Intelligent Transportation Systems, Boston, MA, November 1997.

10. G. Adorni, M. Gori and M. Mordonini, Just-in-Time Sign Recognition, Real-Time Imaging, Vol.4(4), 1998, In Press.

Learning Accurate Engineering Models from Shown Examples

Bob Fisher

Department of Artiøcial Intelligence
University of Edinburgh
rbf@aifh.ed.ac.uk

This talk describes the reconstruction of 3D engineering parts to high tolerances (eg. 25 micron feature position accuracy) while not simultaneously degrading the reconstruction of unconstrained surfaces. The data used was multiple range images taken from all sides of the part. To avoid problems arising from fusing symbolic descriptions, we have adopted an alternative 3 stage strategy that:

1. Fuses the XYZ points from the individual views into a combined view (by using pairwise geometric histograms - a new representation of local surface shape). This allows identiøcation of corresponding surface points in the individual range datasets. The correspondences then vote for the pose that registers the views using a probabilistic Hough transform.

2. Quadric surfaces are øt to the merged 3D data sets based on seed patches extracted from the full 3D description.

3. Improved surface reconstruction is obtained by optimizing the surface øt subject to user declared constraints (such as two surfaces are parallel, or a set of holes are collinear). We formulate the problem as a constrained least square parameter estimation, where the constraints are deøned over the surface shape and position parameters.

Results show that shape and position can be optimized subject to the applied constraints while also improving the reconstruction of surfaces not explicitly subject to the constraints. Local minima of the optimization all seem to be near to the global minimum.

References

1. A. P. Ashbrook, R. B. Fisher, C. Robertson and N. Wergi, "Segmentation of Range Data into Rigid Subsets using Planar Surface Patches", Proc. British Machine Vision Conference BMVC97, Essex, pp 530˘539 September 1997.
http://www.dai.ed.ac.uk
/daidb/staff/personal_pages/rbf/aabmvc97.ps.gz

2. A. P. Ashbrook, R. B. Fisher, "Segmentation of Range Data for the Automatic Construction of Models of Articulated Objects", Proc. IEEE Nonrigid and Articulated Motion Workshop, Puerto Rico, June 1997.
http://www.dai.ed.ac.uk
/daidb/staff/personal_pages/rbf/nam97.ps.gz

3. A. P. Ashbrook, R. B. Fisher, C. Robertson and N. Wergi, "Segmentation of Range Data into Rigid Subsets using Surface Patches", Proc. Int. Conf. on Computer Vision, Bombay, pp ***˘***, Jan 1998.
http://www.dai.ed.ac.uk
/daidb/staff/personal_pages/rbf/iccv98.ps.gz

4. A. Hoover, G. Jean-Baptiste, X. Jiang, P. J. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fitzgibbon, R. Fisher, "An Experimental Comparison of Range Segmentation Algorithms", IEEE Trans. Pat. Anal. and Mach. Intel., Vol 18(7), pp673˘689, July 1996.

5. N. Werghi, R. B. Fisher, A. Ashbrook, C. Robertson, "Improving model shape acquisition by incorporating geometric constraints", Proc. British Machine Vision Conference BMVC97, Essex, pp 520˘529 September 1997.
http://www.dai.ed.ac.uk
/daidb/staff/personal_pages/rbf/nwbmvc97.ps.gz

Integration of vision and reasoning in an airborne autonomous vehicle for traŒc surveillance

Silvia Coradeschi, Klas Nordberg, Lars Karlsson

Dept. of Computer and Information Science
Linkping University
581 83 Linkping, Sweden
silco@ida.liu.se

We have brieÆy presented the WITAS project, funded by the Wallenberg Foundation, which aims at developing a prototype of an airborne autonomous vehicle for traŒc surveillance tasks. During the ørst phase of the project (1997-1999) a simulated environment will be used, and during the second phase (1999-2003) the work will involve a real vehicle.

The project involves research groups for autonomous decision making (lead by Patrick Doherty), Computer Vision (lead by Gsta Granlund), Computer Systems Architectures (lead by Kris Kuchcinski) and Simulation (lead by Peter Fritzson). The leader of the project is Erik Sandewall. One of the most important sensors is vision, which is mainly used for observing the traŒc situation.

We have discussed some issues that should be considered in the integration between vision and reasoning: integration of static knowledge (e.g. from a GIS) and dynamically acquired knowledge; anchoring of symbolic information in visual information; focus of attention; support and guidance of visual skills; and uncertainty.